Technical University of Denmark

Per Bækgaard

Enhancing User Experience in Next Generation Mobile Devices Using Eye Tracking as a Biometric Sensor

PhD Thesis, August 2016

DTU Compute
Department of Applied Mathematics and Computer Science

Enhancing User Experience

in

Next Generation Mobile Devices

Using Eye Tracking
as a Biometric Sensor

Per Bækgaard

Technical University of Denmark
Kgs. Lyngby, Denmark, 2016

Technical University of Denmark
Cognitive Systems
Richard Petersens Plads
DK-2800, Kgs. Lyngby
Denmark
Phone: (+45) 45 25 30 31
Email: compute@compute.dtu.dk
www.compute.dtu.dk

PHD-2016-433
ISSN: 0909-3192

Abstract

A good *User Experience* is not about just “getting the job done” in the most efficient way. It is also about the subjective elements, providing a positive experience to the user while doing so; emotionally and affectively, having the user engage with the service or product.

Knowing when this takes place means we need ways of measuring concepts like attention. The basis for this should preferably be rooted in our understanding of the anatomically based attention networks of the brain.

This thesis looks at biometric markers of cognitive and affective processes; at the overview level *Electroencephalography* (EEG), *Galvanic Skin Conductance* (GSR), *Heart Rate* and *Heart Rate Variability* as well as *Face Expression Detection* – and in much more detail *Eye Tracking*.

A simple framework for relating eye movements and pupil dilations to the visual processing system and to the attentional networks is suggested. It is demonstrated that it is possible to identify components of attention and cognitive load using low cost eye tracking in conventional office settings. It is also shown that aspects of surprise, similar to negativity feedback error coding, is measurable. Behavioural patterns possibly related to time on target, cognitive load, performance or stimuli are inferred. The existence of possibly unique individual gaze patterns related to visual stimuli or to the brain’s *Default Mode Network* are shown.

A way of synchronizing EEG and Eye Tracking is also suggested, and in addition, a few software assets (a Python interface to The Eye Tribe tracker and an implementation of the Attention Network Test (ANT)) have been created.

Resumé (Dansk)

AT FORBEDRE BRUGEROPLEVELSEN I NÆSTE GENERATION AF MOBILE ENHEDER VED HJÆLP AF ØJENSPORING SOM EN BIOMETRISK SENSOR

En god *brugeroplevelse* handler ikke bare om at få ting gjort på den mest effektive måde. Det drejer sig også om de subjektive elementer; at give brugeren en positiv oplevelse undervejs: De følelsesmæssige og affektive aspekter, som får brugeren til at være engageret i et produkt eller i en service.

For at vide hvornår det sker er det nødvendigt at kunne måle begreber som opmærksomhed. Basis for sådanne målinger bør tage udgangspunkt i vores forståelse af det, som anatomisk udgør hjernens opmærksomheds-netværk.

Denne afhandling kikker på biometriske markører for kognitive og affektive processer. Den giver en overfladisk indføring i *Electroencephalography* (EEG), *hudens galvaniske ledningsevne (Galvanic Skin Conductance)* (GSR), *hjerterytme* og *variation af hjerterytmen* og *detektion af ansigstudtryk* – og på et noget mere detaljeret niveau sporing af øjenbevægelser, *Eye Tracking*.

Der fremlægges en enkel model, der kan bruges til at relatere øjenbevægelser og pupil-udvidelser til det visuelle system og til opmærksomheds-netværk. Det demonstreres at det er muligt at identificere komponenter af opmærksomhed og kognitiv belastning under anvendelse af billigt udstyr til øjensporing. Det vises også at elementer af overraskelse, modsvarende kodning af fejl i *negativity feedback*, kan måles. Adfærdsmønstre, som muligvis kan være relateret til opgavetiden, kognitiv belasning, hvor godt man udfører opgaven eller til de anvendte stimuli, bliver udledt. Eksistensen af potentielt unikke individuelle øjenbevægelser, som kan være relateret til visuelle stimuli eller til hjernenes *default mode* netværk, bliver påvist.

En måde at synkronisere optagelser af EEG og øjensporing bliver foreslået. Derudover er der fremstillet nogle forskellige programbiblioteker (en Python grænseflade til The Eye Tribes øjensporingsenhed og en implementation af opmærksomheds-netværks-testen (Attention Network Test, ANT)).

Preface

This thesis is presented in fulfillment of the requirements for acquiring a Ph.D. in Engineering, and was prepared at the Cognitive Systems section of DTU Compute, under the supervision of Associate Professor Michael Kai Petersen and Associate Professor Jakob Eg Larsen.

The thesis deals with aspects of eye tracking as a biometric signal relating to attention and arousal and how it potentially can be used to improve user experience in mobile devices. It includes 4 published papers and 4 in submission or preparation.

Lyngby, August 31st, 2016

Per Bækgaard

List of Publications

Conference Proceedings

(A) Bækgaard, P., Petersen, M. K., and Larsen, J. E. “In the twinkling of an eye: Synchronization of EEG and eye tracking based on blink signatures”. In: *4th International Workshop on Cognitive Information Processing - Proceedings of CIP 2014* [2014], pp. 14–19 [1]

(B) Bækgaard, P., Petersen, M. K., and Larsen, J. E. “Thinking Outside of the Box or Enjoying Your 2 Seconds of Frame?” In: *Universal Access in Human-Computer Interaction. Access to the Human Environment and Culture* 9178 [2015], pp. 186–195 [2]

(C) Bækgaard, P., Petersen, M. K., and Larsen, J. E. “Assessing Levels of Attention Using Low Cost Eye Tracking”. In: *Universal Access in Human-Computer Interaction. Methods, Techniques, and Best Practices: 10th International Conference, UAHCI 2016, Held as Part of HCI International 2016, Toronto, ON, Canada, July 17-22, 2016, Proceedings, Part I*. ed. by Antona, M. and Stephanidis, C. Cham: Springer International Publishing, 2016, pp. 409–420 [3]

Journal Articles

(D) Bækgaard, P., Petersen, M. K., and Larsen, J. E. “Separating Components of Attention and Surprise”. In: *Manuscript submitted for publication* [2016] [4]

(E) Bækgaard, P., Jalaliniya, S., and Hansen, J. P. “Pupillary Measures During an Assembly Task”. In: *Manuscript submitted for publication* [2016] [5]

(F) Cuttone, A., Bækgaard, P., Sekara, V., Jonsson, H., Larsen, J. E., and Sune Lehmann. “SensibleSleep: A Bayesian Model for Learning Sleep Patterns from Smartphone Events”. In: *Manuscript submitted for publication* [2016] [6]

Contributions to Other Papers, Manuscripts and Working Papers

(G) Bækgaard, P., Petersen, M. K., and Larsen, J. E. "The Blank Stare". In: *Manuscript in preparation* [2016] [7]

(H) Laundav, D. K., Jensen, C. B. F., Bækgaard, P., Petersen, M. K., and Larsen, J. E. "Your heart might give away your emotions". In: *Proceedings of the 2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. 2014, pp. 1–6 [8]

(I) Bækgaard, P. "Key Terms in Eye Movement Research (Project Principia; Working Paper)". 2015 [9]

Software Assets (GitHub)

(J) Bækgaard, P. "Simple python interface to the Eye Tribe eye tracker [baekgaard/peyetribe]". In: *GitHub repository* [2014] [10]

(K) Bækgaard, P. "Attention Network Test implemented in PsychoPy [baekgaard/ant]". In: *GitHub repository* [2016] [11]

Acknowledgments

This thesis builds on the wisdom and contributions of many people; some of which I've had the fortune to meet and interact with and others that I've benefited from since they have shared their knowledge in written form.

I would therefore like to thank, first of all, my two excellent supervisors, Michael Kai Petersen and Jakob Eg Larsen! Thank you for creating the opportunity in the first place, for allowing me into your world, for openly sharing your knowledge and networks, for entrusting me with teaching duties and letting me interact with the many clever students, for asking good questions and being curious, for being patient and listening to a great deal of half-baked ideas, and for genuinely inviting me to be an integral part of the team! It's been great working with you guys, and I hope we will continue very much to do so.

I would also like to thank the rest of the section here: to Lars Kai Hansen for hosting me and welcoming me into the team; to Ole Winther, Tobias Andersen, Sune Lehmann and Jan Larsen for good discussions and questions and for sometimes just lending an ear to a stray thought, and for your encouraging way of interacting with me; to Finn Årup Nielsen and Mikkel N. Schmidt for help and support; to Vagn Lundsgaard Hansen for being able to solve any kind of imaginable math question; to Tommy Sonne Alstrøm for many good (and welcome!) interruptions; to Bjørn Sand Larsen, Jens Madsen, Lasse Lohilahti Mølgaard and Matthew George Liptrot for help and assistance and for teaching me how to pronounce Norepinephrinergic (which turned out not to be the correct term anyway); and to many of my fellow students here I've had the pleasure to interact with (or to eat their cakes and cookies): Camilla Birgitte Falk Jensen, Andrea Cuttone (it's been great sharing office with you), Rasmus Bonnevie and Michael Riis Andersen (good to have you guys when getting lost between some Bayesian Prior and the corresponding Posterior...), Simon Due Kamronn, Ditte Høvenhoff Hald, (Kit) Melissa Larsen, Karen Marie Sandø Ambrosen, Sofie Therese Hansen, Benjamin Johansen, and many others... and of course to Marian Probst and later Wanja Andersen for taking good care of all of us!

And to those outside the section I've interacted with: Thanks for your encouragements and ways of being: Michael Rose, Christian W. Probst (who got me to teach at DTU in the first place), to my old co-student Jan Madsen that never left the place, Sven Karlsson, Michael R. Hansen, Ekkart Kindler, Robin Sharp (that hooked me back into academia shortly after I tried to leave after graduation at a young age) and to many others! And to Aki Vehtari for not only reminding me of many good days in beautiful Finland but also for sharing your wealth of knowledge and for coaching me, and for co-authoring the definitive book on a difficult topic.

To all the students I've worked with and have been able to engage in my many weird ideas or get to tinker around in some Python code: You were great, and I sincerely appreciate all your efforts and have enjoyed the interaction! This goes to all of you, and in particular to Emil, Benjamin, Tine, Sung-Won, Oisin, Patrick, Stine, Emilia-Stefania, Robert, Emily, Anastasious, David, Andre, Sigurd, Eleftherios, ...

And to the rest of the extended project team: Thank you for the cooperation and support, and in particular to my eye tracking mentor John Paulin Hansen – you've been really excellent – to my old colleague Rasmush Dahl, to Olav Gjerlufsen, Thomas Visby Snitker, Florian Biermann, Shahram Jalaliniya, Henrik Skovsgaard, Javier San Agustin and Anders Pedersen.

Thank you also to my old M.Sc. supervisor, A.P. Ravn: You may not remember, but many years you asked me when I would come back to the academic world and do a Ph.D., and I never could let go of that thought again. So – well, now I finally did it, and I haven't regretted a second of it!

I would also like to say a heartfelt thank you to Innovation Fund Denmark for the funding, and to Otto Mønsted Fonden for sponsoring my trips to some key conferences to present parts of this work, get feedback and interact with like-minded researchers. I'll do my best to let the research come back to society so that others hopefully can benefit from it as well.

And then lastly, and maybe most importantly, I would like to express my gratitude to my closest family: To my parents for always being there and encouraging me to be curious, to my siblings and their lovely families – that also ask good questions and make me search deeper! And more than anything to our lovely kids and their spouses: Anne Dorte, Simon, Nicolai and Lisa: you guys rock! And to Lene, the love of my life, who has endured a great deal of extra work, but always, always has been supportive and encouraged me to pursue the dream. And you're still here with me!

And Now for Something Completely Different...

Table of Contents

List of Abbreviations

AOI Area(s) of Interest. 19, 20, 21, 22, 25

AUC Area-Under-Curve. 47, 50

BOLD Blod-Oxygen Level Dependent. 4, 41

CON Cingulo-Opercular Neural. 36

DFT Digital Fourier Transformation. 19, 50

DMN Default Mode Network. 32

ECG Electrocardiogram. 5

EDA Electrodermal Activity. 4

EEG Electroencephalography. 2, 3, 4, 7, 27, 40, 86

EMG Electromyography. 27

EOG Electrooculography. 27

ERP Event Related Potential. 3, 27, 33, 34, 40

ET Eye Tracking. 27, 86

FDM Fixation Density Map. 20, 21, 22, 31, 49, 50

FIR Finite Impulse Response. 25

fMRI Functional Magnetic Resonance Imaging. 2, 3, 4, 41

GSR Galvanic Skin Response. 4, 5

HCI Human Computer Interaction. 17

HF High Frequency. 3, 5

HR Heart Rate. 5, 6, 7

HRV Heart Rate Variability. 5, 6, 7, 42

ICA Index of Cognitive Activity. 44

KLD Kullback-Leibler Divergence. 22

LC Locus Coeruleus. 36, 41

LC-NE Locus Coeruleus-Norepinephrine. xii, 33, 36, 38, 39, 41, 42, 43, 45, 47, 55, 57, 62, 63, 65, 72

LF Low Frequency. 3, 5

MSE Mean Square Error. 22

NE Norepinephrine (a.k.a. Noradrenadlin(e)). 36, 38, 41

OPM Oculomotor Plant Model. 10, 17, 19, 25, 31, 47, 48, 52

PPG Photoplethysmograph. 5, 7

QS Quantified Self. 2, 3, 4

RMS Root Mean Square. 20

SEM Standard Error of the Mean. 60, 69

Task-Evoked Pupillary Response Task-Evoked Pupillary Response. 40, 41, 42, 55, 63, 65

UX User Experience. 1, 2, 11

VR Virtual Reality. 17

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Towards a Better User Experience

The quest to put the user at the center of the experience has been ongoing long before the term User Experience (UX) in the 1990s became a buzzword. Keith [12] put it like this back in 1960:

> *The consumer, not the company, is in the middle. In today's economy the consumer, the man or woman who buys the product, is at the absolute dead center of the business universe. Companies revolve around the customer, not the other way around.*

To some, this may still be wishful thinking!

While traditional Western manufacturing was centered around materials resource planning, the Japanese concept of *lean*, which originated as early as 1940 with Toyota [13], focused on what adds value to the end customer and ultimately aimed for perfection. Lean has since spread to other parts of the world and to many other domains; we now even have *Lean UX*, which focuses on creating products in lean startups (and maybe also elsewhere) while having the creation of value for the user at the center.

Indeed, to the author of this thesis, UX is about putting the experience of the *users* at the center of thinking, in planning, execution, delivery and maintenance. But not only that! It is also about moving beyond the instrumental; away from naïvely thinking about how to most efficiently allow the user to solve a set of tasks and towards providing a more holistic *experience* to the users engaged with a product or service!

In an influential paper, Hassenzahl and Tractinsky [14] looks at UX from three different perspectives of:

- Beyond the instrumental: Holistic, Aesthetic and Hedonic
- Emotion and affect: Subjective, Positive, Antecedents & Consequences
- The experiential: Dynamic, Complex, Unique, Situated, Temporally-Bounded

This type of thinking has profound impacts on the way we design our products and services. Forlizzi [15] even puts it like this: "*Emotion is at the heart of any human experience*".

The author of this thesis can share a personal experience of living through a business transformation with the decline and fall of a company that failed to understand (or at least to execute) in a global market where delivering the "best" performing technical solution was not enough, and where at least one competitor much better understood the dynamics of these elusive, intangible and emotional drivers [16, 17]! If nothing else, this now provides ample motivation for doing better.

Hassenzahl and Tractinsky [14] conclude as follows:

> *UX is about technology that fulfils more than just instrumental needs [...] UX is a consequence of a user's internal state (predispositions, expectations, needs, motivation, mood, etc.), the characteristics of the designed system [...] and the context [...] within which the interaction occurs [...].*

This thesis will take an extremely narrow focus on improving UX in mobile systems; we will single in on the aspect of the user's internal state, and look at ways through which systems can get additional insights into the cognitive and/or affective processes that take place internal to (and sometimes unconsciously to) the user. And we will exclude technologies that are not (becoming) mobile and have a fair chance of becoming ubiquitous.

It is not a thesis that digs up completely new findings in neuroscience; we aim to build on foundations that others have prepared. Rather, we will in try to utilize findings "from the lab" *in-the-wild*, with all the complexities and challenges that may bring.

1.2 Biometric Markers of Cognitive Processes

In looking for insights into cognitive processes, we have seen remarkable steps forward in the last decades due to advancement of technologies such as Electroencephalography (EEG) and Functional Magnetic Resonance Imaging (fMRI)[1], which

[1] ...even though we may also have suffered from a few potential setbacks too [18]!

have brought us a much better and anatomically grounded understanding of many cognitive and affective processes. Some of these technologies are becoming portable, in part helped by the Quantified Self (QS) movement [19]: ([20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32] and many others). One review that sums up several of the advantages and challenges of the different modalities, although with a primary focus on psychophysiological methods in games development, is done by Kivikangas et al. [33]. We will have a very brief look at some of these before looking further at the technology of choice in this thesis: *eye tracking*.

Not all of them are, however, equally practical from a mobile or a QS perspective, even though they are known to have strong cognitive neural correlates; as we shall see, EEG Event Related Potential (ERP) components are relatively well described, but currently not so practical from a mobile or QS perspective. Hence, our interest is also as to whether some of the same underlying signals can be detected in contact-less and non-invasive eye tracking measures.

The value of some of these technologies may, incidentally, not just lie in what they can do alone, but even more in what can be done when combining several modalities.

Other technologies which are not going to be portable in a foreseeable future will only be referred to briefly as/if needed.

1.2.1 EEG

EEG has long been a “tool of the trade” in neuroscience and elsewhere, and has been around for a long time. It has almost the opposite characteristics of fMRI; it is characterized by high temporal resolution, but a relatively low spatial resolution [34]. In its most basic form, an EEG is simply a recording in multiple channels of the electrical potential with respect to some selected point of reference on the outer surface of the scalp. It originates from the firing pattern of (groups of) neurons inside the brain, but is also heavily influenced by powerful artifacts, such as muscle activity and eye movements. Fig. 2.5 (chapter 2) illustrates this well: The large peak is an artifact related to a blink, whereas the hardly visible activity is the desired signal from the brain. Additional processing is thus required to remove the unwanted parts of the signal in order to estimate the original source(s), such as Independent Component Analysis [35] or other techniques needed to solve the so-called “Inverse Problem” of Source Localization [36, 37].

Modern EEG studies are typically based on either ERP paradigms or are done in continuous mode, where focus is on fluctuations in frequency bands over time (e.g. by looking at the power spectrum).

Just to give a few examples; Coull [38] lists a number of attentional neural correlates that are visible in the EEG: In the ERP paradigms, some of the more well-known may be 1) the N100/P100 modulation by attention/early selection processes in the early visual or auditive stages [39], 2) the P300, positive deflections elicited around 300 ms, seen when attending to task stimuli and modulated in magnitude and latency (P3a and P3b components) by the degree of the unexpectedness of the event and by how difficult discrimination is against the expected target [40, 41, 42] (this includes the so-called “oddball” paradigm), or 3) the N400, which is a negative waveform around 400ms, indicating continued allocation of resources to semantic processing [43]. Arousal is visible in the High Frequency (HF)/Low Frequency (LF) ratio; activity in the HF bands (α and β) declines with decreasing vigilance or gradual sleep onset and LF activity increases in the (δ and θ bands) [38].

EEG has the disadvantage that it requires direct skin contact and unless special measures are taken, for instance as proposed by Debener et al. [44], it is somewhat unpleasant to wear for extended periods. To the authors knowledge, very few continuous EEG recordings are done in the QS domain, and probably for a good reason.

1.2.2 Galvanic Skin Response (GSR)

The idea of looking at electrical signals through the skin dates back to 19th century [45], (originally involving zinc sulfate solution), but the more modern concepts probably were introduced by Johnson and Lubin [46], and soon after (1971) efforts to standardize measurements were brought in place [47]. In its modern form, it typically uses two electrodes to measure a signal (part of what is also called the Electrodermal Activity (EDA)), which essentially is the electrical conductivity (or resistance) of the skin, which in turn is strongly impacted by the activity of the approximately 3 million sweat glands.

Fundamentally, according to Critchley et al. [48] and Boucsein [45], skin conductance corresponds to the level of sympathetic activation, and thus can, in our context, be used as an (indirect) measure of attention, effort or arousal[2]. Critchley et al. also demonstrate how changes in the GSR signal is related to activity in specific parts of the brain, as seen via fMRI Blod-Oxygen Level Dependent (BOLD) activity. A more detailed summary is found in [45].

One of the reasons GSR is of interest outside lab settings is that GSR sensors can be worn unobtrusively. Dr. Rosalind Picard of MIT is, at least in some circles, known for pitching the idea of constantly monitoring this signal[3] in order to find deviations

[2] As everyone knows, it is of course also an important part of the body’s thermal regulation system.

[3] https://nihrecord.nih.gov/newsletters/2016/04_22_2016/story2.htm

from a “normal” baseline. Healey and Picard [49] henceforth tested out GSR in a real-world setting and reported strong correlates between skin conductance and perceived “stress” when driving a car. In a later work, Sano and Picard [50] used daily monitoring of GSR to detect stress, and used this in a study with other stress markers (some collected by a mobile phone) to check accuracy of the markers.

It is one of the more established markers of emotional variance and cognitive load, and Holmqvist [51] suggests it to be used in connection with some of the eye tracking parameters like pupil dilation and saccadic amplitude, as a multimodal signal can help to deduce the underlying causality. However, in a more recent work by Picard et al. [52], differences between left and right side upper body measurements are seen, and should lead to some caution when trying to interpret data and infer a traditional emotional arousal index. The paper suggests extensions of previous theories and has lead to additional commentaries and discussions, also addressed by Picard et al. [53]. So even if the situation may at present be somewhat inconclusive, it may be wise to be careful with emotional interpretations of GSR.

1.2.3 Heart Rate (HR) and Heart Rate Variability (HRV)

Another technology that has seen some diffusion into the general society is measuring HR and HRV data. Portable trackers used for fitness and smart-watches can often collect heart related data; originally maybe with the intent of tracking how much people exercised, but there are many other uses too.

The HR, which can be measured in several ways, is essentially the number of heartbeats pr. minute. Originally this was referenced with respect to the distance between the two strongest peaks of neighbouring heartbeats in an Electrocardiogram (ECG), the R-R[4] distance. An alternative to a direct electrical measurement (as used in some portable chest-strap monitors like the Polar H7[5] chest band), is using some variation of a Photoplethysmograph (PPG), possibly built into a watch or a fitness tracker or a similar wearable device[6]. A PPG typically uses one or more LEDs and photo sensors to measure variations in light absorption in the skin, and it can estimate the heart rate based on fluctuations caused by the pressure pulse sending blood through the tissue of the body.

[4] A normal ECG has three noticeable peaks in the central and most prominent wave, called Q (-), R (+) and S (-), with the polarity indicated. The R-R distance is normally easy to identify and therefore probably the best candidate for a standardized measure.

[5] http://support.polar.com/en/support/H7_heart_rate_sensor

[6] As an example, the Apple watch, according to an iFixit tear down/review, reportedly uses a PPG; see http://www.cultofmac.com/320322/apple-watch-sensors-are-capable-of-measuring-blood-oxygen/ which summarize the iFixit review.

HRV is a measure of the variability of the R-R intervals. A signal can be formed from the R-R distances, and the power spectrum of this signal is what is used to estimate HRV. The HRV is often divided into it's LF and HF content; The Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology [54] recommends using power in the range of 0.04–0.15Hz (LF) and 0.15–0.4Hz (HF) to characterize the HRV in a comparable way. Variations in HR have (at least) two sources, that both stem from the autonomous nervous system: Both the sympathetic[7] and the parasympathetic[8] nervous system influence the HR. There is not complete agreement on the content of the LF component, but it is thought to reflect either both the sympathetic and the parasympathetic systems or predominantly the sympathetic system. However, the HF power is mainly thought to reflect the parasympathetic system. Thus, the ratio of the power ratio LF/HF is an indicator of the balance in activity between the sympathetic and the parasympathetic nervous systems [54]. It may also be worth mentioning that there is a relation between HR and respiration.

As examples of studies: Wallentin et al. [55] reported that the LF/HF ratio correlated with the perceived intensity of a narrative ("The Ugly Duckling") being read out loud; hence an increased arousal with increased sympathetic activations leads to an increased LF/HF ratio. Another example is Taelman et al. [56], that intends to use HRV to assess "stress" by demonstrating increased LF/HF caused by mental tasks. Löw et al. [57] used a prey-predator scheme to demonstrate several body reactions, explaining that the underlying hypothesis is that motivationally relevant stimuli leads to a heart rate acceleration, and tasks demanding vigilance for response cues result in a heart rate decrease (not completely unlike orienting bradycardia or fear bradycardia in animals). Lane et al. [58] looked at correlation between specific brain regions thought to be involved in emotional processing, and especially the HF components, and found that increased activity in certain regions leads to a decrease in HR (which may sound counter-intuitive, but consistent with theories proposed by the authors). A later study by some of the same group, Thayer et al. [59] has a comprehensive meta-analysis, and amongst other things suggest that HRV might serve as an index of how strongly autonomic activities are shaped by top-down processes.

Kemper et al. [60], while looking at how to deal with outliers, also notes that an increase in HRV (up to a point) reflects adaptability and resilience or can be the result of therapeutic interventions, while decreases reflect rigidity and poor health status and poor prognosis with a wide range of clinical conditions – so many factors

[7] Sympathetic: the "fight-or-flight" system; the excitatory system or the "gas pedal".

[8] Parasympathetic: "rest and digest" system; the inhibitory system or the "brake".

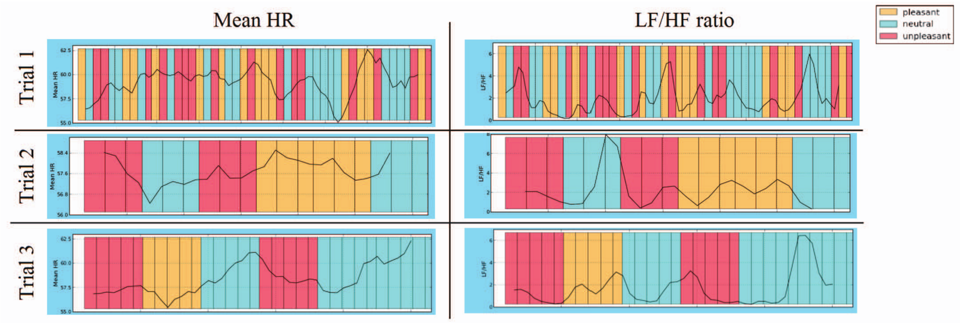


Figure 1.1: Heart Rate and Variability from three trials showing pleasant, neutral and unpleasant images in different permutations to the subject. Variations in the Mean HR and the ratio of LF to HF content in the HR may provide clues to the emotional and cognitive state of the subject – although in this case, even though there may be a pattern in the Mean HR, it turned out to be difficult to relate the LF/HF ratio to the presented stimuli... Illustration from [8].

can impact the measurements! Other studies that look at more practical issues and tools are Vila et al. [61], Rodríguez-Liñares et al. [62], and Pradhan and Islam [63]

Finally, Porges [64] has a very good overview of the history of HRV and presents a comprehensive framework ("the polyvagal perspective") that links neurophysiological activity seen in the autonomous nervous system (and reflected in the HRV) to behavioural patterns, and discusses the aspect of respiratory modulations of the HR (Respiratory Sinus Arrhythmia, RSA) in more detail.

Using a PPG instead of performing electrical EEG recordings to derive HRV is possible in some situations when subjects are at rest, as demonstrated by Giardino et al. [65], but results do not translate directly between PPG and more conventional EEG recordings.

One interesting and final twist on HR, although it may not fully be in our present scope here – despite using mobile tools for recording HR and involving LEGO® construction kits so we'll mention it anyway – Social interaction impacts the heart rate; Mitkidis et al. [66] found signs that heart rate synchrony was associated with higher trust in social teams.

An example of HR and HRV recordings using a finger-PPG can be seen in Fig. 1.1. Judging from the authors experiences, it is not always easy to reproduce larger studies in smaller settings, and especially if dialing in on single-subject performance.

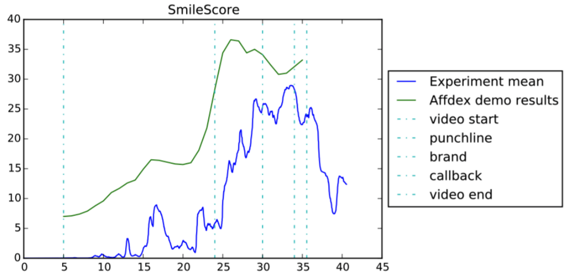


Figure 1.2: AVERAGE SMILE SCORE of people watching a commercial, compared to an average value published by Affectiva. Illustration from a M.Sc. thesis done by Bunk [69].

1.2.4 Face Expression Detection

Dr. Rosalind Picard, mentioned earlier in the context of wearable devices, also co-founded Affectiva, which – on the basis of applying machine learning on a large body of images (described by McDuff et al. [67]) – has created a service to interpret facial expressions [28].

The basis for this is the works of Ekman and Friesen [68], who created the Facial Action Coding System; an objective set of metrics that characterize facial expressions in terms of individual components, decoupled from any emotional meaning. Some simultaneously occurring components are known to correlate with specific emotions, like disgust, anger or surprise. This can be used to give a semantic interpretation to registered expressions, although it may be dependent on specifics of the person or the context[9].

Face expression detection is a new modality that could also serve as relevant input, possibly in connection with other modalities. In some preliminary work done by students working with the author, like Bunk [69] and others, it was possible to reproduce some of the findings reported by Affective using their services on data recorded in our settings (see Fig 1.2). It was also found, however, that social interaction can have a strong impact on facial expressions, so it may not be as straightforward as it could otherwise sound to bring this technology into practical use.

[9] As anecdotal evidence: In some student work performed in our lab, it was for instance noted that some people appeared – at least to the Affective system – to maintain a constant smile where a human observer familiar with the subject interpreted it more like focused concentration or some other expression not identified as a smile signalling joy.

1.2.5 Eye Tracking

We will end our short summary with a couple of considerations on Eye Tracking before we dig deeper into the details.

Eye Tracking has the advantage of being contact-less and non-intrusive in most settings. It is also becoming mobile; the latest Samsung Note 7 phone reportedly has a built-in Iris scanner[10], which means it in theory might be capable of estimating pupil size; a metric that is of particular interest.

There are many correlates between eye tracking metrics and cognitive and affective processes – but it would be fair to mention also that there may not be unanimous agreement on the applicability of all of the technologies, and in particular of using pupillometry.

Based on the author's own experiences, and further anecdotal evidence, there can be at least four different approaches to eye tracking:

- Eye tracking in controlled, clinical conditions, is a useful tool (most would probably agree to this view).

- Some see a big potential in eye tracking (or even a possible business case) and may jump too quickly to unfounded conclusions, possibly extending the eye-mind hypothesis and thus over-interpreting gaze patterns to equate interest (i.e. "you looked at it" may become "you have an interest in it" or even "you will buy it").

- Others, maybe having seen the previous group or maybe themselves having difficulties reproducing previous studies in less controlled circumstances, declare it a failure and run away, sometimes warning others not to go in the direction they gave up on (as may be the case regarding pupillometry).

- Finally, there is a group that humbly believes there is a lot of hard work to be done, aiming to understand the scientific basis, building on previous studies and daring to test out "wild ideas in-the-wild" to check the robustness, with the risk of failing and openly admitting so.

The author has met people of all 4 opinions, but would like himself to be counted as part of the last group. We may certainly not have done everything "by the book", and many things should have been done differently had we been wiser. But we hope we're now a bit wiser due to all the things we didn't get right the first time! And in the process, we believe we have found bits and pieces of technologies that may be robust enough to also have a practical use outside controlled settings.

[10] http://www.samsung.com/global/galaxy/galaxy-note7/security/

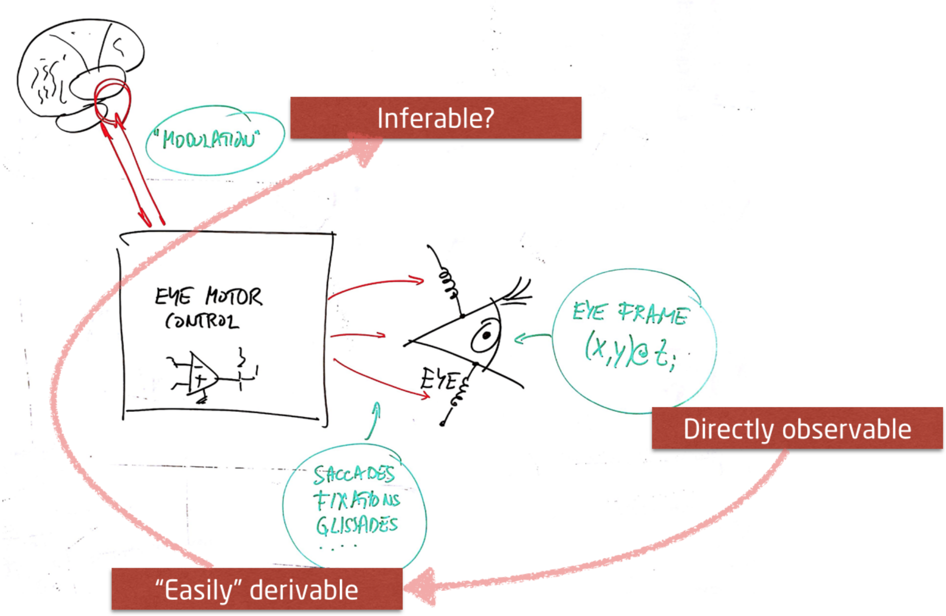


Figure 1.3: The Eye–Mind Framework: The eye (figuratively illustrated at the center), can be directly observed by means of eye tracking technology, that can measure the current gaze point and pupil size and detect blinks. The underlying saccades and fixations, which are driven by muscles controlled by neuronal activity (box left of the eye), can be determined based on analysing the observed gaze points. The brain (top left), with its cognitive and psychological processes, modulate and control this system. Where the eye itself and the motor control to a large degree is part of the OPM, the focus of this thesis is more on gaining information about the cognitive processes in the brain that control and modulate these aspects.

It's also not all theories that we have met that are easy to reconcile. In fact, some of them may be inconsistent or mutually exclusive with others. It may even not only be the author that struggle to understand them. Some of what is presented here may fall in any of these categories too, but hopefully it can serve as a bank of ideas to pursue or at least as questions to clarify. Certainly the author have some of those!

1.3 The Eye–Mind Framework

The drawing on Fig. 1.3, being a verbatim copy (i.e. a photo) of an original whiteboard drawing[11], has been guiding this work for a long time. It may be helpful to view the eye–mind as a three layer model, with attributes on each layer:

- At the lowest level, we consider the eye having a gaze point (the foveal center of the eye) that can be directly measured at a given time, as can the size of the pupil (chapter 2).

- At the next level is the circuitry that moves, and controls the movements of, the eye. This is partly described by the Oculomotor Plant Model (OPM) proposed by Komogortsev and Khan [70], building on earlier works by Robinson [71], [72], [73], [74], [75], Young and Stark [76], Sparks [77], Leigh and Zee [78] as well as probably also others. Together with the lower levels of the neuronal systems inside the brain, this system takes care of handling the next level of eye activity: keeping the eye still for long enough to take in data (fixations), moving it to the next position (saccades), closing it when needed (blink), regulating the amount of light that hits the retina (pupil constrictions), etc. These are usually not conscious processes. (Chapters 4 and 2).

- At the highest level we find cognitive and affective processes that modulate the lower levels[12]. These are the systems that e.g. regulate attention, choose an appropriate top-down search strategy and interrupts it when salient enough stimuli appear (chapter 3).

It is our hope that we, by observing the lowest levels and deriving the “easy” next level parameters eventually will be able to infer not only the presence of individual finger-print like signatures on the second level, but also infer and understand some of the variations at the cognitive and affective levels and thereby learn something about the *users experience* in the present context.

1.4 An Overview of This Thesis

From this brief introduction to UX and to a few important biometric signals, we will continue looking at, first the mechanics and methods of Eye Tracking, in order to

[11] Internally in the research group it was quickly coined “strudsen” (the ostrich) due to the similarity between what the author intended to be an eye and an ostrich.

[12] Our main focus here is on the attentional systems rather than the visual processing systems, and we will to some extent conveniently leave out discussions of the latter systems.

build an understanding of not only the current state and the technological basis but also of how cognitive processes can influence some the specifics (Chapter 2).

Next, we will look at attention and arousal; not primarily as psychological constructs but more as concepts rooted in an anatomical understanding of the brain's networks, and we will investigate how it modulates the diameter of the pupil. This further leads to an introduction to phasic and tonic activations related to the sympathetic division of the autonomous nervous system, and how the pupil can be used as a proxy for assessing such activations (Chapter 3).

Before looking at the details of these activations, we will make a short de-tour into the land of individual differences, acknowledging that we're all different and that measurements need to allow for such differences before we can say anything sensible with respect to the underlying causes of signals we see (Chapter 4).

Then we're ready to look at application of measuring phasic activations (Chapter 5), in particular in the context of paradigms where stimuli timing is known, trying to infer elements of the cognitive impact we may temporally induce. From there, we are ready to look at applications of measuring the tonic level (Chapter 6), which can give us a further insight into arousal and maybe into levels of a subjects' "engagement" or experienced "cognitive load".

Before concluding, we will look at other ways "weak signals" can be used: what can be learned from just knowing if a subject is looking at the screen or not, what can we learn about a subjects' sleeping patterns just by looking at whether the screen on his/her smartphone is on, and how can we use (normally discarded) blinks to synchronize between the modalities of EEG and Eye Tracking (Chapter 7).

1.5 Contributions in This Thesis

This thesis brings contributions in the following areas (primarily based on low cost eye tracking):

- It demonstrates that it is possibly to identify components of attention and cognitive load in decision tasks based on pupil dilations (chapter 5, Sections 5.1, 5.2 and related to Chapter 6).

- It additionally demonstrates that aspects of surprise in decision tasks, similar to negativity feedback error coding, can be detected in pupil dilations (chapter 5, Section 5.2).

- It identifies the existence of possibly unique individual gaze density patterns related to both visual stimuli and to the brain's default state mode, based on

DBSCAN clustering, further proposing spatial Fourier filtering to extract these (chapter 4).

- It shows that blink detection for synchronization of eye tracking and EEG may be a feasible option (chapter 7, Section 7.3).
- It infers behavioral patterns from time on target, further attempting to relate these to differences in induced cognitive load and own performance (chapter 7 Section 7.1 and Chapter 6).

While doing so, it proposes a framework, based on established theories, in which eye movements including pupil dilations can be interpreted in terms of how they may be affected by cognitive processes (Section 1.3 and Chapter 3). In addition, a few software assets has been created that may be helpful to others: An interface [10] to the low cost eye tracker from The Eye Tribe, a re-implementation [11] of the Attention Network Test [79] in the PsychoPy [80] environment, as well as clustering algorithms and tools for working with eye tracking data that have been used within the project and in related student works.

Chapter 2

Eye Tracking Mechanics

The human eye is a remarkable and very complex instrument. Luckily, despite all of it's complexity, most of us don't need to worry about how it works in order to use it in our daily activities. In this chapter, however, we will nevertheless look at some of the mechanics of the eye and of eye tracking.

2.1 Eye Tracking

Eye Tracking as a discipline dates back to the late 1800s [51], and the first solutions were of limited practical use and very invasive, essentially gluing a ring[1] to the eye temporarily; one of the earliest pioneers Delabarre [82] sums up his experience as follows:

> *As to whether there is any danger to the eye to be feared from using it in this manner, I cannot say with assurance. I have myself always suffered a little temporary inconvenience [...] I have also found it necessary to allow a considerable interval to elapse between experiments [...] The unpleasant effects have always soon passed, and now, a full year since my last experiments, I can detect no ill effect.* [82]

Obviously measuring this way impair the free movement of the eye significantly. Dodge and Cline [81] were the first [73] to suggest a more practical way using recordings on film instead, eventually paving the way for less invasive technologies. Today there are even better, less cumbersome and time consuming and promising solutions on the market, but we will refer to the excellent books of Duchowski [83] and Holmqvist [51] for more thorough historical and technological reviews.

[1] This is what Dodge and Cline refer to as the "*Delabarre eye cup*" [81]

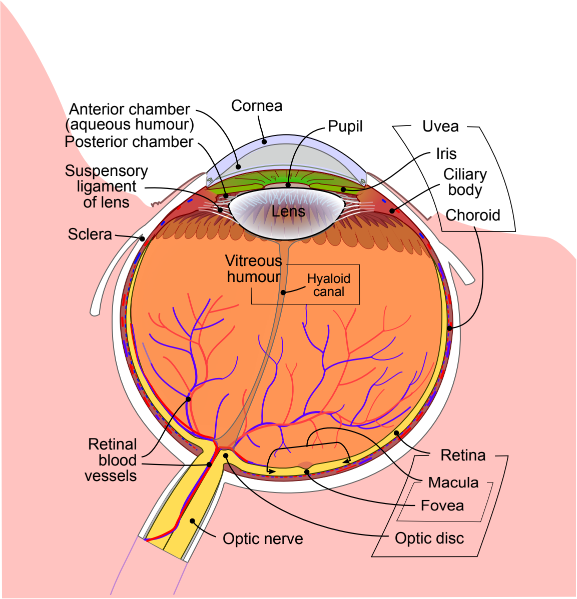


Figure 2.1: THE HUMAN EYE. Light from the external scene passes through the lens, building an image on the retina, from where it is encoded by light-receptive rods and cones, and passed on to the brain for further processing. At the center of the retina is an area of densely packed cones, *fovea*, providing high acuity. The point in the external scene that corresponds to the central position of the fovea is often referred to as the *gaze point* or *point of gaze*. The pupil's main function is to regulate the amount of light that hits the retina; in high light scenarios the normal pupil constricts whereas dim scenes causes the pupil to dilate. The front of the eye is protected by the cornea. Illustration in the public domain, created by Rhcastilhos/From Wikimedia Commons.

What has caused the recent increased interest in eye tracking is partly due to the new data collection options that allow us to measure "*how we view, see and experience the world*" [84, p. v], likely amplified by the availability of off-the-shelf solutions – and now even at an entry price in the 100-200 USD range [26] for solutions that can be used for research [85, 86].

Technology in the devices we have used is based on reflecting infrared light off the pupil and cornea, determining the position of the pupil center relative to (multiple) cornea reflections known as *Purkinje reflections* [87, 88, 89, 83, 90]. Multiple light sources reflecting off the cornea allow the tracker to distinguish eye movements from head movements, giving the user some level of freedom to move around, as in the "head-free" or remote The Eye Tribe system [26], where the tracker is placed in a fixed position relative to a computer screen. Alternatively, it can be built into a pair of glasses and thus mounted in a fixed position relative to the head, with a front-facing camera to determine the head pose, recording the scene in front of the user, like the Tobii glasses [91] do.

Both technologies are mobile or could be used in mobile applications. Eye trackers can also be built into Virtual Reality (VR) systems with the additional purpose of enabling foveated rendering, like proposed by Guenter et al. [92] or recently demonstrated by SMI [93, 94]. This, however, requires not only estimation of the current gaze point, but also an effective eye movement prediction scheme [95], which in turn needs eye tracking with high sampling rate and low jitter, and the use of elaborate models like the OPM [70, 96, 97] or some other means at modelling the trajectories [98]. Even though we suspect the use case of foveated rendering may become the one that will drive a more widespread use of eye tracking, it's details are outside the scope of this thesis.

Duchowski [99] provides a hierarchical division of eye tracking applications which at the top has *interactive* systems and *diagnostic* systems. At the next level, interactive systems are divided into *selective* systems where the gaze point resembles the use of e.g. a mouse, and *gaze contingent*[2] systems where e.g. scene rendering is contingent on the gaze point – like now seen in foveated rendering in VR glasses. Examples of diagnostic systems outside of neuroscience and psychology are e.g. found in human factor studies or within marketing and advertisings. Further, Karray et al. [100], looking at aspects of Human Computer Interaction (HCI), describes *adaptive* HCI designs which adapt to the "*cognitive and affective levels of user activity*".

Hence, for our purposes, it may be helpful to consider applications of eye tracking as broadly falling into one or more of three main categories (that are *not* mutually exclusive):

- *Active systems*, where the main focus is on the user who interacts with the system, giving commands or entering input.

- *Passive systems*, where data is collected for offline research and analysis, possibly impacting later revisions of the system.

- *Adaptive systems*, where the system uses signals from the relevant modalities to adapts its behaviour to e.g the cognitive and affective level of the user.

Naïve ideas of gaze control in active systems may have lead to naming the *Midas Touch* problem in eye tracking, described by Jacob [101]: If gaze activates a command, then you cannot look anywhere without activating a command, which quickly makes it very hard to navigate a user interface! Better solutions have been proposed,

[2] Duchowski further divides gaze contingent systems into *screen based* and *model based* systems.

which can be divided into fixation, saccade or smooth pursuit based. Møllenbach et al. [102] has a summary of the methods and suggests a unifying taxonomy.

Similarly, naïve ideas of using passive systems can also lead to incorrect conclusions. For instance, directly coupling the gaze point to attention is prone to mistakes: As shown by Posner et al. [103] and Posner [104], the attentional focus[3] is not directly coupled to the foveal focus (or gaze point in our terminology). And even if it would have been, equating gaze points or foveal attention to interest in that particular area cannot be done; as noted by Hyrskykari et al. [105], a long gaze does not mean the subject has difficulties comprehending the meaning of, or recognize and has an interest in, the area. It might as well indicate the subject is considering other aspects of an ongoing task or finds some parts confusing. Bol et al. [106] surveys how eye tracking is currently used in communication research, and in particular in advertising, with often very large number of participants, and politely points out that there is no commonality to how results are reported, making them incomparable, and that there is a need for identifying other ways of using eye tracking to maximize its benefits. Thus, taking a more rigorous approach, based on established cognitive paradigms related to e.g. activation of attention– and default mode networks in the brain, as well as improved methods of classifying gaze density patterns and relating these to individual baselines, is called for.[4]

Taking heed in the warnings from these areas, we will nevertheless in the present work focus on aspects of eye tracking that can lead to insights into the user's level of cognitive effort. As such we're mainly driven by its potential use in adaptive systems, although the same insights can also be of value when doing offline analysis in passive systems.

One final comment is called for: Since we're particularly concerned about mobile applications, we're looking for potentially robust signals that translate well between different environments and thus can be used "in-the-wild". It also often implies a scarcity of computational resources and maybe lower resolution, precision and accuracy. We may not be able to deliver on this promise just yet, but it is our hope that building on existing and well understood paradigms, showing how they can translate in-the-wild, is a valid scientific approach that can lead to new applications.

2.2 Classification of Eye Events and Key Metrics

Eye movements can conveniently be grouped into three main classes, as detailed in textbooks like Holmqvist [51] and Duchowski [83], which this summary is

[3] Posner et al. calls this the "*attentional spotlight*".

[4] From a computational perspective, a recent summary of visual attention modelling with a taxonomy of nearly 65 previous models can be found in Borji and Itti [107].

based upon and to which we will refer for additional details beyond what is briefly mentioned here. We will include also blinks and pupil size metrics, and hence cover:

- *Fixations*: when looking at Area(s) of Interest (AOI)
- *Saccades*: when moving between fixations, and
- *Smooth Pursuit*: when following an object, as well as
- *Blinks*: when the eyelids cover the eyes.
- *Pupil Size*: Changes to the pupil size.

2.3 Fixations

Fixations are periods where the eye is (almost) at rest, and all gaze points are densely grouped around an AOI, for at least some tens of milliseconds and up until several seconds, but typically in the range of 0.2–0.3s[5].

Microsaccades are small movements of the eye during fixations. These *fixational eye movements* have been thought to be a "bug-fix"; a way of counteracting neuronal adaptation in the spatial domain, as documented by Martinez-Conde et al. [109] and Martinez-Conde and Macknik [110]. This has been somewhat disputed [111], and in a recent work by Rucci and Victor [112] it is suggested that "the unsteady eye" may be more of a "feature" than a "bug", helping to encode spatial information in the spatio-temporal domain, facilitating higher-level processing in a more complex way rather than simply a means of encountering neuronal fading.

Fig. 2.2 visualizes a 10 second section of a high sample-rate eye tracking session. The gaze point of each data frame from the eye tracker is marked with a dot; green dots are part of saccades and pinkish dots are part of fixations. Microsaccades are clearly visible, and here predominantly vertically oriented. There is anecdotal evidence[6] that eye tracking researchers can differentiate between a smaller group of participants just by looking at their microsaccades, so possibly these are influenced by the specific characteristics of the OPM. Methods exist or are being proposed that can characterize sequences of fixations (section 2.3.2); we would suggest that some of these could also be used to characterize a fixation as a sequence of individual gaze points (or by some other means) in order to highlight the internal structural

[5] Velichkovsky et al. [108] reports that fixations can vary between 100ms and up to several seconds in free viewing conditions (as seen in e.g. top-down visual processing). The fixation durations are distributed with a positive skew; the modal value is often in the same range as the saccadic reaction times: saccadic latencies (often seen in the context of salient stimuli) has a much narrower range, typically between 150–250ms.

[6] Confirmed to the author in private discussions with some eye tracking experts.

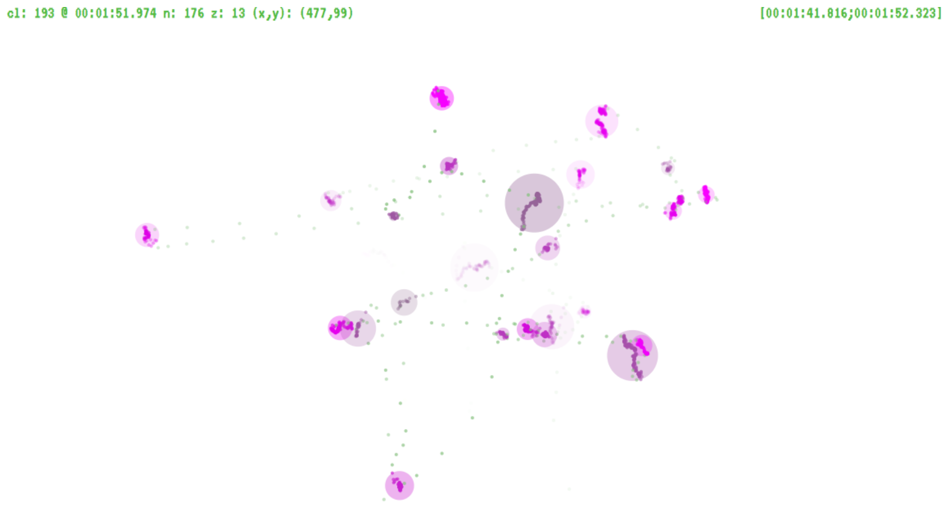


Figure 2.2: FIXATION CLUSTERS from a high-resolution, medium sampling-rate (250 Hz) eye tracking capture, showing both captured gaze points and derived clusters for a 10 second period, with the oldest parts fading away. The green dots are individual gaze points classified as saccades, and the pinkish dots are fixations that belong to clusters, which themselves are marked as shaded areas. The more dense a fixation is, the more coloured it also appear. The illustration is a screen-capture from a web-based tool for visualizing gaze points, fixation clusters and saccades, created by the author. The clustering used here is the DBSCAN-based algorithm also implemented by the author (see Section 2.5.2).

elements and their relationship. A related attempt is found in Paper G; using a similar Digital Fourier Transformation (DFT) may also be used to characterize the gaze points within one fixation.

Individual fixations can be characterized in many ways; at the most basic level they each have an average spatial position (x, y) and a temporal duration, t. Sometimes the dispersion is also reported in the form of an Root Mean Square (RMS), Variance or Standard Deviation value. Additional moments can be included, as can many other metrics. We again refer to [51] for an exhaustive list.

Fixations (or by extension AOI) can also be compared to others of the same kind and reported with metrics like number of fixations, fixation rates or proportion of fixations.

If a fixation is within an AOI, it is often counted as a *dwell* within that specific area, and the duration of the fixation contributes to its weight compared to other AOIs

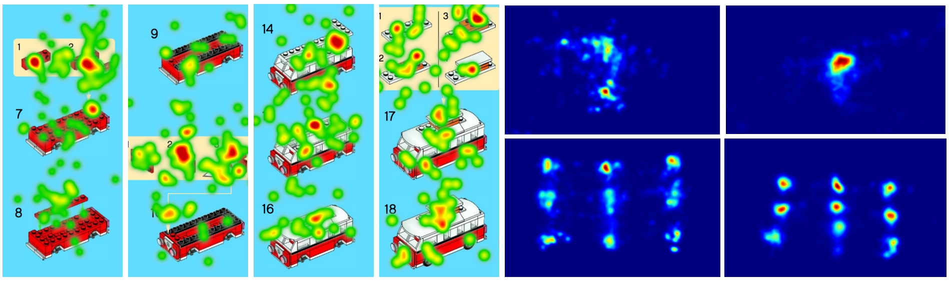


Figure 2.3: Traditional Heatmap (Left), as output from The Eye Tribe's EyeProof system, showing the density of fixations overlaid the stimuli for multiple users. Taken from the experiment described in Section 6.4. Fixation Density Maps (Right) that similarly show fixation densities for two different subjects, in two different situations. At the top, no specific stimuli is presented, so eye movements may reflect activity in the brain's Default Mode Networks. At the bottom, the subjects tend to salient stimuli at the presented locations. Apart from the apparent miscalibrations of the eye tracker, there are also significant differences between the signatures of the two persons. Generated by the author as an illustration included in [2].

as well as to the overall Fixation Density Map (FDM), which shows the aggregated spatial distribution of fixations:

2.3.1 Fixation Density Maps

Even though there is not a direct coupling of fixations to attention, it is often imagined that the sum of fixations somehow reflect areas we attend to: The central foveal area has the highest resolution, and we tend to put things into this central view when we want to study their details. As Duchowski [83] notes, "*most often*" we also tend to put our cognitive attention to this area.

For this reason a *Fixation Density Map* (FDM) or a *Heatmap* can be used (Fig. 2.3). It is also sometimes assumed in passive eye tracking studies, that even though there can be large individual differences in fixation patterns, when many subjects' data are summed, the most important AOIs becomes visible. Fig. 2.3 shows such an example (left), but adjacent to the data (right) are shown sample heatmaps from two different individuals looking at the same stimuli, where the large individual variation becomes very visible.

FDMs are also good for summarizing the results of many persons. Given the same scene, multiple FDM can simply be averaged in order to create a between-subject aggregate. Differences between two different FDM – in their original form or after a transformation – can also be calculated.

The Mannan similarity index [113, 114] was one of the first attempts at quantifying similarity between two sets of positions, using a least-square measure of similarity between two sets of fixations. It has a number of shortcomings, and other ideas have since been added to the toolbox. We will show a practically use hereof in Chapter 4, where 1-MIN, Mean Square Error (MSE), Kullback-Leibler Divergence (KLD) and Euclidean based metrics are used.

2.3.2 Fixation Scanpaths

Even though the duration of a fixation impacts the FDM, the order of the individual fixations disappear. A person looking first at object A and then at object B may appear very similar to a person inspecting the objects in the reverse order, when comparing their resulting FDM. In this case, a *fixation scanpath* can conveniently be used. Fig 2.4 shows an example of the visits paid by a person counting the number of screws needed to assemble an oven.

As an exploratory tool, scanpaths and similar visualizations can be helpful, but they can be difficult to compare with other scanpaths. Comparing the order of visited AOI or dividing the sequence into consecutive FDM that can then be compared has been proposed [51], as has been counting the number of permutations to make one sequence equal to another (the Levenshtein distance, [115]). Other ideas exist as well, and one we suggest might be worth exploring is the HDBSCAN based clustering algorithm proposed by Wilson et al. [116]. Although it originates in another context, it looks to us as if it could be a promising idea for comparing the similarity of two (sections of) scanpaths between visited AOI or a set of fixations – or even to characterize the internal structure of one fixation, as previously suggested.

2.4 Saccades

Saccades are fast movements of the eye from one fixation to the following.

Holmqvist [51] considers saccades to be the fastest movement the body is capable of producing, typically moving at angular velocities of 30–500°/s and lasting 30–80ms. The eye seldom moves in a straight line, but can move at different latencies and speeds in the horizontal and vertical dimensions, and may wobble a bit before coming to rest (sometimes referred to as a *glissade*, which is a small adjustment to the eye position, typically less than 1°).

If the initial target is somehow missed, e.g. due to an inaccurate long initial saccade, one or more *corrective saccades* may follow.

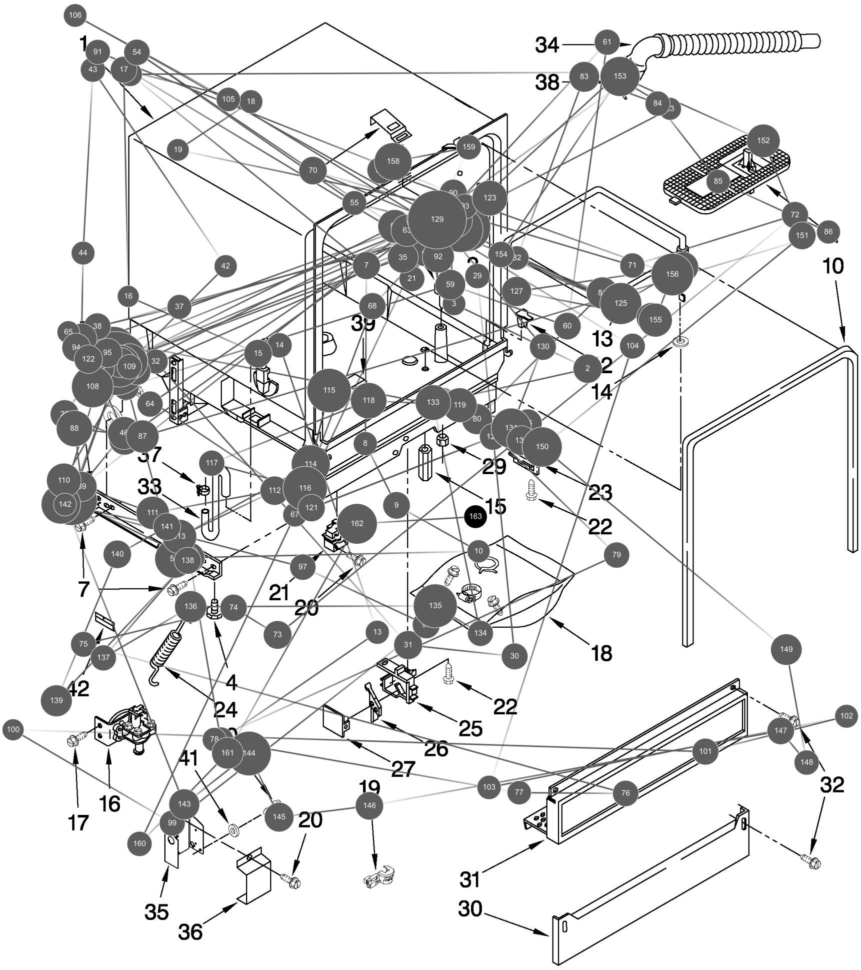


Figure 2.4: FIXATION SCANPATH: An adult subject trying to count the numbers of screws needed to assemble an oven made 163 fixations in 65 seconds to do this. Each circle shows a fixation of more than 125 ms, with the diameter proportional to the duration. Illustration from [5].

Visual processing of low spatial frequency content is suppressed during saccades [117, 118]. Had it not been for saccadic suppression, the world might not have remained as stable as we purport it to be!

Bahill and Troost [119] provide a taxonomy with a detailed hierarchical description of saccadic eye movements.

Individual saccades can be quantified in many different ways, such as by their direction, duration, amplitude, velocity, acceleration, trajectory and latency/delay from stimuli to onset or arrival time. They can also be quantified by their average rate.

2.5 Classification of Gaze Points as Fixations or Saccades

Even though as a concept fixations are not difficult to understand, comparing fixations from one experiment to another is not always easy. What should a minimum fixation length be so that short stops between inaccurate and corrective saccades are not counted as fixations? And if we insist on sticking to a minimum fixation length as a key parameter, how long should it then be?

Different manufacturers provide algorithms that classify fixations, often based on proprietary methods. Manufacturers may suggest different settings, sometimes dependent on what kind of study is being performed. Holmqvist [51] has a detailed discussion of these issues.

Further, all systems have different levels of noise and precision. Tole and Young [120], Behrens et al. [121], and Nyström and Holmqvist [122] suggest using systems that adapts to an estimated noise level. But this may again make comparison difficult if events are not classified in similar ways between recordings.

To illustrate some of the practical issues: The Eye Tribe tracker has a flag that is passed on to the application as part of each data frame, but given the constraints of having to provide near real-time information from the relatively low sample rate, it cannot flag the start of a fixation until it has seen some part of it. It therefore may miss the first data frames of the fixation. The accompanying Eye Proof tool has a better clustering (even though it is also not publicly documented how it works), but it is meant to work only within its own environment.

For that reason, for many of the experiments described herein using The Eye Tribe tracker, a modified DBSCAN [123] based algorithm has been deployed as a workable solution for the lower sample rate data.

Two main classes of classification methods will briefly be mentioned below. Salvucci and Goldberg [124] provides a taxonomy and an overview of representative methods, and Holmqvist [51] and Duchowski [83] also have good and more recent summaries of the field. Note that work very much seem to continue in this field, and more recent proposals using probabilistic methods like a Bayesian Mixture Model [125] may eventually prove to be more robust, requiring fewer prior assumption (like setting parameter values).

2.5.1 Velocity Based Methods

The basis for classification are the gaze point estimates (x_{t_i}, y_{t_i}) at the sampled intervals $[t_i, t_{i+1}, ...]$. From this, the (angular) velocity (v_{x,t_i}, v_{y,t_i}) $(°/s)$ as well as acceleration (a_{x,t_i}, a_{y,t_i}) $(°/s^2)$ can be derived. Simply calculating the velocity by subtracting gaze point estimates and dividing by the sample interval creates a very noisy signal. Based on ideas originally proposed by Tole and Young [120] (in a hard to come-by article), Duchowski [83] shows how using Finite Impulse Response (FIR) filters matching idealized saccade profiles improves saccade detection. It is likely that other (more advanced) models of the OPM system could also be used for this purpose, further improving event detection.

As it is required to estimate high velocities, a sufficiently high sampling rate[7] and low jitter is needed for the velocity or velocity/acceleration based methods to be effective in identifying saccades. For lower sampling rates, as seen in current mobile applications, density based methods may be more feasible at present:

2.5.2 Density Based (Clustering) Methods

Density based methods, or dispersion and duration based methods, work by locating fixations which are characterized by a high spatiotemporal density of gaze points.

A method often used is the I-DT method, which is outlined in e.g. [124, 126].

We have chosen, however, to use a modified version of DBSCAN [123], where the temporal order is included as well so that the presence of more points than a given specified noise parameter outside the current cluster will break it, and no additional points can then be added to that cluster. Although it is a parametric method (it needs a minimum density and a noise tolerance parameter), it has, for our purposes, the advantage of being able to pick up smooth pursuit and larger-than-normal microsaccades, as well as drift[8] caused by head movement, inside one cluster. This may make sense here, since visual processing continues in ways that may more resemble fixations than during saccades. A very similar parametric algorithm, based on the same idea, has recently been published by Li et al. [127].

[7] Holmqvist [51] set a lower limit for velocity and acceleration based methods around 200 Hz.

[8] The Eye Tribe tracker now has improved 3D tracking, but earlier versions benefited from the ability to maintain a fixation even if the head position moved forward or backwards towards the screen.

2.6 Smooth Pursuit

Smooth pursuit are slower movements of the eyes, typically 10–30°/s, in pursuit of an AOI that moves relative to the head pose.

Visual processing continues during smooth pursuits, which can only take place when there is an object to follow.

Smooth pursuit has been proposed in active gaze control systems by e.g. Lorenceau [128], Møllenbach et al. [102], and Esteves et al. [129], and may help to solve the Midas Touch problem in that it may be easier to avoid accidental activations compared to more naïve fixation based schemes.

We will not here consider smooth pursuit any further, but from an attentional point of view regard it in similar ways to normal fixations (and knowingly disregard a lot of additional complexity).

2.7 Blinks

Blinks are short periods where the eyelid covers the eye. The endogenous eyeblink is a cortically controlled event with very specific properties, and different from the reflexive and the voluntary closure of the eyelid. It is well described by Stern et al. [130].

Blinks are often seen as an nuisance in that it occludes the pupil and thus makes it impossible to track eye gaze, pupil size and other measures. Blinks also tend to take place shortly after a stimuli has been presented or reacted to, further confounding recording of the "more interesting" metrics: Siegle et al. [131] reports that blinks are more frequent after stimuli presentation or after an extended cognitive processing period. In our results, we have also seen this, and although it is not reported as a separate metric, Fig. 3.3 hints at blinks occurring after a response has been given: As an example, looking at the left eye congruent data (green dashed line), the SEM grows as a consequence of fewer data points representing the pupil size in the period of 1.3–1.8s after the initial stimuli, which corresponds well with an increased endogenous blink rate in this range.

Blinks may also carry meaning in themselves, and correlated with other factors, such as e.g. time on task [132, 133, 134, 135] or fatigue – as everyone that has tried a long night-drive or in heavy rain [136] may have encountered first-hand. A phenomena opposite to the blink-after-stimuli-presentation also exists; it is not uncommon to see periods of blink inhibition when subjects actively engage with a scene, and Shultz et al. [137] shows that in natural viewing conditions inhibition could be used as an index into how important a subject perceives a visual scene to be. The blink signature itself is also modulated by fatigue; Caffier et al. [135]

report an increase in blink duration from 202ms to 259ms comparing alert to drowsy conditions, and especially the reopening time seems affected from 138ms to 187ms. In addition, the rate of long-closure blinks increases significantly from 9% to 29%.

The most common aggregated measure is the blink rate, but blink velocity and duration of eyelid closure and opening times may also be of value. We suspect that, as these are not so easy to measure with optical eye tracking equipment and may require the use of Electrooculography (EOG) or EEG equipment, they appear less frequently in conventional eye tracking.

2.7.1 The Eye Dipole

The eye holds an electrical potential between the (positive) cornea and the (negative) retina. Moving the eyes causes not only Electromyography (EMG) activity but also a signal caused by moving this eye dipole. In addition, a blink also causes a strong signal when the eyelid closure changes the electrical fields around the eye dipole [138].

This is an unwanted signal in the EEG [139] and usually needs to be corrected for in order to get to the weaker EEG signals. It is particularly confounded by the fact that blinks occur more frequently just after cognitive processing, which may be a period of particular interest for ERP paradigms.

It also, however, means that there are other ways of measuring a blink and estimating blink velocities, and it further provides a way of synchronizing EEG and Eye Tracking (ET) recordings, as we will detail in Chapter 7. Fig. 2.5 shows an example of blink signatures recorded in the EEG modality.

2.8 Pupil Size

The main function of the pupil is conventionally ascribed to regulation of the amount of light that hits the retina, similar to a diaphragm in a photographic lens that controls the aperture stop. The *pupillary light reflex* [142] causes large variations depending on the scene luminance. This is controlled by the parasympathetic division of the autonomous nervous system: the Edinger-Westphal nucleus controls the rapid constriction phase, and the slower dilation is likely caused by a subsequent parasympathetic relaxation, as suggested by [143].

The pupil is, however, also influenced by cognitive processes, and is therefore of particular interest to us. This will be detailed in the following chapter.

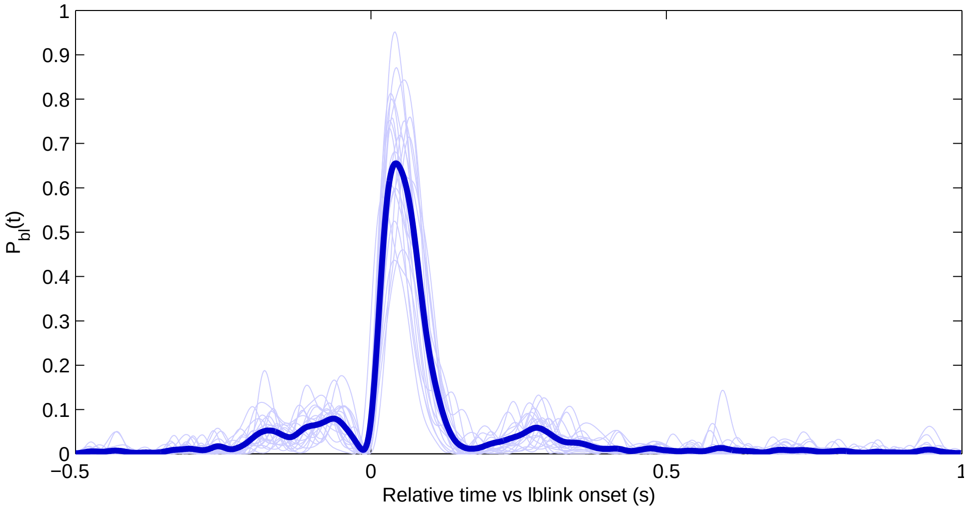


Figure 2.5: Typical blink signatures, here shown as normalized EEG power in the electrodes FPz and AFz for 21 blinks in [140], aligned to when an eye tracker sees the left eye closing (the overall mean value is shown darker). An electrical dipole between the positive cornea and the negative retina [141] generates an electrical signals in EEG and EOG recordings when the eye moves or blinks (which is significantly stronger than the normal EEG signal originating from the brain). A faster closing time is seen when comparing to the reopening time. Illustration from [1].

The basic pupil measure is the size[9]. Often (as with The Eye Tribe tracker) a linear measure is returned, which is essentially the diameter of the pupil in some arbitrary scale (sometimes just in pixels). It is also not uncommon to estimate the area of the pupil, possibly by measuring the diameter and correcting for any optical distortions when the eyes look away from the center of the eye tracker camera, which causes the pupil to appear more or less elliptical.

[9] The size sounds like a very simple concept – but there are different traditions here in the literature; some (like many of the cited earlier studies) use absolute size and express e.g. a pupil dilation in (fractions of a) millimeter in addition to sometimes also reporting the baseline pupil, whereas others only report a percentage of the baseline, like we generally have done in this thesis. The later has the advantage of making it easier to aggregate across different baselines in the varying conditions we generally see in our settings. Also, there has been no way of establishing a calibrated measurement until now, so our choice has largely been relative measurements out of convenience. Beatty and Lucero-Wagoner [144] deals in more details with the methods, and we would like to point out that as pupil dilations appear more constant in absolute size (i.e. mm) across different light conditions Bradshaw [145], it may however be the preferential choice.

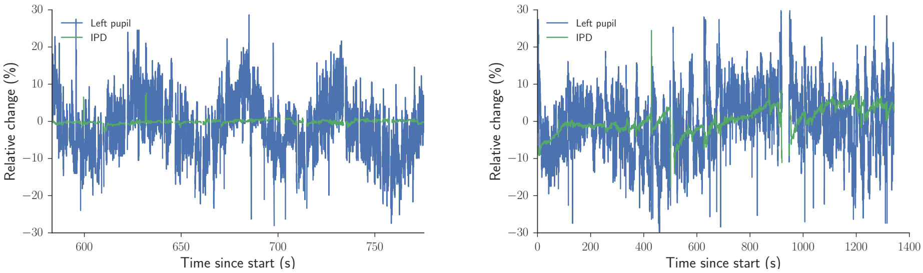


Figure 2.6: Naturally occurring variations in Pupil Size (Left) over part of an experimental session. A combination of task-evoked pupillary dilations and noise is visible, as is a slower varying modulation on a timescale of 30-60 seconds. The blue curve is the relative pupil size and the green the distance between the eyes, as reported by the eye tracker. Illustration from [3]. Head Movement (Right) over a full experimental session, consisting of 3 individual blocks of reaction time tests. It is likely that the subject moves closer and closer to the screen, becoming more and more tense within each block, but takes a deep breath and restarts at a longer distance when the next block starts. This makes it important to correct for head movements. Note that the slow modulation is also visible here, although confounded by the (uncorrected) head movement. Right illustration originates from the same study [3].

Some trackers like the EyeLink 1000 series [146] can utilize a paper target of known size fixated on the forehead to estimate the absolute measure (mm) of the pupil, independent of any accidental head movements. Others like the Tobii trackers [91] use an eye model to estimate the eye position and distance and also return an absolute measure.

The absolute size is, however, not a requirement for many types of measurements. Thus, in our case, we have resorted to reporting the pupil size as a percentage deviation from the mean or median value over a section of the experiment, and when required, the inter-pupil distance reported by the eye tracker (also in arbitrary units) has been used to compensate for head movements towards or away from the tracker. Whereas a normal standard- or z-score normalizes with respect to the mean and the standard deviation of the measurements, $z = (p-\mu)/\sigma$, we favour normalizing with respect to the mean (or median) only, $p_n = (p-\mu)/\mu$, as we are then more likely able to compare pupil dilations between experiments or subjects.

Fig. 2.6 shows two examples of pupil size variations over a 3 minute and a 23 minute period respectively. The variations are large and slowly occurring in the example on the left side, that also has a stable inter-pupil distance. To the right, however, variations in head-to-tracker distance causes variations in the inter-pupil

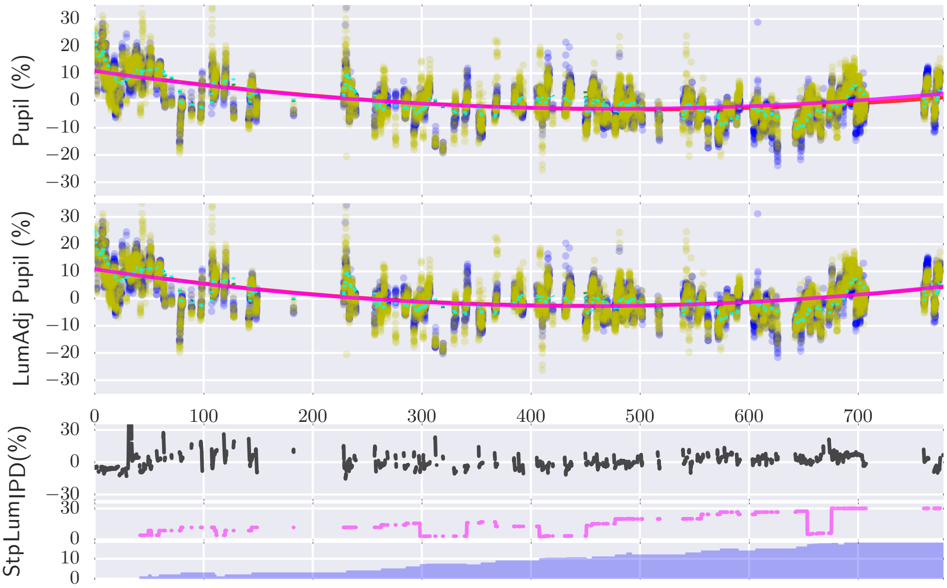


Figure 2.7: Pupil Size and Luminance over time. The (Top) graph shows the pupil size from left and right eye (blue and yellow) after correcting for changes in inter-pupil distance, with an exponentially averaged mean value (cyan) overlaid. The magenta lines show a 2nd order polynomial approximation to the curve. When the subject is looking away from the screen, no pupil data is recorded. The (Middle) graph shows the same pupil measures after compensating for screen luminance levels. The (Bottom) graph shows the reported inter-pupil distance vs the median value (black), the estimated luminance of the screen in magenta, and the progress through the steps of the assigned task(s). See study in Section 6.4 from where these data originated.

distance, and the pupil size needs to be corrected for some types of comparisons to be meaningful. In both cases, the luminance changes on the screen are small (estimated to be less than 3%) compared to the pupil size variations measured.

Fig. 2.7 shows another example of pupil size variations. The data here are from an experiment with children in a LEGO® assembly process. In this case, the screen luminance can be estimated, and a regression versus the pupil data (after correction for head movement) can estimate the effect of luminance changes,[10] and the residual can be used as an approximation of the non-luminance related pupil size changes.

[10] This is obviously only a first level approximation. A better method, as also suggested by Lang and Bradley [147, p. 445] might be to detect the unique signature of the pupil constriction caused by the pupillary light reflex, but due to the many other confounding factors, we settled for a first-level approximation.

In this particular case, we found the effect of luminance on the pupil size to vary considerably between subjects; in a few cases more than 35% of the variance could be explained by luminance, but values around 10% were more typical.

Other approaches exist for extracting the cognitively induced signal. In the body of work presented here we have in some cases divided tasks into smaller subtasks, and used within-subject comparisons between subtasks. Especially in cases where variations in conditions change slowly this may work well. As similar methodology, dividing tasks into subtasks and comparing between these, was also demonstrated by Wang et al. [148].

In general, pupil size measurements contain a lot of random variation. Duchowski [83] addresses some general issues of noise, and Klingner et al. [149] specifically deals with noise in pupil size measurements using remote eye trackers. Some may be caused by the measurement method, but the pupil itself is known to undergo small fluctuations all the time, as reported e.g. by Stark et al. [150]. Stark et al. find that it is cannot just be caused by a badly tuned servo-loop, and speculates that the $\pm 15\%$ fluctuations are small enough not to disturb the visual acuity of the eye, and it could be "*an example of economy in construction*", although it is not ruled out that it could have a specific role in the visual process. Compare also [150, Fig. 1] with Fig. 3.4.

It is, however, unlikely that the pupil size jumps e.g. 30% or 50% in a 16ms timespan, and therefore in many experiments herein, we have applied a hampel filter with a window size of approximately $\approx 2 \cdot 80$ms and a threshold of 3σ, to remove outliers. It is also important to filter out blink-affected periods. We find that pupil size estimates immediately before and after blinks deviate strongly from the neighbouring values; an effect likely caused by not correcting for the now distorted pupil. These periods also needs to be removed from the data. After this filtering, for many experiments we find a high degree of correlation between left and right eye pupils size measurements, typically with Pearson Correlation Coefficient R between 0.8 and 0.95. This indicates that there is a common causality for the variations.

We will mention also just briefly another pupil related measure that is sometimes used, the pupil (dilation) latency, which is the time it takes for the pupil to react to external stimuli (light or cognitively induced, as we will detail later).

2.9 Modulation by Cognitive Processes

Komogortsev et al. [151], Holland and Komogortsev [152], and Rigas and Komogortsev [153] have shown that there are several apparently unique physical and neurological characteristics of the OPM, as well as individual behavioural traits, that leads to unique signatures that can be derived from eye movements or simply from FDMs recorded under many different kinds of stimuli.

It is therefore not unexpected that individuals have very different measures under similar conditions, even though within-subject stability can be very high. Holmqvist [51] reports this for e.g saccadic amplitudes and fixation durations.

When comparing within-subject, variations can be caused by cognitive or affective processes. There are often multiple other factors that can influence eye related metrics, and it is important to disambiguate the signals in order to account for other factors as well, for instance by combining combining several metrics or modalities and/or by careful experimental design to separate other factors.

Some of the maybe more commonly proposed metrics when investigating cognitive or affective processes are:

- *saccadic amplitude (extent)* is impacted by cognitive workload:

  May et al. [154] finds reduced saccadic amplitude in both saliency-driven and spontaneous saccades when workload increases;

  Recarte and Nunes [155] reported spatial gaze concentration in a simulated driving scenario when “mental workload” increased.

- *pupil dilation* is also related to various cognitive processes; we will deal with aspects hereof in the following chapter, so just a summary here of some of the relations reported in the literature:

  Kahneman and Beatty [156] found pupil dilation to relate to task difficulty;

  Beatty [157] reviewed the body of evidence at the time of writing and concluded that phasic pupil dilations reflect cognitive processing load;

  Hyönä et al. [158] reports a correlation of pupil dilation and processing load caused by interpretation/language tasks;

  Einhäuser et al. [159] reports that perceptual selection and rivalry is reflected in pupil dilations;

  Privitera [160] finds pupil dilations relate to target detection,

  Preuschoff et al. [161] finds pupil dilations signals surprise (and so do we [4]);

  Siegle et al. [162] reports depressed individuals respond to emotional processing tasks with sustained pupil dilations;

  Gee et al. [163] found that an extended decision process results in a sustained pupil dilation with a smaller transient component when the decision is effectuated;

Kuchinsky et al. [164] finds pupil dilation correlates with indices of tonic alertness and inversely correlates with Default Mode Network (DMN) activity;

Kuipers and Thierry [165] finds a greater pupil dilation relates to a less negative N400 in the ERP pointing (within the experiment) to a correlation to matching meaning between pictures and words;

Mill et al. [166] reports increased pupil dilations in relation to unexpected memory recognition vs expected recognition.

Bradley and Lang [167] however warns that pupil diameter is not a straightforward indicator of previous occurrences in natural scenes;

Zekveld et al. [168] finds pupil dilations to be larger for degraded speech compared to normal or less degraded speech;

Siegle et al. [131] reported pupil dilations to correlate with cognitive load in digit sorting tasks;

Partala and Surakka [169] reported larger pupil dilations when listening to positive or negative highly arousing sounds compared to neutral sounds;

Stanners et al. [170] investigated the relation of arousal and cognitive factors on pupil dilation, and reported that cognitive demands take priority and reflect in pupil dilations, and arousal only becomes visible when the cognitive loads are small

Bradley et al. [171] further found emotionally arousing pictures to solicit a larger pupil size;

Geng et al. [172] found pupil diameter to reflect uncertainty in a visual selection and search process;

Laeng et al. [173] used a Stroop task to find larger pupil dilations for incongruent tasks compared to congruent tasks (and so did we in another task [3]);

Beatty [174] investigated the relation between phasic and tonic pupil dilations and found only phasic response to vary with vigilance performance;

Alnæs et al. [175] found pupil dilations to reflect activity in LC-NE and thereby related to how the brain allocates its processing resources; a similar finding was also reported by Murphy et al. [176];

Wilson et al. [177] found a correlation between pupil size and cognitive load, but no correlation between pupil size and fatigue (and speculates that it could be caused by increased cognitive load in the experiment);

Liao et al. [178] tested an auditory oddball scenario and found phasic pupil dilations as a response to some types of oddball sounds, dependent on the tested conditions;

Gabay et al. [179] found that more demanding discrimination tasks results in phasic dilations;

Holmqvist [51] sums up a range of research, and lists cognitive and affective processes related to mental workload, emotion, anticipation, drowsiness and fatigue as factors that impact the pupil size (and so do diabetes, age, pain and certain types of drugs);

finally, we should note that Eldar et al. [180] found that measures of the pupil size reflected the degree to which learning was focused on stimuli that the subjects were more predisposed to, and although the relationship here may be the other way (i.e. the pupil size, reflecting underlying neural activity, impacts the degree of learning), it may still be of importance also for this discussion.

- *saccadic speed* is reduced when the subject deactivates (i.e. due to fatigue, time on task, etc), as reported by [181]; Holmqvist [51] warns there can be confounding factors based on work by McGregor and Stern [182], but more recent surveys by Di Stasi et al. [183] suggests that peak velocity may be a good indicator for arousal.

- *blink rate* goes up when subjects are fatigued or have increased time-on-task (as mentioned previously), as reported by Stern et al. [132, 133], Schleicher et al. [134], and Caffier et al. [135].

  Holmqvist [51] also references literature that find blink rate to be impacted by mental workload and time of day.

- *fixation duration* was found by Dambacher and Kliegl [184] to correlate to the amplitude of the N400[11] in ERP reading paradigms (though see also [185]).

Even though this thesis focuses on pupil dilations, there is a large body of (not always easily comparable) research that investigates correlations between eye tracking measures and the cognitive or affective level of the subjects. We will, however, again have to refer to e.g. Holmqvist [51] that devote large sections to these subjects. We will instead now narrow our focus more on pupil dilations caused by cognitive processes.

[11] Known to relate to semantic processing; see Section 1.2.1.

Chapter 3

Attention, Arousal and Pupil Size

3.1 Attention and Arousal

Attention and Arousal are frequently used and helpful psychological and physiological concepts. They may, however, not always be associated with a precise definition. In our context, we perceive attention to be the process of allocating resources to relevant stimuli. James [186] puts it like this:[1]

> *Every one knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Localization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others, and is a condition which has a real opposite in the confused, dazed, scatter-brained state which in French is called distraction, and Zerstreutheit in German.*

Duchowski [83] has a nice historical overview of how we have perceived visual attention, from the "where" of von Helmholtz [187], the "what" of James [186] and the "how" of Gibson [188] via Posner's [103] "spotlight", that separated the foveal focus from the attentional focus, onto more modern elaborations building here upon.

3.1.1 Attention Networks

Coull [38] attempts to divide attention into several subprocesses depending on how one attends to stimuli, but this appears not to be linked to an anatomical view of the brain's networks. We prefer to relate attention more to functions of specific networks and processes in the brain, based on insights gained during the last decades, and it

[1] ... and everyone that try to define attention appear to include the very same quote!

turns out that Posner's concepts has evolved into a neurophysiological model. This was originally proposed by Posner and Petersen [189] and updated again by Petersen and Posner [190] with new insights from the 20 years in-between. The model has three premises:

- It anatomically separates the attention system from the processing systems, which take care of dealing with the input and generating the output.
- Attention uses different anatomical networks.
- Each network has a unique role that can also be specified in cognitive terms.

The three elements that are suggested are the networks of:

- *Alerting*, which functions as a "warning bell" preparing the subject for a stimuli that may require a response.
- *Orienting*, which helps to focus sensory input to the location or modality from where a stimuli is expected.
- *Conflict Resolution* or *Executive Control* (which may turn out to involve two anatomically separate networks), playing an important role in what is also called "focal attention", selecting between multiple inputs or cues, directing focus (and the ability to respond) to the relevant place.

Based on the concepts of *goal-directed* (top-down) and *stimulus-driven* (bottom-up) attention, Corbetta and Shulman [191] and Shulman and Corbetta [192] proposes that two segregated anatomical networks are involved in these attentional processes, and further ties this into the model proposed by Posner and Petersen, [190], in particular how it affect the orienting: the (dorsal) top-down visiospatial and the (ventral) bottom-up reorienting networks. This is mentioned here as it may impact how we should treat eye movements: if two different networks are involved between e.g. search-strategies vs attending to more salient stimuli, we may also see different modulations of the movements. See also [193].

Arousal is broadly understood as the level of being "awake"; i.e. somewhere between drowsiness and panic at the two extremes. With the basis in Petersen and Posner [190] and Coull [38], we suggest arousal to be understood as a level of vigilance which modulates the attentional system, with the goal of being able to perform sufficiently during expected tasks.

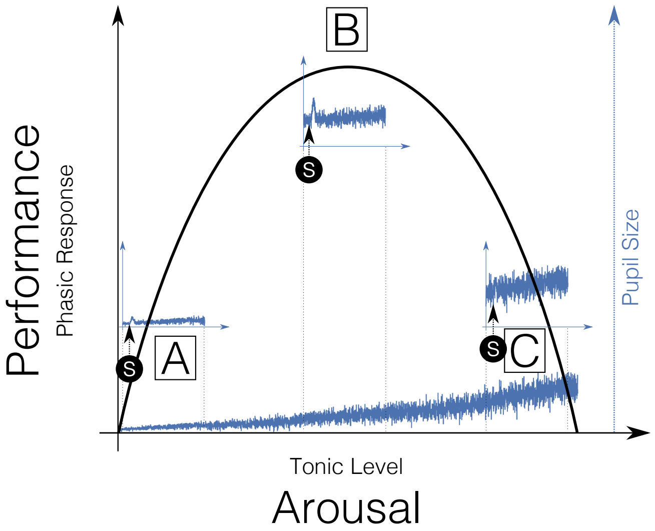


Figure 3.1: Arousal and Attention. Illustration from [3]. This graph is a synthesis borrowing from especially Aston-Jones et al. [194] and Joshi et al. [195]: As the "arousal" (x-axis) increases from drowsiness [A] via exploitation with highly focused attention and peaking task-specific performance [B] to a state of highly distractible explorative behaviour [C], the LC-NE tonic activations increase and the pupil dilates, as illustrated by the blue curve. The task-related performance (y-axis) in response to relevant stimuli increases from [A] to [B], but declines again hereafter as the tonic level increases. At the peak [B], relevant stimuli results in the largest phasic pupil response. At [A] and [C], task specific pupil responses are very small or absent. The small blue graphs illustrate the resulting visible pupil dilation that combine the output from both phasic and tonic processes.

3.1.2 The Role of Locus Coeruleus

Aston-Jones and Cohen [196] have investigated the role of the LC-NE system in attention[2] and arousal. They suggest that LC cells can function in two (overlapping)

[2] As additional examples of some of the links between attention (networks) and the LC-NE system we refer to: Petersen and Posner [190] who links the alerting network to Norepinephrine (Noradrenadlin(e)) activations; Bouret and Sara [197] who sums up the Norepinephrine (a.k.a. Noradrenadlin(e)) (NE) system as having an important role in regulating attention, learning and memory as well as suggesting a role of Locus Coeruleus (LC) in regulating attention and promoting behavioural flexibility; Aston-Jones et al. [194] who mentions that the LC-NE system is coupled to brain areas linked to attention, and give as examples the parietal cortex (involved in the orienting network [190]) as well as pulvinar nucleus (also linked to the orienting network [189]) and superior colliculus; and to Kuchinsky et al. [164] who links attentional processes and the Cingulo-Opercular Neural (CON) network to LC activations. In another recent work, which incidentally builds on similar ideas to what is analysed in our work, Geva et al. [198], with an important commentary given by Ambrosini et al. [199], found that the LC-NE system is involved in all three networks, as evidenced by two different pupillary reactions: an early component *Pa* and a

modes: *tonic mode* and *phasic mode*. With little activity in LC-NE, subjects will appear drowsy. As the tonic level increases, so does the ability to perform, until an optimal level, with regards to the specific tasks, is reached. When responding to stimuli, phasic activity is seen, and at peak performance, there is a strong phasic response to relevant stimuli. This is the point of optimally focused, or selective, attention, and grants the subject the ability to exploit the present scenario. However, if the tonic level increases further, the ability to respond to the ongoing task diminishes, and task-related performance drops[3], although at the same time there can be an increased response to other stimuli. At the higher level of tonic activity, phasic response is again at a minimum and the mode of the subject becomes predominantly explorative.

This resembles the classical Yerkes-Dodson relationship [200], and may in fact be an important element in explaining the observed behaviour. See Fig. 3.1 for a graphical explanation, with A corresponding to drowsiness, B to optimal performance[4] and C to a state of very distractible and explorative behaviour or even some levels of panic. Jepma and Nieuwenhuis [202], using a gambling paradigm, has found empirical evidence that relate pupil size to the choice of explorative or exploitative behaviour, consistent with the theory that LC-NE is having an important role in regulating the explore–exploit trade–off. So rather that trying to explain the elusive concept of arousal we will adopt a slightly more mechanistic view here, and focus more on the tangible anatomical function of the LC-NE system.

The resulting LC-NE activity at any given time can be thought of as regulating the overall gain in the innervated brain systems, as the neuromodulatory role of releasing NE is to increase gain in the cortical circuitry involved in the tasks [196]. In normal conditions, the balance between the two modes is governed by a desire to optimize utility, i.e. to seek the required level of performance with the minimum amount of resources [196]. This system is aptly called the *adaptive gain* theory.

later component, *Pe* that is modulated by incongruency, which presumably involves the executive control network.

[3] An analogy, however awkward it may be, is to think of a guitar amplifier. At low gain settings, the signal is not loud enough, and gain needs to be increased for the sound to become louder. Increasingly, noise will also grow, and at a certain point the signal-to-noise ratio is optimal. If, however, the gain is further increased, distortion will become more and more prominent and random noise will also grow. At some point the original signal from the guitar can hardly be recognized and all that is left is internal noise from the amplifier.

[4] Note that performance is always in relation to a specific task, and different tasks may cause different ’curves’ describing the relationship in the diagram. An example, shown by Kahneman [201], a simple task may peak at a higher arousal level whereas a more complex task, requiring a more focused performance, peaks earlier.

In the context of a specific task, a too high tonic level leads to decreased performance. This may not be optimal seen in isolation, but Aston-Jones and Cohen [196] suggests that it may nevertheless serve a wider role of disengaging performance from the specific task, allowing us to sample other things around, and possibly seek alternatives that could be more rewarding. Constantly operating at a too high tonic level is not beneficial, though, as is likely seen with some attention deficit disorders.

The LC-NE system thus is, or becomes, involved in many cognitive processes, of which some has already been listed in previous chapters. In addition, Chmielewski et al. [203] found the LC-NE system to have an important role in response inhibition and how it is modulated by working memory load. We should also note that Corbetta and Shulman [191] mentions that the LC-NE system not only plays a role in vigilance and arousal, but also has a role in selective attention and especially when it come to salient unexpected stimuli (see also [4], where strong phasic responses are caused by errors or surprise, and e.g. [161, 172]). And further, Eldar et al. [180] reports, as partly mentioned earlier, that the brain-wide gain level correlates with learning: An increased tonic level (gain level) helps to strengthen already established neural connections in favour of weaker ones, thus reinforcing these already strong network – even if they are not related to the immediate task, as long as stimuli is salient enough. Conversely, lower gain states, where stronger phasic responses are seen, facilitates flexibility when surprising or unexpected stimuli appear.

3.1.3 Regulation of the Adaptive Gain System?

Lastly, we couldn't help but notice that not only do we very often see slow pupil size variations in our own recordings, with cycles of 30s to several minutes, as exemplified by Fig. 2.6, but also note it elsewhere, such as in Aston-Jones and Cohen [196, Fig. 7] and partly in Joshi et al. [195, Fig. 2]. Stark et al. [150] and Stark and Sherman [204] concludes that the variations are not caused by the closed-loop properties (of the light reflex he investigated), and pointed out that the cause for variations are induced from outside the loop and must be of central origin. Klingner et al. [149] found that correlation between the pupil size of the two eyes is very high at low frequencies but drops of sharply to below R=0.3 above 5Hz, and also attributes low frequency variations to a central source. We would therefore hypothesize that 1) the source of these variations *might* be the adaptive gain system, and that 2) what is seen could be related to how the system regulates itself, adapting to tasks and optimizing utility, thus balancing required (or anticipated) effort and allocation of resources, and that 3) there could therefore potentially be clues to the present state of this system that could be told on the basis of monitoring these possible tonic level variations.

We should caution, though, that even if Privitera and Stark [205] also conclude that the source of the noise must be in the brain in a part that is common to both eyes, it was also reported that the pupillary noise has the largest variance for mid-range pupil sizes and diminishes as the pupil becomes larger or smaller. This would for instance be consistent with the adaptive gain theory if the source of the ’noise’ is either phasic activations to exogenous or endogenous stimuli or some other activity that causes a response from the LC-NE system, but it may not be entirely consistent if the ’noise’ is internal to the adaptive gain system. It is, however, not clear to us if the concept of ’noise’ here also includes the slower variations found, and more work is needed to understand this system.

Currently we only have observations that so far appear uncorrelated to any independent and otherwise measurable variables, and we cannot therefore offer a more coherent theory at this point.

3.2 Pupil Size

Modern pupillometry [144], [170], [206] probably started with an experiment performed by Hess and Polt [207] in 1964. In a famous experiment using mental multiplication problems, they demonstrated that not only (as shown earlier [208]) did the pupil dilate when subjects had an emotional interest in presented material, it also dilated in relation to the cognitive load of the multiplication tasks. Kahneman and Beatty [156] confirmed this finding and did further studies, and suggested on the basis hereof that pupillary responses functions as an index of memory and processing load.

Often such responses cognitively induces responses are smaller than light induced variations [144, 142], in typical conditions they amount to maybe 15% of the baseline pupil size, i.e. around 0.5mm of linear variation where light induced variations can be significantly larger, of the order of several mm[5]. For this reason it is common to present the same stimuli repeatedly, and calculate a stimuli-time-locked average, similar to how it is conventionally done with ERP[6]. Such responses are

[5] It might be worth noting that the pupil constricts rapidly when the pupillary light reflex kicks in, with latencies on the order of 200–400ms and full effect in around 1s [209, 142] but dilates somewhat more slowly hereafter when light levels allow for it. Dilations caused by cognitive load also kicks in relatively swiftly, on the order of maybe 500ms after neuronal activity [195], and in some of the paradigms used by the author, it peaks 1-2s after stimuli onset, sometimes with a somewhat slower constriction. See Fig. 3.3 for an example.

[6] If measuring the phasic Task-Evoked Pupillary Response (Task-Evoked Pupillary Response) in eye tracking in some sense resembles the methodologies used for ERP paradigms in EEG, the equivalent of analysing the tonic level in eye tracking might be not unlike continuous EEG paradigms, looking for instance at the spectral power distribution.

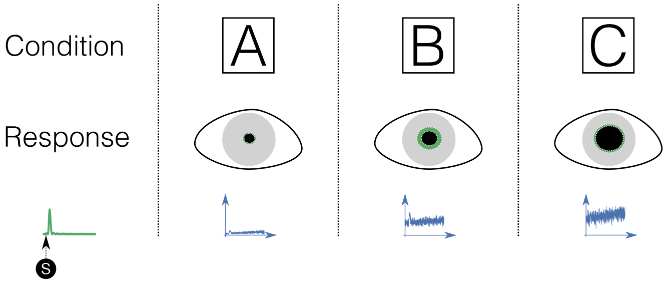


Figure 3.2: Arousal and Pupil Size: Task-Evoked Pupillary Dilations. In each of the 3 conditions corresponding to [A] drowsiness, [B] focused attention (exploitative mode) and [C] scanning (or explorative mode) the baseline pupil size is as illustrated in black. A stimuli attended to will solicit a pupillary response of the magnitude illustrated by the green area. The blue curves show the combined baseline pupil size and the task-evoked response. The illustration is not to scale and is meant to illustrate the concepts in a simplified way. Illustration from [3].

called Task-Evoked Pupillary Responses, likely coined so by Beatty [157] and Ahern and Beatty [210].

Aston-Jones and Cohen [196] correlate LC-NE activity to pupil diameter, and Murphy et al. [176] later report a positive correlation between fMRI BOLD activity in (specific parts of) the LC-NE system and the pupil diameter. Further, based on invasive testing in monkeys, Joshi et al. [195] have suggested that LC-NE mediated activity is related to pupil diameter changes[7]. Such activations in LC-NE correlate with pupillary responses with fast onsets and lasting on a timescale of seconds, consistent with what is seen for cognitively induced pupillary responses [212]. Varazzani et al. [213], while looking at the roles of NE and Dopamine in a reward/effort scheme in monkeys, also reported correlation between pupil dilation, LC-NE activity and effort/energization of behaviour (see also the commentary on the relation to uncertainty by Ang et al. [214]).

Fig. 3.2 tries to illustrate how the pupil respond in each of the three conditions shown in Fig. 3.1, corresponding to drowsiness, maximum performance and scanning behaviour. The base pupil size corresponding to the tonic activity at the condition

[7] It is speculated that the pupil diameter and LC activations may have a common causal factor, which for external events could be the paragigantocellularis nucleus (PGi) and/or, with reference to Wang and Munoz [211], that pupil size changes caused by attention related aspects including orienting to salient stimuli are related to a projection from superior colliculus (SC) to the pupillary control circuit. These theories are not mutually exclusive, and the suggestion remains that LC-NE mediated activity (in some parts of the brain) is reflected in pupil diameter changes that are not related to luminance.

is shown in black, and the green area denotes the Task-Evoked Pupillary Response driven by the phasic activations.

3.3 Estimating the LC-NE activity

It thus appears that the pupil can be a window into the mind – or to be more specific: we can estimate the activity of LC-NE based on pupil dilations.

We will now look at how pupil dilations can help us to estimate LC-NE activity, but we need remind ourselves that LC-NE activity is not the only regulatory mechanism that impacts the pupil size, and that it might be a good idea to "*disambiguate*" such signals (as suggested by Holmqvist [51]) using other modalities, such as HRV. Further, Corbetta and Shulman [191] suggest that the baseline/tonic level is representative (a "*proxy*") of the LC-NE system gain level, but cautions that comparing essentially non-normalized values between-subjects is very difficult, and that relative phasic responses are better suited for between-subject comparisons.

Different proposals have appeared for how to essentially measure LC-NE activity, or "quantify levels of attention or workload" in more fuzzy terms. Armed with the knowledge above, we can divide these into methods of estimating:

- The Combined Levels of Phasic and Tonic Activity
- The Phasic Response
- The Tonic Level

3.3.1 Combined Levels of Phasic and Tonic Activity

As long as the impact of other factors, in particular variations in luminance, can be accounted for, measuring the combined level of activity in LC-NE is straightforward. We simply measure the pupil size during our experiment, average it over some suitable periods to remove some of the inherent "noise", and compare it to a baseline that must have been established in some reference condition[8]. This appears similar to the *global processing load* or the *pupillometric estimate of mental load* suggested by Hyönä et al. [158], the *percentage change in pupil size* of Iqbal et al. [215] or the *mean pupil diameter change* found in [216].

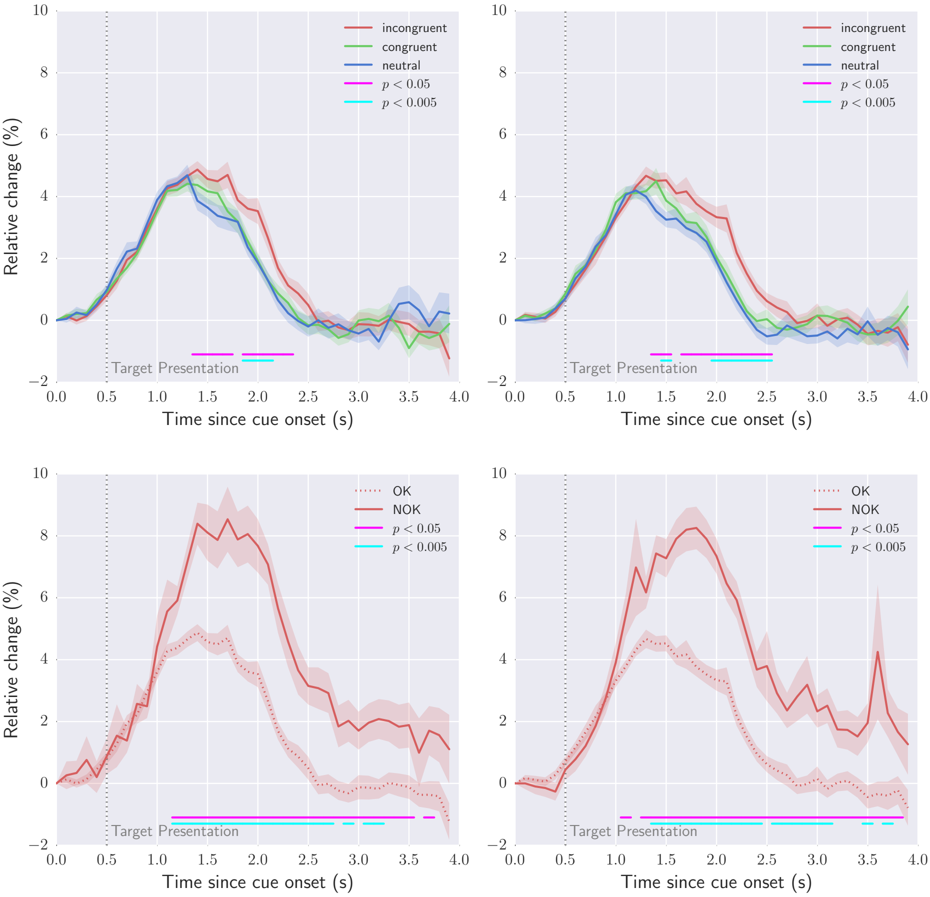


Figure 3.3: (Top) Task-Evoked Pupillary Responses. The graphs show the difference between three congruency conditions (left and right eye). The incongruent response may have a larger magnitude and has a longer temporal duration compared to the two other responses. (Bottom) “surprise” pupil dilations vs correct responses for left and right eye, in the incongruent condition. Note how significantly stronger the pupil dilation to an incorrect response (surprise reaction, labelled NOK) appear; almost twice as large as for correct responses. Illustrations partly from [4].

3.3.2 Stimuli-driven Paradigms

If the timing of the stimuli is known, we can use the methods proposed above by Beatty [157] and Ahern and Beatty [210] to measure the Task-Evoked Pupillary Response, possibly averaging a number of similarly conditioned responses to improve the signal-to-noise ratio. It may be important to bear in mind that slow oscillations often are seen in the tonic level, so an appropriate baseline needs to be selected for the intended purpose, representative of the conditions immediately before the response is invoked. Fig. 3.3 shows an example of measuring such Task-Evoked Pupillary Response. The OK conditions are each averaged over approximately 1500 trials and the NOK conditions around 150. These are averaged over different subjects, so any individual differences will tend to disappear. Note also that different stimuli may invoke responses of different magnitude and/or extent.

We can also measure the periods where no phasic responses are expected, to find the mean tonic level and, for instance, see how it evolves over time or relates to the task performance otherwise; one example of this is seen in Hyönä et al. [158]. An alternative to consider is to use an indirect measurement of the tonic level by estimating the gain level based on comparing the magnitude of a response to a given stimuli to previous within-session responses to the same stimuli. This might give an indication of whether the gain has increased or decreased, and thereby indirectly an indication of the tonic level. As the relationship between gain and tonic level is not monotonous, this method may only work within certain confined ranges. We have yet to see this used in practice, though.

3.3.3 Unconstrained Tasks

In unconstrained tasks, where the stimuli activating responses are not well timed, unknown, or even endogenous to the subjects, we need other methods to differentiate the phasic and tonic levels. Marshall [217] and Marshall [218] has proposed a (patented) method using wavelets tuned to the expected pupillary response to calculate an Index of Cognitive Activity (ICA) that appears to count the amount of phasic activations. As long as the tonic level is between [A] and [B], and not too close to [A], this may work well. If phasic activations start to diminish, they may be harder to tell from the tonic level, and the ICA may no longer reflect the perceived cognitive load or focus in the situation.

Lew et al. [219] also proposes the use of wavelets to estimate the power spectrum for pupil dilations, as do Pinzon-Morales and Hirata [220] and Pedrotti et al. [221]

[8] An alternative to a baseline comparison is to divide the recording into suitable subtasks (as we do in several of our studies here) and compare in-between the subtasks.

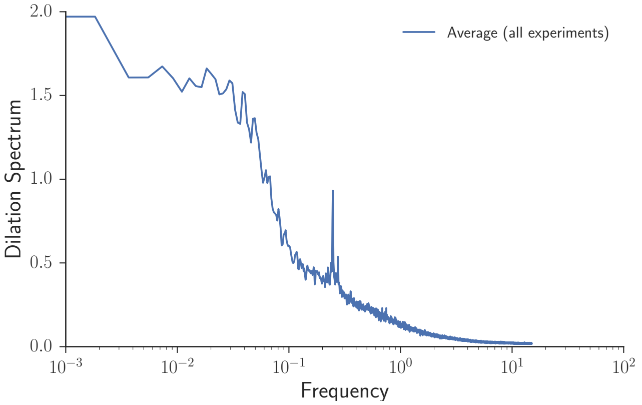


Figure 3.4: PUPIL POWER SPECTRAL DENSITY (after filtering outliers and blinks and using a blackman window, using Welch' method), averaged over 17 sessions (each approximately 25 minutes long), for one subject. The stimuli timing is once every 4 second, leading to the peak at 0.25 Hz. Data originally from [3].

whereas Lew et al. [222] (somewhat inconclusively) earlier had proposed the use of short-time Fourier Spectra and Nakayama and Shimizu [223] also uses a frequency domain based comparison; in the later case finding different power spectra in the ranges of 0.1–0.5Hz and 1.6–3.5 Hz – likely a result of phasic dilations and subsequent constrictions – depending on the estimated workload. Incidently, Lüdtke et al. [224] earlier proposed the use of frequency domain analysis of pupil size, and found differences between sleepy and alert states, and further suggested the use of a *pupillary unrest index*.

One example of the power spectrum of the pupil is found in Fig. 3.4. In the particular scenario, stimuli occur every 4 seconds, and thus a very sharp peak is seen from the phasic dilations caused by this stimuli.

Depending on the experimental paradigm, one alternatives to consider is whether fixation aligned averaging, as proposed by Klingner [225], could detect phasic responses. Another idea that has initially been suggested by Bradley et al. [171] and Lang and Bradley [147, p. 445] is to detect the rapid pupil constrictions associated with the pupillary light reflex – unlike those from cognitive processes, which cause rapid pupil dilations – and account for such changes in the experimental setup or when analyzing data. This could possibly be combined with fixation alignment, although we have not yet any experiment evidence for such a setup.

3.3.4 Temporal Resolution

A final concern to mention is the temporal resolution. Pupil dilations take place in the 1 Hz range[9], whereas the underlying LC-NE processes operate at a higher speed. Hoeks and Levelt [226] used deconvolution to model the attentionally driven pupil dilation system and derived the impulse response of the pupil, and determined the system to be linear, with additive effects of the neuronal pulses. Wierda et al. [227] further developed this idea, using iterative convolution of attentional pulses to match the measured signal in order to derive at the original neuronal signal; this method may possibly be more robust when presented to a recorded “noisy” signal. In a very recent work, Gollan and Ferscha [228], further suggests a modification of the proposal by Wierda et al. and specifically intents this to be useful in online, low computationally resourced environments.

[9] Stark and Sherman [204] reports a significant loss of gain in the pupil system around and above 2 Hz, at least when it comes to the pupillary light reflex

Chapter 4

Individual Differences

People are different... and so are their biometrics, which exactly for that reason becomes useful for identification and verification purposes.

This is true also for many metrics of the eye. The most well-known is probably iris recognition and identification, which surfaced in practical applications in the 1990s [229, 230, 231] and now is moving into mobile consumer devices. Hill [232] patented retina scanning, which relies on the individual characteristics of the retina in a similar way as for the iris, although it has a lower collectability and permanence [233].

Kasprowski and Ober [234] suggested using eye movements solicited by highly salient jumping points, and used as feature vector “reaction time” and “stabilization time”, which in our terminology would include characteristics such as the saccadic latency, saccadic velocity and some measure that includes inaccuracies and corrective saccades. However, as we have seen, with the more elaborate OPM proposed by [151, 235], we are able to model eye movements to a significantly higher degree. It is possible to estimate the individual 18 parameters[1] of the OPM based on the response of the eye when stimulated by visual patterns. The identification accuracy was over 90% Area-Under-Curve (AUC) scores in the best cases tested with proper stimulation.

Bednarik et al. [236] found it was possible to use pupil dilation measurements as a method of identification, even with a low sample rate system (50Hz). The experimental setup used (for this specific part) required participants to view a static fixation cross (which would probably not involve a lot of cognitive load) while the naturally occurring variations of the pupil diameter were recorded. Data was analysed in several domains, including the frequency spectrum. Hence, we assume that what may have been measured and compared was essentially variations of

[1] The parameters are constants describing e.g. the elasticity, length-tension relationship, force-velocity relationship, viscosity, mass, (de)activation timing, neural pulse characteristics, etc [70, 235].

the tonic level LC-NE, shaped and filtered by the characteristics of the individual eyes. Using this method, with a small set of people (N=12), a cross validation gave above-chance level identification results.

Based on our experience, it may additionally be possible to detect differences in microsaccadic patterns. In discussions with others in the community there appear to be anecdotal evidence that researchers can identify individuals based on the characteristics of their microsaccades within fixations. We saw a similar trends in a small group of test subjects where a high-resolution tracker was used: Not only were there visible differences in the dispersion (which could have been an artifact of the eye tracker used), there also appeared to be a preferential directions (up-down, left-right, ...) for each individual. This, however, is not unexpected, given our understanding of the OPM.

Some of the methods mentioned require high-resolution and accurate equipment. For example, to estimate saccadic speeds, an eye tracker with a very stable sample rate and low jitter is required, otherwise accuracy will be lost. We have instead focused at how some of these results translate into a mobile, low-cost and low-accuracy setting.

In order to perform the experiments, a Python interface to the eye tracker was developed by the author [10].

4.1 Lessons Learned and Questions to Ask

A lesson learnt from these studies is that it is important to account for individual differences when using eye tracking data that involves comparison between subjects. Even though aggregating the results of many user tend to average out individual differences, comparisons between groups can be misleading due to individual differences.

A question that it raises is of how stable these individual signatures are over time? We may assume that the parameters of the OPM remain relatively stable over time, since these are grounded in dimensional and physiological characteristics of the eye. But some of the metrics may be modulated by higher level processes, as may be the case for pupil size variations, and these may change over time and condition. That is the topic of Papers [B] and [G].

4.2 Summary of Thinking Outside of the Box... [B]

To answer parts of the previous question, we recorded a series of eye movements in a simple scenario: we presented a uniformly coloured screen for 4 seconds, followed by a contrasting target that appeared at 9 different position in semi-random order, each time for a few seconds. The colour scheme changed between different colours. The experiment was repeated approximately twice a day over a week, to see if there were any variations in simple metrics we could detect.

We did not detect any measurable differences caused by the different colour schemes.

We did, however, see very different fixation patterns between the subjects. Fig. 4.1 shows Fixation Density Maps (FDM) for each experiment and subject when the subjects look at a blank screen. As the subjects wait for the next stimuli to appear, it may be a good strategy to keep the eyes towards the middle of the screen, as the chances of quickly detecting the target – directly or with a peripheral view – is then probably largest. However, the fixation density maps we observe when the subjects attend to the blank baseline screen, throughout the experiments, could indicate that individual signatures in the gaze patterns are specific to each subject and preserved across sessions, even in a state likely to involve the default mode network in the brain

Not only are there differences in dispersion between subjects; there is also a tendency for especially subject A to look towards the left side of the middle. Unfortunately we did not analyse the pupil size variations, but we should have done so, to compare with Bednarik et al. [236].

Further, when attending to the stimuli we see a difference between the subjects, as evidence by Fig. 4.2. These are large differences; a human observer can pick almost any of the plots and tell whether it is from subject A or subject B.

We were not able to detect any statistically significant variation in the dispersion of the FDMs that we had hoped to relate to e.g. time of the day or other subjective factors.

Additional studies (done by some of our volunteering students) also showed similar individual differences with the paradigm, as illustrated by Fig. 4.3.

In addition to this analysis in the spatial domain, we also looked at variations in fixation time. Fig 4.4 shows how it varies throughout a week for two conditions (and we also see differences between subjects).

Even if the material here from a statistical point of view is quite small, it provided an insight into a few unexpected phenomena. Also, after having done the experiments, another paper by Rigas and Komogortsev [153] appeared that also looked at FDM. Even if the paradigms here were different, similar results appeared, and it caused an interest in repeating the experiment also to see how stable the signatures were over a longer period:

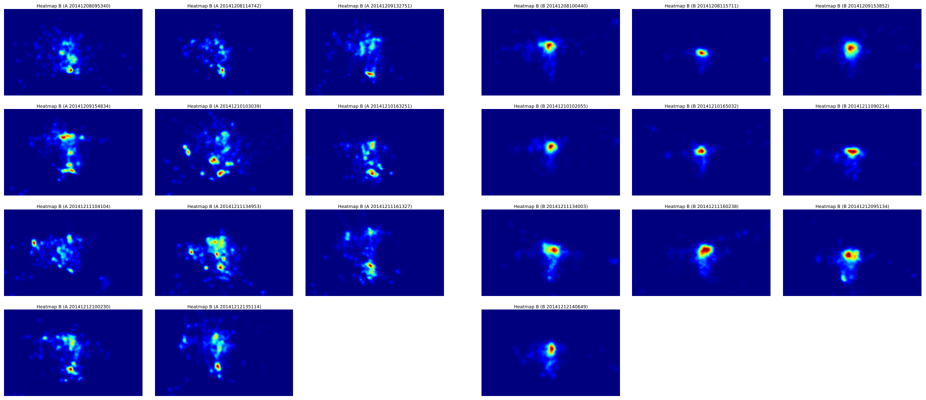


Figure 4.1: ASSUMED DEFAULT MODE NETWORK FIXATION DENSITY MAPS for two subjects A (LEFT) and B (RIGHT). Illustration from [2].

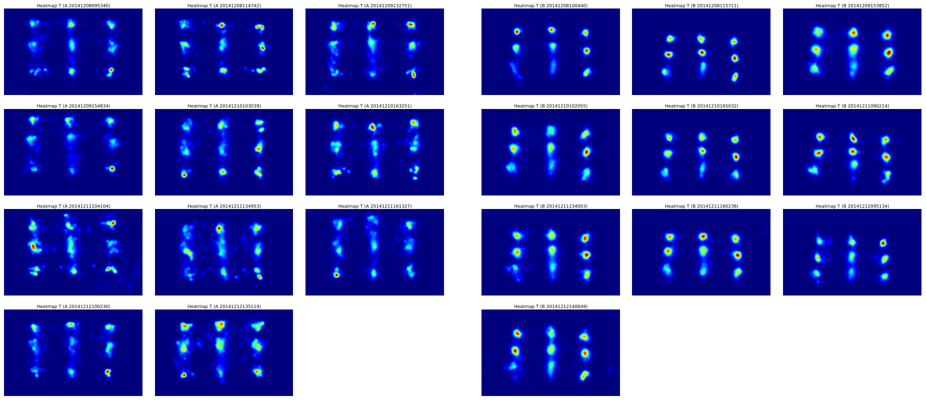


Figure 4.2: SALIENCY-DRIVEN FIXATION DENSITY MAPS for two subjects A (LEFT) and B (RIGHT). Illustration from [2].

4.3 Summary of The Blank Stare [G]

The previous experiments were repeated again 3 months later, and inspired by Rigas and Komogortsev [153] a more formal comparisons of the FDM (in particular) was done.

In addition to comparing the FDM in similar ways to what was done by Rigas and Komogortsev, a DFT was also applied in the spatial domain. This was done to be more invariant of translation errors in the calibration, and also with an idea of capturing "patterns" of the dispersion of the fixations. See Fig. 4.5.

The analysis showed a remarkable stability in the FDM within-subject. Depending on the metric chosen, we could characterize any fixation density map as belonging

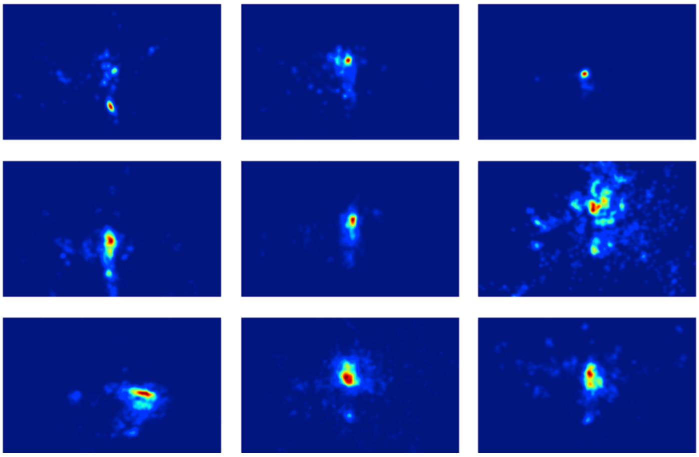


Figure 4.3: Additional Default Mode Network Fixation Density Maps for 9 additional subjects. Illustration from a student paper by Andersen [237].

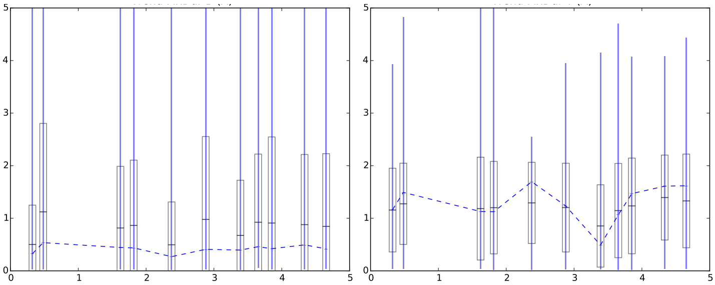


Figure 4.4: Variations in Fixation Time over two weeks of recordings, for one subjects, in two different conditions: (Left) looking at a uniformly coloured screen, and (Right) attending to salient stimuli. Variations between sessions over a week appear larger in the presumably salient mode. Illustration from [2].

to either subject A or B with high accuracy; the AUC was 0.996 for the classification using a Kullback-Leibler [238] divergence based distance metric and an Equal Error Rate of 2.4% even between the two weeks. Fig. 4.6 shows the confusion matrices for the 4 proposed metrics. The DFT based ones improved performance compared to not using it.

Seen as an argument for proposing new biometric system, it may not have much value to compare only two persons. But as an experiment looking at the stability of a signature over an extended period (3 months with 16 respectively 18 trials for

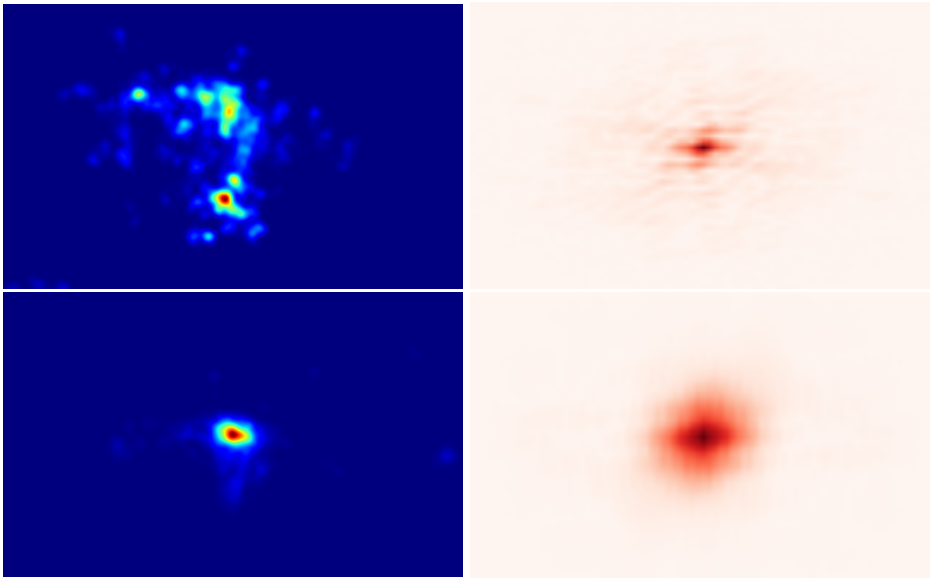


Figure 4.5: Conventional Fixation Density Maps (Left) recorded during presentation of solidly coloured screens, somewhat typical of subject A (Top) and subject B (Bottom). Fourier Transformed versions (Right) (in the spatial domain) corresponding to the Fixation Density Maps to the left. The fixation density maps here represent aggregated values and contain multiple fixations; the same transformation could also be applied to single fixations to visualize microsaccadic movements. Illustration from [7].

each subject) it does provide valuable insights into the robustness of the individual patterns which are seen in the fixation density maps.

We also hypothesize that what we see is not only the result of the OPM circuitry and the lower visual processing layers, but may also reflect activity in the attentional top-down processes that modulate the eye movements.

We also looked at saccadic latencies. Fischer and Boch [239] found a bi-modal distribution of saccadic latencies when testing monkeys, possibly indicating two different types of saccades exist: *express saccades* and *regular saccades*; see also Holmqvist [51]. Even though our experimental setup is not directly comparable, it is worth noticing that we did not see any such indications

We also didn't see much variation between sessions in saccadic latency, even if the subjects self-reportedly at times felt more tired due to other activities. We did, however, see significant differences in the average values between subjects and also a dependency on the direction. This is in line with what we would expect from the OPM, and it is worth noticing that even with the relative low sampling rate used here (60Hz), we were able to measure some aspects of the saccades. See Fig. 4.7.

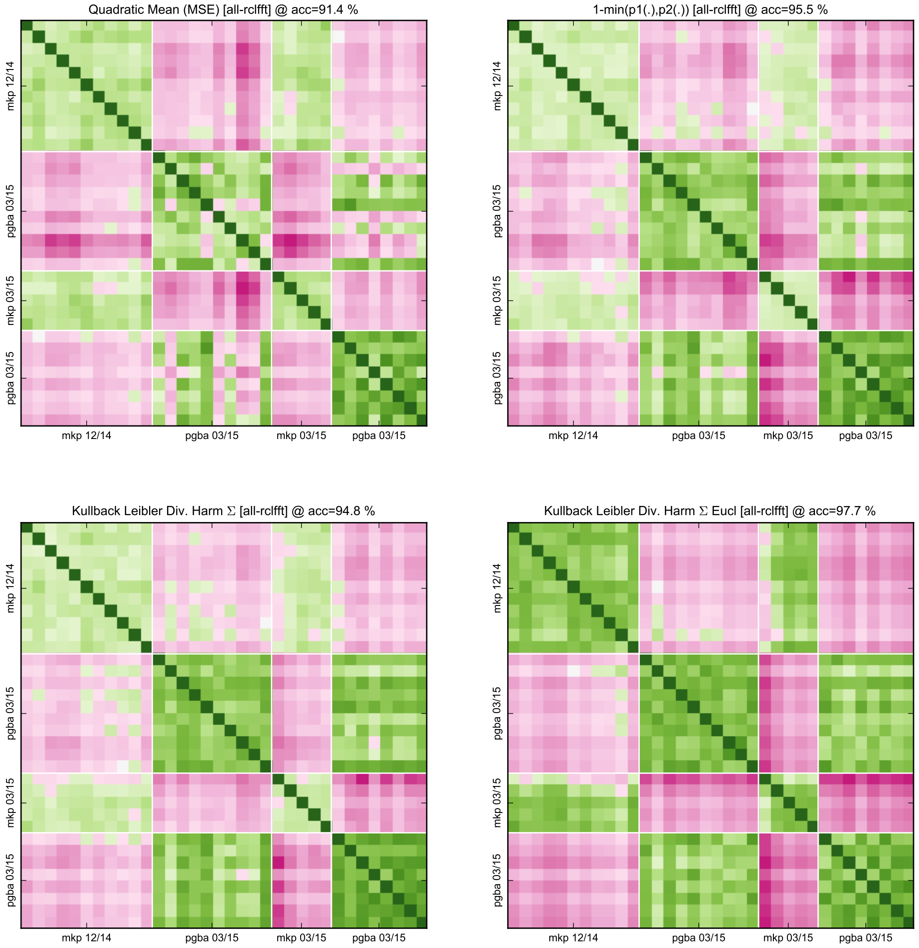


Figure 4.6: CONFUSION MATRICES over all trials, based on computing the 4 dissimilarity scores: Mean Square Error (TOP LEFT), 1-Min (TOP RIGHT), Kullback-Leibler Divergence (BOTTOM LEFT), and the Euclidean distance (BOTTOM RIGHT) over Fourier Transformed Fixation Density Maps across all trials. Each matrix is divided into 4 subgroups horizontally and vertically. The first two SUBGROUPS (TOP and LEFT) represents week 1, and the last two SUBGROUPS (BOTTOM and RIGHT) week 2. Within each week, the FIRST of the two SUBGROUPS is from subject A and the LAST from subject B. Each line (row or column) within the subgroups represent a single trial. The colours are chosen so that stronger red indicates a higher score, i.e. decreasing similarity, and darker greens indicate a lower score, i.e. increasing similarities. The split between red and green has been chosen at a threshold value that corresponds to the highest F1 score achievable. Illustration from [7].

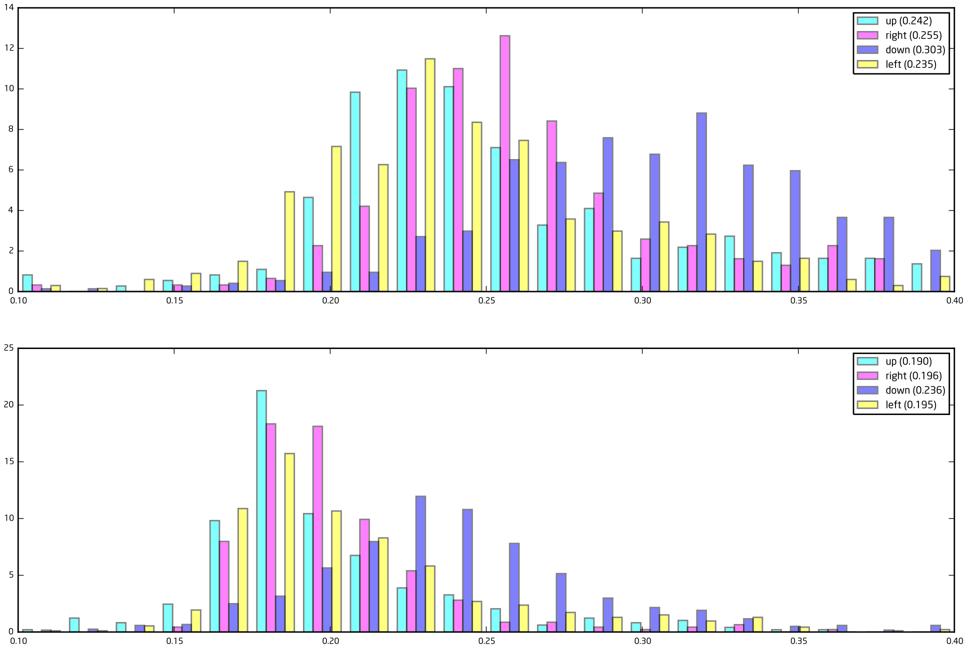


Figure 4.7: Split histograms of all time-to-target values for all trials for subject A (Top) and subject B (Bottom). Cyan, magenta, blue and yellow bars correspond to saccades in the upwards, right, downwards and left directions respectively. The distributions overlap but are clearly different between subjects and partly between directions (downward is slower). Illustration from [7].

Chapter 5

Phasic Pupil Dilations

As we have seen, pupil dilation responses are correlated with LC-NE neurotransmitter release which, according to the adaptive gain theory, reflect a fluctuating balance between phasic and tonic aspects related to the activation of attentional networks in the brain, as modulated by the arousal and impacted by stimuli attended to.

In task-evoked paradigms, such as Task-Evoked Pupillary Response, it is essential to determine the timing of the stimuli. Since we intended to measure levels of attention, the established Attention-Network Test, developed by [79], has for our purpose been chosen as a point of reference in our studies. It was not available in any form that could be used with our setup, so the author re-implemented [11] it in the scope of PsychoPy [80].

The test, which is essentially a behavioral experiment focused on teasing apart the three attention network components based on observed differences in reaction time, is illustrated in Fig. 5.1 and works as follows: The subject looks at a fixation cross, and after a (variable) delay, a cue is normally presented for 100ms and 400ms later a target is shown, to which the subject must respond. The target holds an arrow pointing to the left or to the right, and a key must be pressed on the keyboard in the direction of the target. Sometimes the target has congruent flankers (arrows point in the same direction), sometimes neutral flankers (no arrow head) and sometimes incongruent flankers (pointing in the other direction). The target can appear either above or below the initial fixation cross. Also, the cue that appears before the target presentation may be either a central star, a star at the location where the target will appear, two stars (above and below), or sometimes simply missing, so the target appears without any cue.

This is thought to invoke either of the 3 networks, alerting, orienting and/or conflict resolution [189, 190].

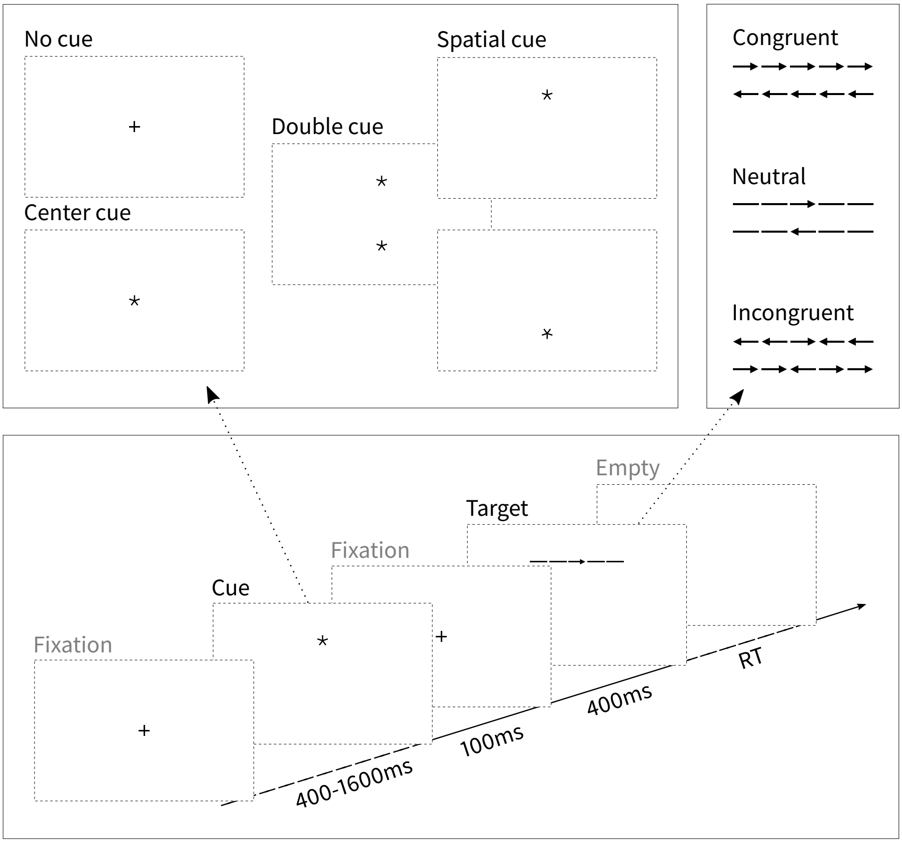


Figure 5.1: THE ATTENTION NETWORK TEST PROCEDURE: Every 4 seconds, a cue (either of 4 conditions (TOP, LEFT)) precedes a target (either of 3 congruency conditions (TOP, RIGHT)), to which the participant responds by pressing a key according to the central arrow. The reaction time differences between cue- and congruency conditions form the basis for calculating the latencies of the attention, orientation and conflict resolution networks. This figure and description is from [3].

Comparing the average reaction time of all the non-cued conditions with those that are double-cued allows for an estimate of the impact of having invoked the alerting network. Similarly, comparing the center-cued to the spatially-cued conditions allows an estimate of the impact of the orienting network. And finally, comparing the incongruent condition to the congruent condition allows an estimate of the impact of the conflict resolution network.

This is all a standard procedure, so what we did was to record the pupil dilation during the experiments, and we additionally performed it approximately twice a day

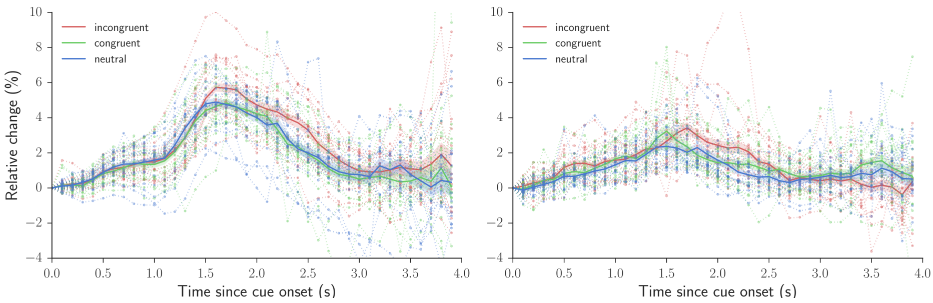


Figure 5.2: AVERAGED LEFT-EYE PUPIL DILATIONS for each session, coloured according to congruency (A (LEFT) and B). The all-session average shown in bold, with the shaded area representing the standard error of the mean. The average incongruent (Red) pupil dilation is stronger than the others, possibly indicating a higher cognitive load. These plots are from [3].

over a period of 2 weeks as a type of longitudinal study. Such an experiment has not, to our knowledge, been done before.

5.1 Summary of Assessing Levels of Attention... [C]

The outcome of the ANT test in terms of reaction time was in line with previous experiments; these are however typically an average over many subjects, so the absolute numbers can be different in a small population of only two subjects. We found the effect of being alerted to be around between 27–35ms and of being oriented to 22–49ms. The conflict resolution network apparently causes a delayed response of 81–85ms.

More interestingly, we also found differences in the pupil response, as shown in Fig. 5.2. The solid line represents the average of all congruent, neutral and incongruent pupil responses over all sessions, with individual session average values shown as a thinner line. For both subjects, the incongruent response is stronger compared to the two other.

Note how noisy the signal is, even after a relatively strong filtering and averaging over many trials.

We also looked at how large the response was for each session. Fig 5.3 shows the area under a (selected) part of of the pupil response curve, representative of the level of the response.

For both subjects, the two experiments on the first day has a much stronger response than the following ones. This could indicate a training effect; the stronger phasic response may indicate that the gain setting of the LC-NE adaptive gain system

Table 5.1: Pearsons correlation coefficients between key metrics for A (Top) and B. A shows negative correlation between mean reaction time and error rate ("speed vs. accuracy trade-off"). 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. Data from [3].

	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.

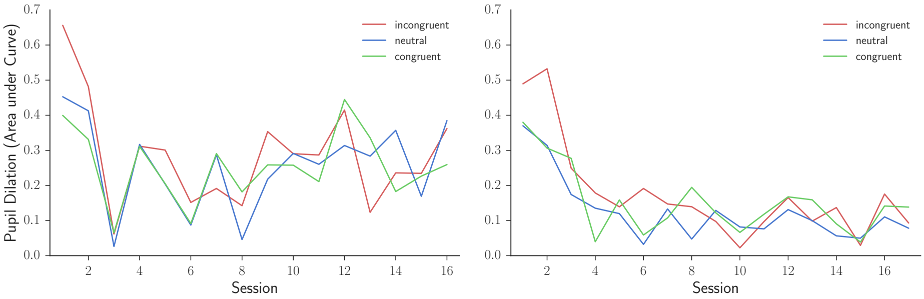


Figure 5.3: Area under left-eye pupil dilation curves $[1.5, 2.5]$s for each session, indicative of the size of the phasic response, grouped according to congruency. Both A (Left) and B show initial training effects; only A however show an increasing trend (in assumed cognitive load) for the remaining sessions. Illustration from [3].

is higher during the first experiments. Note that even though we measure the Task-Evoked Pupillary Response, since we're using similar types of stimuli between sessions, and only compare within-subject, we may be able to infer knowledge about the tonic level/gain settings as a higher gain results in a comparatively stronger response. The 3 responses can also be seen to co-vary (correlation coefficients are over 0.7 between different indices of the pupil dilations; see Table 5.1).

It is also worth noticing that subject B remains at the lower levels whereas A show an increasing trend. This could indicate different levels of adaptation, motivation, fatigue or many other factors – we simply don't know.

However, if we compare the correlations between mean reaction time and error rate, for A we see that there is an inverse relationship. This is not be unexpected; faster reaction times correlates with an increase in error rate, which is the classical speed vs. accuracy trade-off. There is no relationship, however, between the error rate and the size of the pupil dilation.

For B, this is different. There may be a (weaker) speed–accuracy correlation, but it is not statistically significant. There is, however a statistically significant correlation between the pupil dilation and the error rate. Whether this simply indicates (as we soon shall see) that an increased error rate leads to a stronger pupil dilation in response to having made an error, or if it also could be a signal that a perceived bad performance leads to an increased effort and thus a higher gain setting cannot be said on the basis of the available data at this point.

In any case, the difference between A and B is noteworthy.

We should also briefly note that the significance levels here are uncorrected. It would be appropriate to use a correction such as Bonferroni or Holms-Bonferroni within the families of related hypotheses here, so it is best consider mainly those that have a low uncorrected p-value, such as those marked to be below 0.25%.

5.1.1 Additional Note

After publication of the original paper we have become aware of a similar study done by Geva et al. [198], albeit in more controlled conditions than ours. The objective findings are consistent with what we see here: Geva et al. identify two different pupillary reactions, referred to as *Pa* and *Pe*; the latter is hypothesized to be modulated by incongruency, and then appear more prominently. Even when accounting for components of surprise related to incorrect replies (see Section 5.2) we also see the effect of the incongruent condition resulting in a stronger/more prolonged response (see also Fig. 5.5).

Some of ideas suggested by Geva et al., especially on the relation of the explorative and exploitative modes to the attention networks, has been questioned by Ambrosini et al. [199], and it remains inconclusive at this point.

With that in mind, some of the proposals listed by Ambrosini et al. on analysing baseline pupil diameter and looking at performance harmonize with some of the ideas that independently lead us to the following study:

5.2 Summary of Separating Components of Attention 1/2 [D]

The same experiment was repeated once with a larger group (N=18) of subjects. This was partly done to verify the previous findings.

Fig. 5.4 shows the average between-subjects pupil response for all trials which were correctly replied to. The magenta and cyan lines mark the confidence level of the differences between the incongruent response vs the neutral response. There is, as in the previous experiment, a difference, in this case from approximately 0.8s–2.0s after target presentation, with possibly a slightly longer difference in the right eye response.

It is not clear from the present data if the slight difference in left vs right eye response could be related to lateralization of the networks involved; the difference may not be statistically significant but comparing the significance levels between the neutral and the congruent condition in the left eye to that of the right eye shows the difference may be significant for a longer period, hinting at an extended reaction in the right eye.

For most, we see a high correlation between left and right pupil size (typical R-values are in the range of 0.80–0.95) although for some the recording quality or the tracking between eyes may be less than optimal.

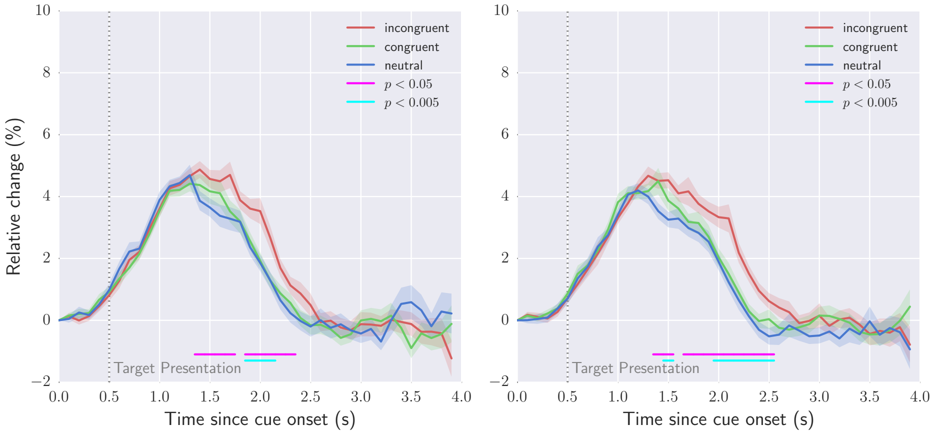


Figure 5.4: TASK-EVOKED PUPILLARY RESPONSES. The graphs show the difference between the three congruency conditions for the N=18 subjects of the study, for LEFT and RIGHT eye respectively. The incongruent response is larger compared to the two others. The statistical significance level is marked with magenta and cyan, when comparing the two set of data with a non-parametric test. The shaded area is ±1 SEM. Illustration from [4].

We conclude that the we can replicate previous findings of a stronger response for the incongruent condition also with a larger group of a different age distribution. The mean reaction times between the younger group was lower (505ms) compared to the older group from the previous experiment (569ms), but the limited amount of participants in the older group makes it difficult to rule out it could be related to individual variation. Also, accuracy cannot be compared meaningfully between the groups since, in the longitudinal study participants may have been entrained to the test, and if only including the results of the first experiment, the amount of measurements in the group becomes too small (i.e. comparing 2 measurements only with 18 other group).

Williams et al. [240] also finds differences in reaction times between age groups. Even if some behavioural differences were reported between the age groups with respect to alerting and not with respect to orienting and executive control, the ERP components related to alerting and orienting (P1, N1 and CNV) showed similar modulations between the groups whereas those related to executive control (N2 and P3) differed.

Overall, this might indicate that the test and methodology is relatively robust and translates between individuals although with some variation between those (caused by physiological or other ingrained differences, possibly influenced by current affective condition, as reported by e.g. Bellaera and Mühlenen [241]). In our (limited) studies, we can only conclude that the differences visible in the pupil dilation related to the incongruent condition (and to surprise; see below) are seen in both age groups.

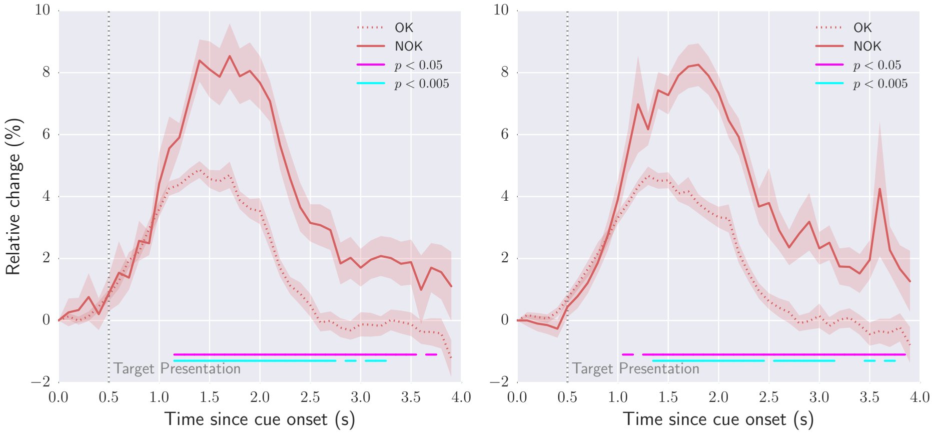


Figure 5.5: AVERAGE INCORRECT (SURPRISE) INCONGRUENT PUPIL DILATIONS vs correct responses for LEFT and RIGHT eye, respectively, for the N=18 subjects. An incorrect response invokes a statistically significant stronger pupil dilation, almost twice as large as for correct responses. Illustration from [4].

Comparing the incorrectly vs the correctly replied responses in the incongruent condition shows a large difference, which is statistically significant from 0.6s after the target presentation. The large difference may indicate that the subjects continue at an elevated phasic LC-NE stage in realization of having made an error – although not necessarily consciously. A similar result was also found with the original dataset from the longitudinal study (Paper [C]).

This could be related to the adaptation and required reorganization reported by Bouret et al. [242], or it may be related to what Preuschoff et al. [161] reports as surprise, and allow for an "adjustment" of the next expected response.

5.3 Concluding Remarks on Phasic Responses

We see that the phasic response differences can be divided into two components, either:

- related to the incongruent condition, or

- related to the incorrect response/surprise and the need for adjustments of subsequent responses

We hypothesize that, to the extent the Task-Evoked Pupillary Response are cognitively induced signals related to the LC-NE activity:

- engagement with a task, for instance through a decision process, causes a phasic pupillary response
- the magnitude of the response reflects some degree of the induced cognitive load, that different networks are involved or possibly that an extended temporal processing is involved
- committed errors or surprise causes a further increased response, possibly with the purpose of being able to reconfigure and be prepared for a (better) response to the next event

We shall now look at the tonic pupil size dilations.

Chapter 6

Tonic Pupil Dilations

We have previously seen that the tonic pupil size, in addition to modulations by light, also reflects the current gain level in the adaptive gain system controlled by the LC-NE system, which projects to many other important networks and is an important element in the sympathetic nervous system.

In this chapter, we shall look at some of the work we have done to measure and analyse elements of the tonic level; alone or as a combination of tonic and phasic activity.

6.1 Summary of Separating Components of Attention 2/2 [D]

In Paper [D], we additionally looked at elements of the tonic activity.

Fig. 6.1 shows the mean[1] pupil dilation of each of the 4 divisions of the experiment (training round and the 3 test blocks). The mean includes both tonic and phasic activity. Since the foveal focus is mainly at the center of the screen (fixation cross, cue and targets are all close to one another), we can assume that luminance changes are minimal and – if nothing else – have an equal chance of being distributed to any of the sections.

What we find for subjects in Paper [D] (Fig. 6.1, bottom), is that they have a significantly larger pupil dilation in the training round compared to any of the other three which, in between the three test blocks, do not differ statistically significantly. For the subjects in Paper [C] (Fig. 6.1, top), we see a similar statistically significant difference for Subject B, but not for Subject A. We hypothesize that this may again

[1] It would, of course, have been beneficial to separate the phasic activations from the baseline tonic level, but this was not possible to do reliably in the training round; further in the unconstrained task conditions found in Paper [E] we also have no possibilities of separating the two easily, so to make the studies somewhat more comparable, the overall means was chosen for this part.

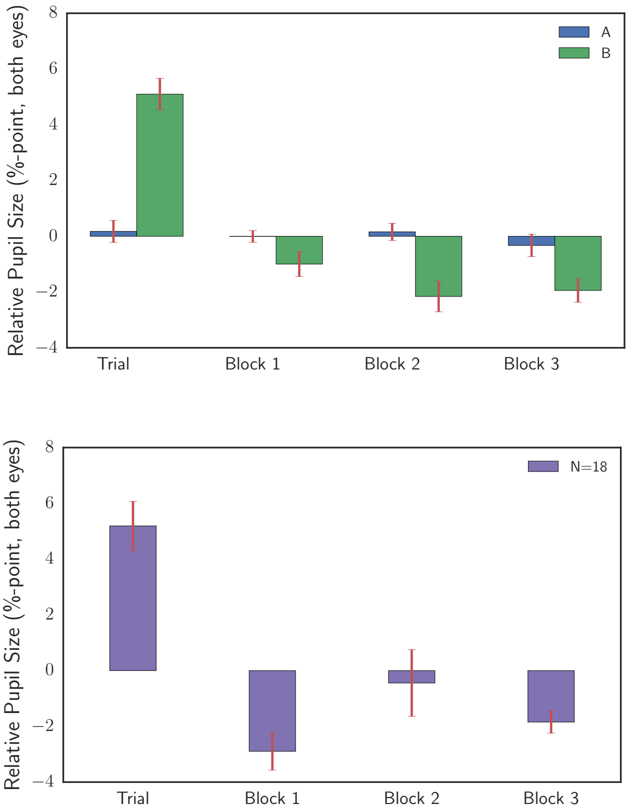


Figure 6.1: AVERAGE PUPIL SIZE for the initial trial run and the three test blocks for the original longitudinal study [3], for Subjects A and B (TOP) and for the more recent study [4] of N=18 subjects performing the test once. For B and the N=18 subjects, the average pupil size is larger in the trial block compared to the rest of the blocks whereas A has very little variation in the mean pupil size. For comparison, A's Task-Evoked Pupillary Response is typically larger than B's. Illustration from [4].

point to individual differences in characteristics, behaviour or preferences, and warrant further investigations.

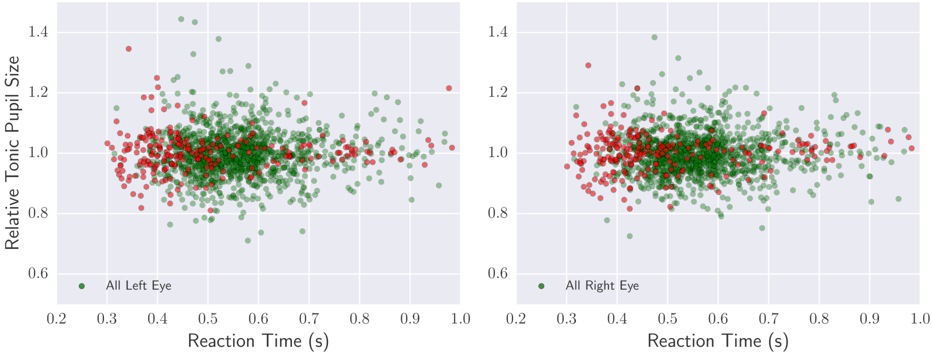


Figure 6.2: Scatter plot of the tonic pupil size relative to the session mean vs reaction time for incongruent conditions for all N=18 subjects (Left and Right eye), colour coded according to correct (green) and incorrect (red) responses. The mean reaction time between correct and incorrect responses are significantly different for the incongruent condition (Welch t-test $t = 7.00$, $p < 0.000001$). The mean relative tonic pupil size between correct and incorrect responses do not differ significantly. Illustration and data from [4].

In addition, the baseline tonic activity level was estimated by calculating the average normalized pupil size before the target was presented, and it was plotted vs the relative within-subject reaction time for that trial, in all incongruent conditions, grouped into those with a correct and an incorrect reply; see Fig. 6.2.

The distributions of the relative reaction times differs between the two conditions; the average value is lower for the incorrect replies, which is consistent with the normal speed–accuracy trade-of. This also applied to both subjects of the longitudinal study; see Table 6.1.

There were no statistically significant differences between the baseline pupil size for the present study nor for subject B of the longitudinal study; however subject A showed a difference (p=0.014, uncorrected) between the two conditions, so that a smaller average baseline pupil size is seen before incorrect replies.

We do not know why there is this difference, but again suspect that there can be individual behavioural or motivational traits behind. It should be noted that for the subjects of the present study, it cannot be ruled out that a further analysis and more data would reveal that some may have similar correlations.

Table 6.1: RELATIVE TONIC PUPIL SIZE AND REACTION TIME. Left eye tonic pupil size, as measured immediately before target presentation, relative to each session's mean, and the reaction times, are listed across all N=18 subjects of the present study [4], and for both Subjects A and B over all sessions of the longitudinal study [3], for the incongruent conditions, divided into groups of correct and incorrect responses. The mean reaction time (μRT) differ between correct and incorrect responses in a significant way (Welch' t-test $t = 7.00$, $t = 5.99$ and $t = 21.29$ respectively, p<0.000001) for both A and B. The means of the tonic pupil size (PSz) differ significantly between correct and incorrect responses for A (Welch' t-test $t = 2.47$, $p = 0.014$); for B and the subjects of the present study, the means between the conditions do not show a statistically significant difference. Almost identical results are found for right eye pupil sizes (not listed here). Table from [4].

			μ	SEM	N	p()
All (N=18)	PSz	OK	0.998	0.002	1472	
		NOK	1.010	0.007	254	
		δ	0.017			0.116
	μRT	OK	0.568	0.003	1472	
		NOK	0.496	0.010	254	
		δ	-0.072			0.000
A	PSz	OK	1.003	0.003	1333	
		NOK	0.982	0.007	198	
		δ	-0.021			0.014
	μRT	OK	0.631	0.003	1333	
		NOK	0.572	0.009	198	
		δ	-0.059			0.000
B	PSz	OK	1.000	0.002	1434	
		NOK	1.001	0.004	197	
		δ	0.001			0.845
	μRT	OK	0.612	0.002	1434	
		NOK	0.519	0.004	197	
		δ	-0.093			0.000

6.2 Concluding ANT Related Remarks

We have seen individual differences in several of the studies. Comparing in particular Subjects A and B from the longitudinal study [4] (including some aspects only reported in Paper [C] and not summarized here), we see a difference in that:

- The magnitude of the phasic response of Subject B stays at a comparatively lower level after the initial 2-3 experiments whereas for Subject A it increases again hereafter

- The accuracy of Subject A trends towards a decline (increasing error rate) whereas it does not for Subject B

- Subject A do show the expected accuracy–speed trade-off relationship whereas Subject B (instead?) show a correlation between accuracy and phasic pupillary responses

- The baseline tonic level of Subject A differs between correct and incorrect replies, whereas this is not the case for Subject B

- Subject B has a difference in mean pupil size between the training round and the following blocks whereas the mean pupil size stays on a flat level for Subject A

We cannot at this point offer any clear explanation of why this is so, but may hypothesize that Subjects A and B operate out of different strategies towards the experiment, possibly with different levels of motivation, or at different arousal levels compared to the task.

6.3 Summary of Pupillary Measurements... [E]

Even when performing eye tracking studies in an everyday context, where pupil data may be very noisy, insights related to aspects of attention might still be assessed based on aggregation of data from many subjects or even just based on the amount of time subjects attend to visual stimuli, such as to building instructions, when performing an assembly task.

This study involved approximately 60 children playing with LEGO®, following two different kinds of instruction manuals based on a Tablet and a Booklet (the media), in real "in-the-wild" conditions. We knew from the outset that this would be a difficult case for pupillometry, but nevertheless decided to see what could possibly be learned from such an experiment.

The children were divided into two groups, and given 10 minutes of building time with first one type of media, and were then switched to another type of media, continuing from where they were at that time in the building process. One group started with the Booklet and the other with the Tablet before they switched to the other media.

Fig 6.3 shows a graph similar to what was seen in the preceding section. At the top is shown the relative (normalized) pupil dilation for each group (Booklet First (blue), or Tablet First (green)) in each 5 minute block of the experiment, with two measures for each media of 5 minute pupil dilation averages, referred to as a New condition and a Known condition respectively. The statistically significant differences are marked together with their significance levels.

Table 6.2: CORRECTION TERMS FOR MODELS A AND B: These are percentage-points that can be summed when estimating the model mean pupil size for a given block of the experiment. The * indicates that the term differs from 0 at a significance level of p<0.05. The ‡ indicates a p<0.001 significance level. Note that Model B, when compared to Model A, should also include an additional factor of +0.21 to be added to the baseline for the slowest performers (not shown in the table). Table from [5].

Legend			*Model A: All*			*Model B / All but Slowest*			*Model B / Slowest*		
Time	Novelty	Media	Time	Novelty	Media	Time	Novelty	Media	Time	Novelty	Media
First 10 min	New	Booklet									
		Tablet			-1.28‡			-0.71			-3.03‡
	Known	Booklet		-0.89*			-0.89*			-0.89*	
		Tablet		0.22	-1.28‡		0.22	-0.71		0.22	-3.03‡
Last 10 min	New	Booklet	-1.71‡			-2.40‡			0.33		
		Tablet			-1.28‡			-0.71			-3.03‡
	Known	Booklet		-0.89*			-0.89*			-0.89*	
		Tablet		0.22	-1.28‡		0.22	-0.71		0.22	-3.03‡

At the bottom is another way of visualizing the same underlying data: For each subject, a 2nd order polynomial was fitted to the pupil data of the screen at each step, to achieve a very strong low-pass filtering effect. The average value of the filtered version (i.e. the average polynomial points) between the group was then calculated and plotted with the SEM shown shaded.

As can be seen, the Booklet First group starts out at a higher pupil size (approximately 3% at the beginning of the experiment) and gradually drops and flattens out around -1% for the 2nd half of the 20 minute total length of the experiment. However, the Tablet First group has – although with a larger variation – a more flat progress.

This leads to the model shown in Table 6.2. When trying to account for the effects of *media* (Booklet or Tablet), of *novelty* (New or Known, i.e. first or last 5 min with the media) and the overall *time* (first or last 10 min of the experiment), it can be seen that using the Booklet leads to a model 1.71% larger pupil size than using the Tablet, and that the difference between the first and the second part of the experiment is best modelled with a decline in pupil size of 1.28%. There are also differences in novelty: The Tablet stays at roughly the same level (insignificant increase of 0.22%) but the Booklet leads to a decline of 0.89% from the first 5 min with it to the last 5 min with it.

Separating the 25% slowest performers (those which had less than 25 completed steps and needed more frequent help) shows that there is an even larger impact of media type for that group: The Booklet now leads to a 3.03% larger pupil size, and furthermore, the decline otherwise seen over the experiment for the rest of the group (now at 2.40% decline) is no longer present.

So if the mean pupil size would be taken to be representative of the required cognitive effort, the slowest performers need to keep their effort levels constant, and receives relatively larger help from the Tablet than the other groups. The difference between the slow performers is also shown in Fig. 6.4.

Thus, it may appear that even in these "in-the-wild" conditions, we were able to see statistically significant correlations between the effects of media, time, novelty and performance reflected in pupil dilations, and that the simpler explanation in this case is that it is a real effect not just caused by uncorrelated luminance variations in the environment, partly due to the structure of experiment.

6.4 The LEGO® VW Transporter Study 1/2

In a separate study carried out in LEGO® shops in Copenhagen and Frankfurt, a gaze-controlled instruction manual was used by children assembling another small LEGO® model (#10220, a small VW Transporter). In total 59 children participated in the study, boys and girls between 5 and 13 years. This was done to experiment with active gaze-control but at the same time, gaze and pupil data was recorded for later analysis. We will report additional findings in the following chapter[2], but we will look at some of the recorded pupil data here. The subjects completed the building process at different pace (more on this later), so to be able to compare the results, we use as time-scale a relative fraction of the total time needed for each subject.

Fig. 6.5 shows a graph of the mean pupil size during the experiment, similar to Figs. 6.1 and 6.3, where the progress here has been divided, somewhat artificially, into 4 equal-length blocks to make it comparable. In this case, however, we have not only corrected for distance variations between the head and the Tablet[3], but given information on which step is being shown to the user, we can estimate the luminance of the Tablet and do another linear regression[4] with respect to the pupil dilation

[2] A separate paper is in preparation on this study with additional details of the setup.

[3] This is done by using the reported inter-pupil distance as a baseline with respect to which the pupil size should be corrected. We have further validated by a linear regression that there is a strong correlation between the inter-pupil distance and the pupil size (which needs to be corrected for here!), and that the estimated model parameters fits with our theoretical model.

[4] See also Fig. 2.7.

and use the residual as an estimate of the cognitively and affectively impacted pupil signal. This is not perfect, but as a first approximation it may help to remove at least some of the effects of luminance variation due to the differences in how the model is presented. We find that the correlation between luminance and pupil size can vary a lot (with this method), but typical R values may be in the range of -0.1 to -0.2.

We see a similar (statistically significant) difference between the first part of the construction process and the remaining 3 blocks. This could, as previously, indicate an increased initial arousal in preparation for a somewhat new situation, followed by some entraining (familiarity effect).

When looking at the bottom graph in Fig. 6.5 we see that the mean pupil size apparently increases towards the end of the building process. We hypothesize that this is caused by the achievement or near-completion of the construction; an increased baseline level would be consistent with task completion and the fact that the subjects are about to move from exploitation mode (building the model) to more of an exploratory mode (reorienting to other stimuli in the environment).

We cannot rule out other explanatory models; as noted by Kahneman [201], arousal cannot be treated as a unitary dimension and even if the context and terms in which he described it are different from what we discuss here, the warning still applies!

Finally: As an additional test, we also analysed the impact of using data that is not corrected for luminance. We still find similar differences, although with a less pronounced increase towards the end of the construction process, as expected, since the last part of the manual (with more white bricks) has a higher luminance level.

6.5 Concluding Remarks on Tonic Pupil Size

Across four different studies with 3 different age groups (people in their 50s, in their 20s and children between 5 and 13), we have, to varying degrees, seen an enlarged pupil at the initiation of the experiment followed by a gradual decline and – in one case – a slight increase again in pupil size towards task completion. This was seen both in two studies with 2 and 18 subjects, respectively, where screen luminance was relatively stable, in one study where we included 43 children using a different system, a head-mounted eye tracker, and where light conditions were not controlled for, and in a study of 59 children where light conditions to some extent could be compensated for.

We hypothesize that, to the extent these are cognitively induced signals related to the LC-NE activity:

- This common pattern is related to an increased arousal level in anticipation of the task to commence.
- The gradual decline reflects a degree of familiarization with the task.
- Conversely, no decline reflects a need for a continued sustained arousal.
- An increased arousal towards the end of a task might be related to a sense of completion and preparing for a shift from exploitation mode to a more explorative mode.

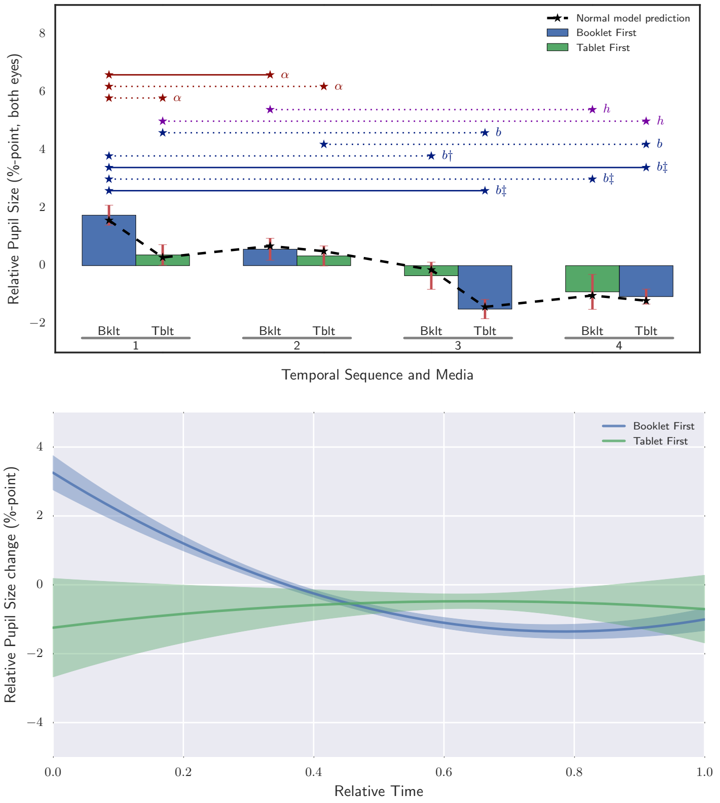


Figure 6.3: (Top) Post-Hoc between the means of the data from each of the blocks of the ANOVA regression colour coded according to the two groups of subjects that had Booklet first (blue) and Tablet first (green), with corresponding significance levels marked as well as the predicted value (dashed line). The red lines indicate one standard error of the means to each side. The lines above the bars indicate where there are statistically significant differences, with solid lines indicating within-group and dotted lines between-groups. The markers α indicate uncorrected significance, h corrected with Holm-Bonferroni factor, and b the most conservative with Bonferroni correction. The levels are at p<0.05 unless marked with a † or ‡, that indicates 0.01 and 0.001 respectively. (Bottom) Filtered trend of how the mean pupil size develops over time for the two experimental groups of subjects that had Booklet first (blue) resp. Tablet first (green). The graph is the result of filtering the data for each subjects using a 2nd order polynomial (i.e. a heavy low-pass filtering) and then calculating the means over all of these. The shaded area represents one standard error of the mean to each side. Illustrations from [5].

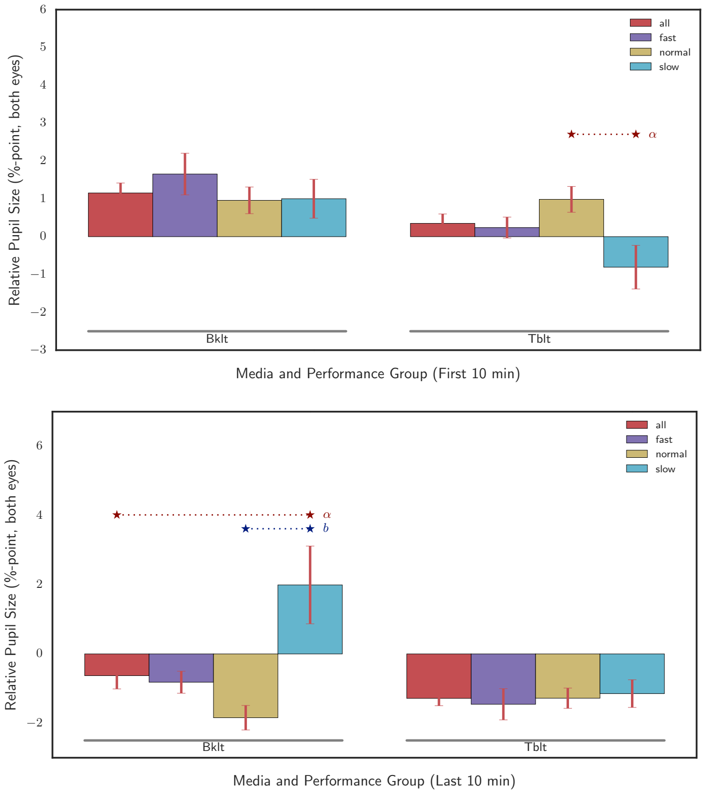


Figure 6.4: Differences between the slowest and fastest performing groups vs all or all others, here shown over the two 10 minute periods of the experiment (first 10 minutes at the Top, last 10 minutes at the Bottom). There are statistically significant differences during the last 10 minutes when continuing on the Booklet (after 10 minutes of Tablet use), as marked: The slowest performers have significantly larger pupil size compared to the normal performers and to the overall mean (bottom). There is also a statistically significant difference between the slowest and the normal performers for the group that start with Tablets during the first 10 minutes (top). An α denotes significance at the $p<0.05$ level; a b denotes significance at the same level after Bonferroni correction. Only comparisons within each media (Booklet or Tablet) and time (first or last 10 min) between the 4 groups of performers have been tested and is shown. Illustration from [5].

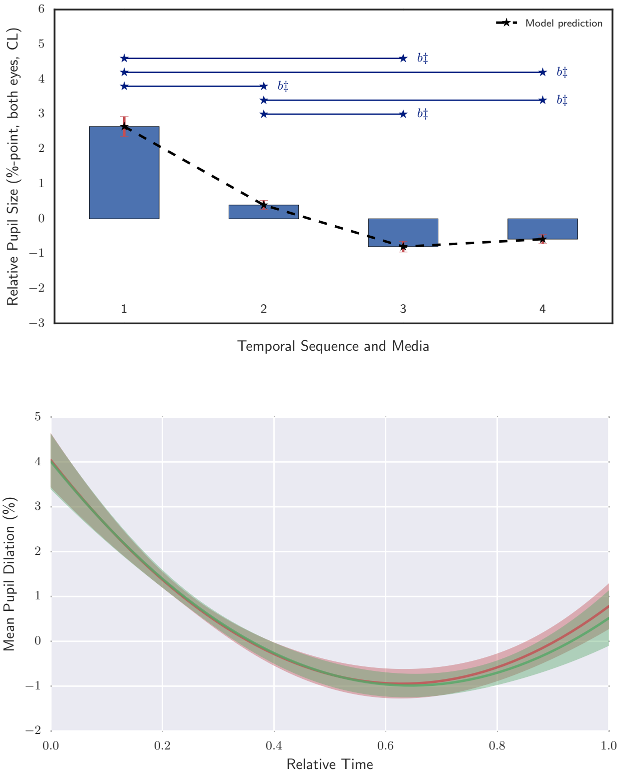


Figure 6.5: Construction Manual Pupil Size Differences (after correction for luminance variations). At the (Top) is shown the aggregated mean pupil size within each quarter of the building progress, relative to the total length for each participant, across participants of the study. The significance levels are also marked. (The "model" here is simply the means of the measured value). At the (Bottom) is shown the aggregated pupil size resulting from a low-pass filtered estimate using a 2nd order polynomial, similar to Fig. 6.3.

Chapter 7

Using Weak Signals

In this chapter, we will look at the use of "weak" signals, by which we mean signals that may not in themselves individually carry meaning, but taken in context and over time, they may be used to construct a more meaningful aggregated signal. Such signals are for instance dealt with in the context of e.g. social-media streams, where they are sometimes treated "*more akin to art than to science*" [243].

Here, we will however try to make systematic use of weak signals in three different contexts, trying to find value in signals that are otherwise typically ignored.

The first signal is the presence of gaze vs no tracking: Is a subject looking at a screen for instructions (where we can then get more detailed gaze and pupil data) or have we lost tracking, and the subject is engaged in some other (possibly related) activity than studying screen content? In combination herewith, we will also look at information on which step is currently being shown, as a proxy for progress.

The next signal is a simple screen-on event recorded by a smartphone. We will use that to infer sleep patterns in a larger population.

The final signal we will consider is a blink, and we will use that and show how to synchronise recordings in multiple modalities in the absence of precise timestamps.

The overall motivation for looking at these signals is that they can often be collected without too much extra effort, and they can in some cases be used to better understand user behaviour and possibly adapt to individual needs.

7.1 The LEGO® VW Transporter Study 2/2

We will continue to analyse data from the LEGO® Study mentioned in Section 6.4.

The gaze controlled manual has 18 steps, some of which contains sub-assembly steps. These are not shown individually here for clarity.

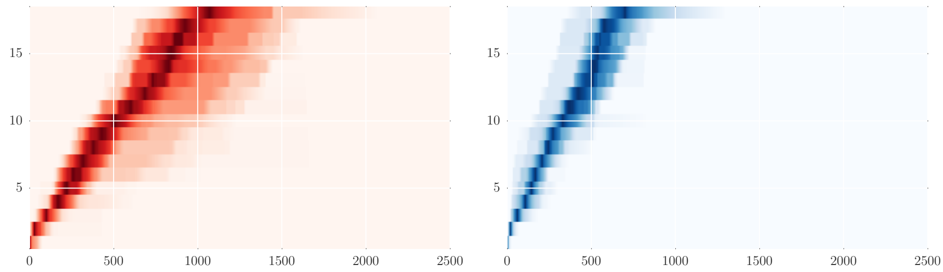


Figure 7.1: LEGO® Building Progress for girls (Left) and boys (Right), averaged over all participants in the experiment. The visualization shows the distribution of time spent within each step: The darkest color shows when 50% of the total time within each step has been spent, with the lower and higher quantiles shown increasingly lighter. Within the analysed group, girls progress slower than boys, and with a larger spread. The current step is shown on the y-axis and the elapsed time in seconds on the x-axis.

Of the N=59 participants in total, there were in total 32 labelled sessions where age and gender is known (20 boys and 12 girls, aged 5 to 12 with mean ages of 8.1 and 8.0 years respectively). Of these, 24 completed the building instructions (14 boys and 10 girls). For these, all steps forward or backwards were recorded, as was the presence of gaze on the tablet that showed the construction manual. From this data, we can build a progress model (with regressions) for each participant, and we can also create one or several aggregated building metrics.

Fig. 7.1 shows the aggregated building progress for boys and girls, respectively. Girls generally appear to progress slower and with more spread compared to boys.

This may be related to many factors which we cannot tell from the current experiment (spending more time playing with the interface, more attention to detail, differences in 3D perception, ...). The average age for boys and girls are comparable, 8.1 respectively 8.0 years; the age span is 5 to 12 years for boys and 5 to 11 for girls.

Assuming that the construction progress can be modelled as a function of 1) gender, 2) age, and 3) difficulty of individual steps, based on the weak signal of the time recorded within each construction step including any regressions, this can be expressed as follows for boys and girls:

$$\begin{aligned} \text{steptime}_{boy}(N, age) &= \text{baseline}(N)(1 + k_{age,boy} * (age - 8)) \\ \text{steptime}_{girl}(N, age) &= \text{baseline}(N)(1 + k_{girl} + k_{age,girl}(age - 8)) \end{aligned} \tag{7.1}$$

where $\text{baseline}(N)$ represents the step difficulty and is considered the time it would take an average boy of age 8 to complete step N, k_{girl} accounts for differences between boys and girls at age 8, and $k_{age,boy}$ and $k_{age,girl}$ accounts for age dependent differences for boys and girls, respectively.

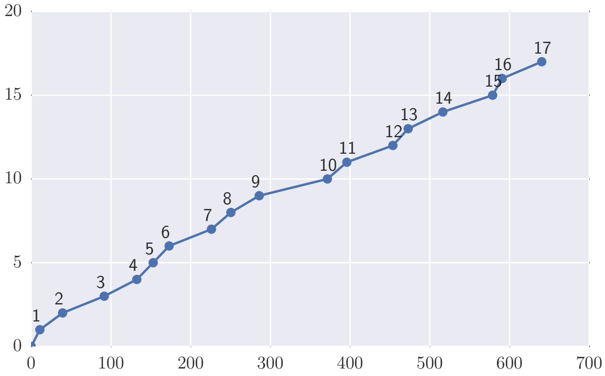


Figure 7.2: Model Building Progress (Step vs Time) showing the derived model, which assumes an average 8 year old boy. To estimate another age and gender, the corrections in Table 7.1 can be applied. Completed steps are shown on the y-axis and the elapsed time in seconds on the x-axis.

Estimates of individual step baseline times and the relative dependencies of age and gender have been derived using both a Bayesian model (using a Markov Chain Monte Carlo simulation on a model that fits all data points individually, with weak priors), as well as a conventional OLS regression (treating each gender independently), fitting the existing labelled data of 14 boys and 10 girls that have completed the construction process. The results are shown in Fig. 7.2 and Tab. 7.1. The Bayesian and the conventional OLS regression models result in similar estimates, especially given the limited amount of participants, but the confidence level of the conventional regression is lower as the data is split into two groups[1].

The results of the Bayesian model indicate that, with this construction set, a girl at age 8 would on average be 68.9% slower than a similarly aged boy. As the children grow older, the differences becomes smaller, as N additional years of age makes a girl $N \cdot 12.6\%$ faster compared to a boy that only gets $N \cdot 4.8\%$ faster. The

[1] We're using the mean value over normalized steps in each age group for the OLS regression and estimated boys and girls separately; the first part to make the results more robust with respect to outliers. However, the confidence interval is more difficult to properly evaluate as the number of data points then becomes small. The explained variance, when comparing the fit to the relative mean value of each age group, are 0.53 and 0.44, for boys and girls respectively. The Pearson Correlation R and p-values are R=-0.730 and p=0.062 for boys and R=-0.665 and p=0.149 for girls respectively. Accordingly there is not enough data to rule out that the correlations are accidental, based on conventional p=0.05 significance testing, when separating the genders.

Table 7.1: BUILDING PROGRESS MODEL FACTORS. This table shows the estimated dependency on gender and age for the progress through each building step vs its baseline: The gender dependency shows the amount of time an 8 year old girl needs in addition to an 8 year old boy following the same construction manual (though not necessarily ending up with the exact same construction result). The age dependency indicates how much each year impacts the amount of time needed to follow the construction manual for boys and girls, respectively. The first line shows the results of a Bayesian model that fits all of the available data in one model. The second line use conventional OLS regression on the mean values (of the normalized steps for all users of each age group). OLS regression on gaze tracker data is performed as two separate regressions, on boys and girls respectively.

	Boys		Girls	
	Age Factor	Gender Factor	Age Factor	Gender Factor
Bayesian Model	-4.8%	-	-12.6%	+68.9%
OLS Model	-5.1%	-	-12.3%	+74.3%

95% confidence intervals of the Bayesianmodel estimates are -0.7% and +0.6% for the gender dependency and approximately +/-0.2% for the age dependencies, based on analysing the trace of the Monte Carlo Markov Chain simulation used by the Bayesian sampler (PyMC3).

The building progress curve should be compared to an overall "designed experience": Should progress be faster initially and slower in the middle section, with a final fast progression, to make it more interesting and maintain some desired level of arousal – or should it remain at equal speed, so all steps take up approximately the same time?

Another view into the building progress is made possible by combining gaze data with the overall progress log, and is shown in Fig. 7.3. This view shows the amount of time spent on studying instructions vs. doing the actual building work (in this case: working with the LEGO® bricks). It also shows how much regression time has been spent, which may point to a failure in comprehending the instructions fully at first visit. In this particular case, steps 8, 9 and 14 appear to have been more difficult to understand at first visit.

Fig. 7.4 shows the relative time predicted by the model for each step (i.e. the $\mathrm{baseline}(N)$) compared to the overall average step time, and is as such an indication of the "difficulty" of each step (or at least an estimate of how long it takes to complete it). Comparing Fig. 7.4 to Fig. 7.3 can be helpful to understand if a given step is just time-consuming (i.e. it may contain many tasks to perform) or if the instructions are difficult to understand at first view.

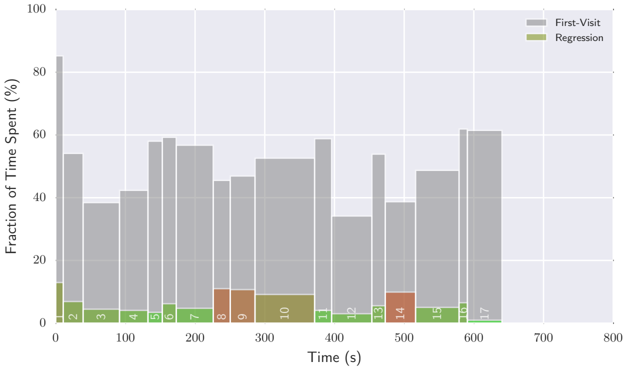


Figure 7.3: Building Progress with Regressions showing time spent on studying the instruction Manual on first visit vs subsequent views going back to a previous step ("regressions"). Overall model time (seconds anticipated for an 8 year old boy) is shown on the x-axis, and relative fractions of time is shown on the y-axis. The height and colour of the coloured bars show the fraction of time spent looking at the instruction manual on regressions: going back to a construction step after having progressed to a subsequent step. The height of the gray bar shows the fraction of the time that was spent studying the instruction manual on first visit to the step. The remaining area, on top of the bars, show the time spent doing the actual construction work, and not looking at the instruction manual. The width of the bars correspond to the time spent in each step.

For each step, the mean step duration and the mean fraction (shown in percent) hereof spent gazing at instructions can be calculated across all users that completed the building process. Fig. 7.5 shows this correlation. When the outliers are excluded from the calculation, a linear regression yields a statistically significant correlation:

$$\text{gaze time}(N) = [-0.44 \cdot \text{step time}(N) + 66.9](\%) \tag{7.2}$$

with an explained variance of 0.73, and a Pearson Correlation R=-0.853, with a p-value of 0.000211. Thus, steps that take longer to complete apparently need increasingly smaller fractions hereof for gazing at instructions

The outliers are the first and last step of the construction: At the first step, more time is presumably taken to familiarize with the interface, and at the last step, the instructor interviewed the participants. The 3 other outliers (steps 7, 10 and 17) turns out to be the (only) steps made up of additional sub-steps, and thus can be expected to take longer to comprehend and navigate. This particular construction manual is likely very well designed and only shows easily explainable outliers. We suggest that had other steps shown up far away from the trend-line, it

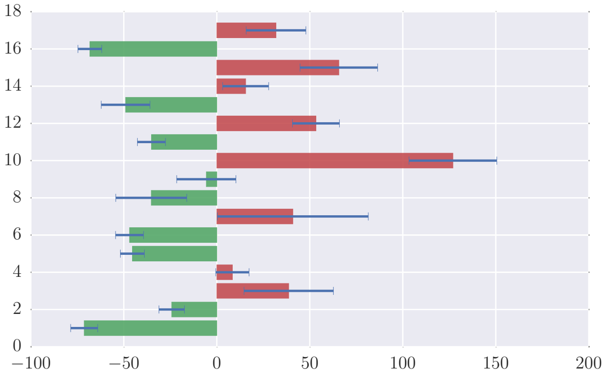


Figure 7.4: RELATIVE STEP DIFFICULTY: The relative time (in percent) spent for each step in the baseline model, which can be taken as an indication of the “difficulty” (to comprehend or complete) each step. Negative (green) numbers are faster than the mean of 37.7s, positive (red) are slower. The blue lines indicate the Standard Error of the Mean, based on the normalized distribution of each step taking into account each participant’s age and gender. The x-axis shows the percentage and the y-axis the step number.

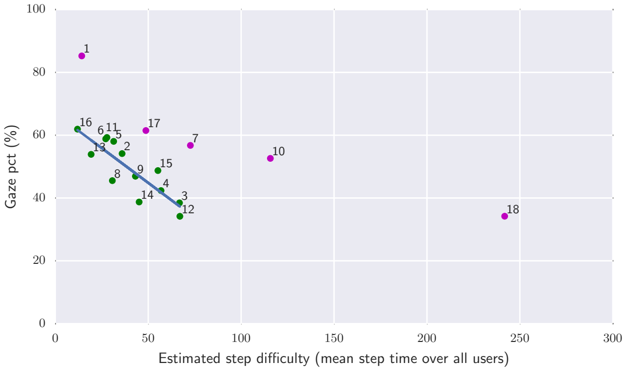


Figure 7.5: GAZE VS STEP DIFFICULTY showing, for each step, the average amount of time (y-axis, in percent) spent gazing at the tablet with instructions, as a function of the mean step time (x-axis, seconds) for all participants at each step. The magenta dots are considered outliers and are excluded from the trend-line, shown in blue. There is a significant correlation, and steps that take longer to complete apparently need a smaller fraction hereof for gazing at instructions.

is recommended to analyze the reasons in more detail to potentially improve the instructions and the overall building experience.

7.2 Summary of SensibleSleep: A Bayesian Model... [F]

We will continue with an even weaker signal, hardly considered a biometric marker: Screen-on events from a smartphone.

Through the Sensible DTU [244] project, a large amount of personal data is collected from the many participating students. We wanted to know if it would be possible to extract data about sleep patterns from only this signal.

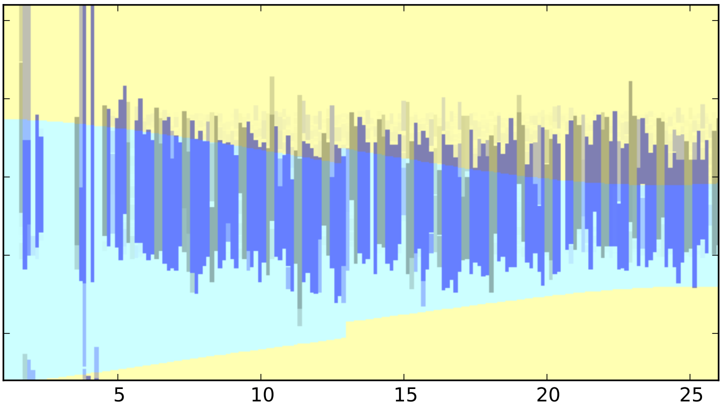


Figure 7.6: SLEEP PATTERN extracted from screen-on events from a smartphone, using a naïve approach of just finding the longest period with no (or only very short) uses of the mobile device during the night. Even in this case, patterns (like waking up around sunrise, including adaptation to daylight savings hour until some time in the spring when the sun rises too early...) are easily visible. The x-axis shows the week number since Jan 1st and the y-axis runs from 4 pm (bottom) to 4 am in the middle to 4pm the following day at the top, with the coloured background indicating either daytime (yellow) or night (light blue).

Fig. 7.6 show a visualisation based on a very simple heuristic analysis that locates the longest non-interrupted period (i.e. with no screen-on events) during the night, using a prior probability that sleep is more likely to take place 12 to 24 hours after the previous sleep period. A single short interruption does not break the registered sleep cycle. This is somewhat similar to what was later proposed by [245].

Some patterns are clearly visible; there appear to be some weekly patterns (getting late to bed happens more frequently during weekends than during weekdays, for instance) and there is also some correlation to sunrise and waking up in the spring and around the daylight savings time change.

However, we wanted to improve the method and build a more robust method. For this, we built a Bayesian probabilistic framework with the assumption that there would be two different rates of screen-on events depending on whether the person is awake or at sleep – which is a rather weak prior belief.

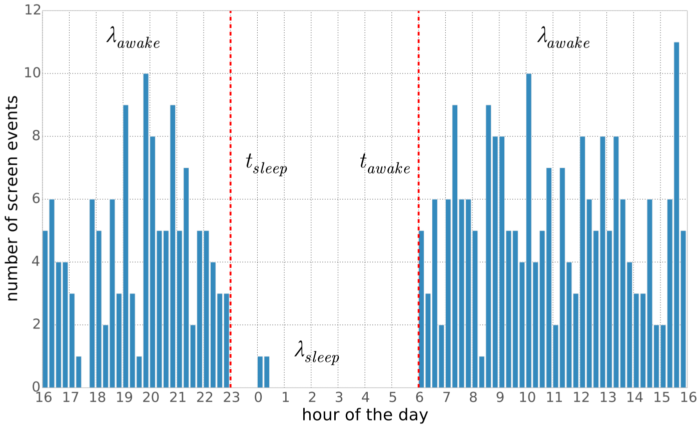


Figure 7.7: SCREEN ON EVENTS counted in bins of 15 minutes for a particular user, showing the two different rates of events between day and night, and the switch-points between the two. We assume that sleeping takes place in the period with the lowest event rate (here between 11 pm an 6 am). Illustration from [6].

Fig. 7.7 shows the rate of events for an example user. It is easy to guess where the user is sleeping based on this. The events seems to come from two different distributions. We assume this would be Poisson distributions with different rates, λ_{awake} and λ_{sleep}, with two different switch-points between the two distributions, t_{sleep} corresponding to "bedtime" and t_{awake} corresponding to morning "wake-time".

We developed several models of increasing complexity, including several hierarchical ones, and tested which ones would give sensible results based on the data we had access to. In addition to Sensible DTU data, we had access to data courtesy of Sony for which we also had sleep-tracker data to use as ground truth, in total more than 400 users. We report several metrics related to the performance of the different models, and also compare them to previous work.

We believe that our framework is more robust with regards to a few interruptions during the night period. In addition, as a Bayesian framework, it quantifies also the assumed distributions of the estimated parameters, and it can potentially encode prior beliefs (when desirable). When using the hierarchical models, it can further capture more complex patterns and dependencies between the variables of the model and it may serve the purpose of also revealing patterns of commonality between otherwise heterogeneous groups.

Fig. 7.8 shows an example of the output of the derived sleep period overlaid a visualization of the screen-on events from one single user.

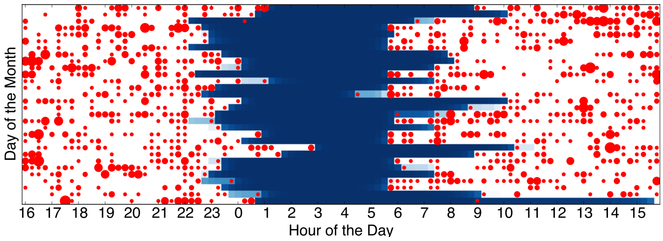


Figure 7.8: Screen On Events and Derived Sleep Pattern for a month for an example user. The red circles indicate one or more screen on events during the particular 15-min time bin, with a larger circle representing more events. The blue area is the probability of the user sleeping, derived with the SensibleSleep method proposed in Cuttone et al. [6], from where this illustration is also taken.

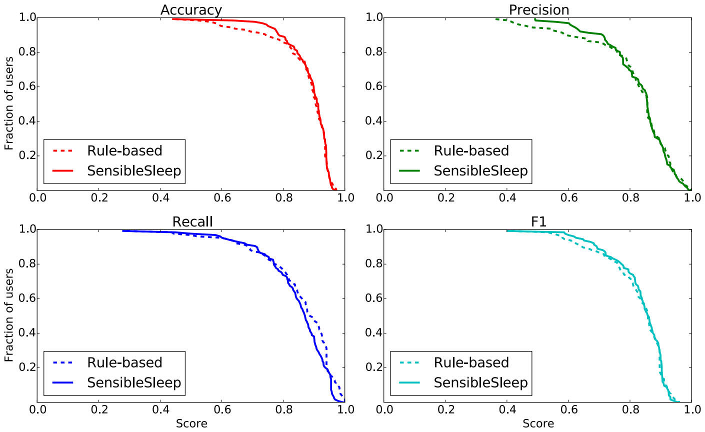


Figure 7.9: Accuracy of the Method shown as complementary cumulative distributions of accuracy, precision, recall and F1 scores, compared to rule-based models proposed earlier by e.g. Abdullah et al. [245] (dashed line). The graphs show the proportion of users on the y-axis that has a score less than, or equal to, the value on the x-axis. Illustration from [6].

Compared to previous work, using sleep trackers as the ground truth, the SensibleSleep model performs at least as well when it comes to accuracy and precision, and has a very comparable (weighted) F1 score, as shown in Fig. 7.9.

Fig. 7.10 shows the aggregated sleep and wake times for a particular set of analysed users. The distributions look somewhat akin to a Gaussian distribution, with what is probably also an expected mean value of the group analysed.

Thus, the model seems to work and serve the purposes we intended it to. A drawback currently, however, is the runtime since analysing longer periods of data is

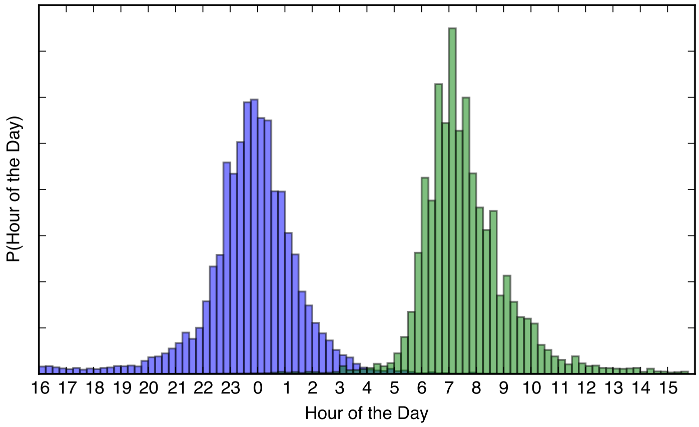


Figure 7.10: AVERAGE SLEEP AND WAKE TIMES for the set of users being analysed. Illustration from [6].

rather slow due to the Markov Chain Monte Carlo (MCMC) sampling required of sometimes relatively complex models.

7.3 Summary of In the Twinkling of an Eye... [A]

A final example of using a weak signal is that of using blinks to synchronize between eye tracking and EEG. The motivation for this came from a rather practical challenge in aligning properly data-stream from heterogeneous sources in multiple modalities in an environment where multiple devices (smartphones, computers, tablets, ...) could be in use.

Fig. 7.11 shows an example of blink probabilities (top, red) extracted from eye tracking data alone, based on absence of pupil tracking. Also shown is the power in two central EEG electrodes PFz and AFz, which provides a very "clean" blink signal[2].

Even if there may not be complete agreement on when a blink takes place or not, judging from the two modalities, as long as the blinks that are present have similar timing, the offset between the two data-streams can be calculated by e.g. a cross correlation and finding the timing of the peak – which then corresponds to the offset. This is shown in Fig. 7.12.

Fig. 7.13 shows the variation in the determined offset depending on the length of the analysed periods. As long as a few blinks are present, the offset can be

[2] Researchers whose main interest is in EEG would likely call this a noisy signal as there is a lot of blink data present that is unwanted in the continued analysis.

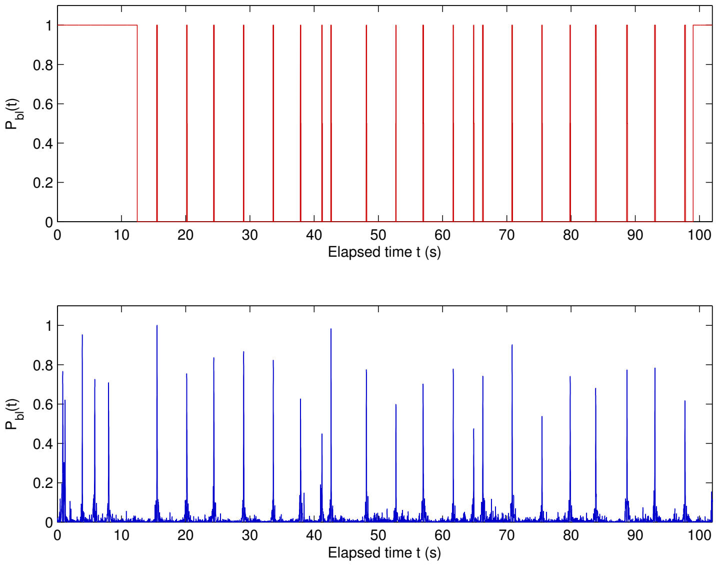


Figure 7.11: BLINK DETECTION in the ET data (TOP, red) and the EEG (BOTTOM, blue). As expected, since there is a common cause for the two signals from either modality, there is a high similarity between the two signals. Illustration from [1].

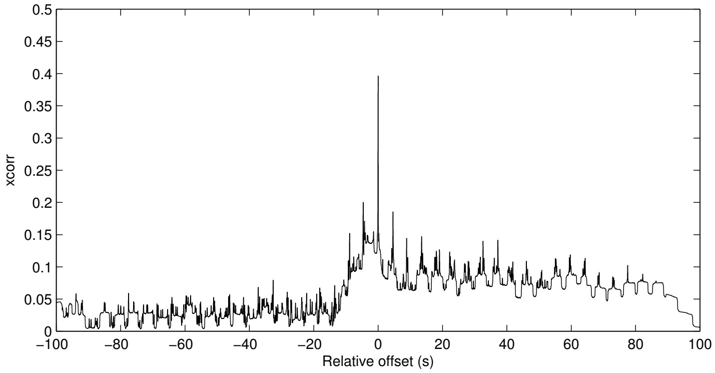


Figure 7.12: CROSS CORRELATION BETWEEN ET AND EEG signals, showing that there is one specific offset that can be used to align the signals so that they have the best match. Illustration from [1].

determined to within a few samples, given the assumptions on how a blink looks (see the paper for details).

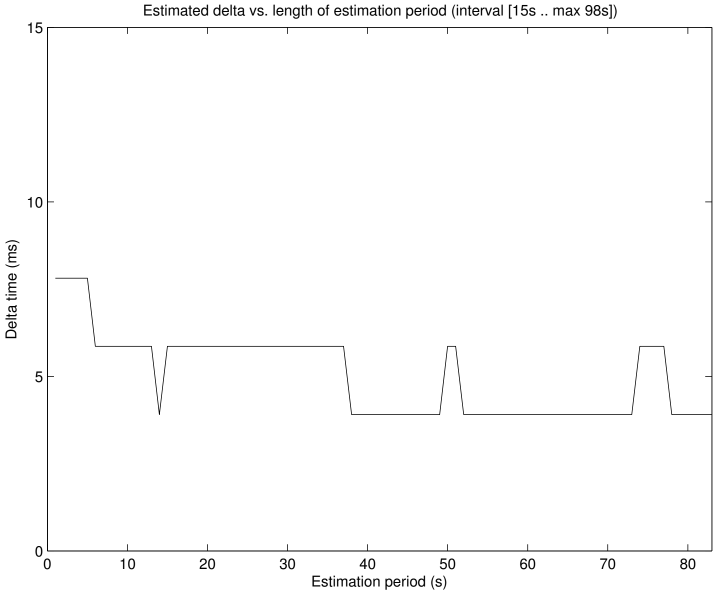


Figure 7.13: SENSITIVITY OF OFFSET TO ANALYSED PERIOD showing that even the derived offset is relatively stable, typically on the order of a one or a few samples, as longer periods are analysed. It may therefore also work in near real-time analysis. Illustration from [1].

7.4 Mentions of Your Heart Might Give Away... [H]

This study was an early attempt of estimating emotional responses using Heart Rate and Heart Rate Variability, based on previously demonstrated paradigms, such as those reported by Bradley et al. [171], Lang and Bradley [147], Löw et al. [57], and Wallentin et al. [55].

We shall refer to the paper for the details, but just highlight the conclusion: That finding a robust signal with a small set of subjects and a noisy signal was not conclusive. Some trends were observed, but a statistically significant result in such almost single-run conditions was difficult.

This lead to a search for other and potentially more robust signals.

Chapter 8

Summary

We set out to identify ways of measuring elements of attention based on eye tracking, with the intent of improving the user experience and pave the way for e.g. adaptive systems that can tailor their behaviour to some elements of the current cognitive and affective condition of the user.

We succeeded in identifying components of attention and cognitive load in pupil dilations, as well as elements of surprise.

We also found large individual differences and variations; some may simply be related to the mechanics of the eye or the lower layers of the visual (control) system, but some may also be related to cognitive and affective processes at higher levels. Variations from the personal baseline between conditions may serve as indicators for load or other cognitive or affective processes, and can in some cases be related to objective measures like performance or time on target.

We did all of this "in-the-wild", using low cost equipment, trying to build on established paradigms and on the insight of scholars before ourselves.

But there is still some way to go. The signals are inherently very noisy, and better methods for extracting the underlying signals of interest are needed. It may also be worth combining signals across modalities to better disambiguate their interpretation.

And even though we're learned a lot during the last 30 years, there is still a need for an improved understanding of the underlying networks and their interworking, and how they modulate other processes.

But the results may turn out to be very rewarding, and could help create much better user experiences for future products and services!

Paper A

In the Twinkling of an Eye: Synchronization of EEG and Eye Tracking Based on Blink Signatures

This paper was presented by the author at the 4th International Workshop on Cognitive Information Processing in May 2014 in Copenhagen.

4TH INTERNATIONAL WORKSHOP ON COGNITIVE INFORMATION PROCESSING, MAY 26–28, 2014, COPENHAGEN, DENMARK

IN THE TWINKLING OF AN EYE: SYNCHRONIZATION OF EEG AND EYE TRACKING BASED ON BLINK SIGNATURES

Per Bækgaard, Michael Kai Petersen and Jakob Eg Larsen

Cognitive Systems, DTU Compute, Building 321,
Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark
{pgba,mkai,jaeg}@dtu.dk

ABSTRACT

ACHIEVING ROBUST ADAPTIVE SYNCHRONIZATION OF MULTIMODAL BIOMETRIC INPUTS: The recent arrival of wireless EEG headsets that enable mobile real-time 3D brain imaging on smartphones, and low cost eye trackers that provide gaze control of tablets, will radically change how biometric sensors might be integrated into next generation user interfaces. In experimental lab settings EEG neuroimaging and eye tracking data are traditionally combined using external triggers to synchronize the signals. However, with biometric sensors increasingly being applied in everyday usage scenarios, there will be a need for solutions providing a continuous alignment of signals. In the present paper we propose using spontaneous eye blinks, as a means to achieve near real-time synchronization of EEG and eye tracking. Analyzing key parameters that define eye blink signatures across the two domains, we outline a probability function based algorithm to correlate the signals. Comparing the accuracy of the method against a state of the art *EYE-EEG* plug-in for offline analysis of EEG and eye tracking data, we propose our approach could be applied for robust synchronization of biometric sensor data collected in a mobile context.

Index Terms— EEG, Eye Tracking, Synchronization

1. INTRODUCTION

As consumer grade wireless EEG headsets are becoming available which turn smartphones into brain scanners enabling real-time 3D neuroimaging [1][2], and with low cost mobile eye tracking sensors integrated into smartphones and tablets, there will be an growing need for synchronization of multiple sensors on mobile devices. EEG neuroimaging and eye tracking data are traditionally combined by using external triggers to synchronize and accurately timestamp signals in standard experimental setups [3]. While this can be extended to mobile environments [4], it may not always be feasible for e.g. continuous recordings from multiple sensors.

Although mobile neuroimaging and eye tracking data may in terms of causality reflect common underlying events, they are often recorded at different resolutions. To further complicate matters, consumer-grade equipment might not perform as reliably as medical equipment. Packets of data could be lost in transmission or sampling rates may drift over time, especially taking into consideration the varying conditions in a mobile context. In order to interpret signals from multiple modalities, they need to be synchronized. For mobile applications it is desirable to synchronize without the overhead of introducing additional signals. Both, to simplify the hardware setup as well as avoiding dedicating channel resources for this purpose. A challenge is to identify a robust and adaptive method to synchronize signals, as a first step towards continuous recording and aligning of data from multiple biometric sensors in a mobile context.

2. METHOD

When several inputs within the same modality need to be aligned, as when synchronizing multiple audio recordings from different microphones in a video production, a related and well-known problem occurs. This can be solved by extracting key signatures (“audio-fingerprints”) from the signal [5]. Aligning these signatures proves to be more robust and effective than a naïve cross-correlation of the corresponding raw audio signals. However, in the present case, we are looking at vastly different signals retrieved from two modalities: An EEG signal frequently consists of 16 to 256 channels recorded at sampling rates typically between 128Hz to 2048Hz, and data from a visual Eye Tracker often reports (x, y) coordinates for the fixation and saccades plus estimated left/right pupil sizes at different sample rates. Yet other types of biometric sensor data might at a later stage include heart rate measurements or a one-dimensional measurement of skin conductance at a low sample rate. It is therefore proposed to identify *signatures* within each modality that arise from the same underlying set of *causal events* and to correlate these

Acknowledgment: This work is supported in part by The Danish National Advanced Technology Foundation through the project Eye Tracking for Mobile Devices

to align the multimodal signals. In order to assure a robust synchronization in a mobile context, we furthermore need to take into account sampling rate drifts and loss of data.

3. IDENTIFICATION OF COMMON SIGNALS

Blinks reveal a lot of information about subject alertness [6] and mental state/type of activity [7], but are discarded as unwarranted artifacts in EEG recordings as they may overshadow the signals of interest related to neural processing.

However, blinks deserve more than being discarded, and in this case our focus is on reusing spontaneous blinks retrieved from both EEG and Eye Tracking data as a common causal event serving to synchronize the two domains.

A mean blink frequency around 16 min^{-1} is typically found in healthy and alert adults with a mean duration of around 200 ms, where the reopening time is around twice the closing time [6] . Blink times can regularly increase to above 400 ms when alertness levels drop, and especially the proportion of long closure duration blinks and the reopening time seems to be affected by drowsyness [6].

Blinks tend to be dependent on whether the subjects are involved in active tasks or simply passively observing objects, leading to large individual variations. However, even for the least blinking subjects involved in active tasks, spontaneous blinks occur naturally multiple times per minute [7].

V. Luca and M. Swirski [private correspondence] have preliminarily found that with consumer grade EEG and Eye Tracking equipment, blinks may be less easy to identify with high confidence. It may thus be worth prioritizing a high specificity in favor of a high sensitivity in blink detection, in order to reduce false positives.

Hence, our focus will be on blink durations in the range of $50 - 500$ ms, with closing times < 150 ms in order "to exclude so-called nonblink closures according to the criteria of spontaneous eye blink defined by Stern et al. [8][9]".

3.1. Blink signatures in the EEG

Blinks appear as a strong and common signature in the EEG. Their signatures have therefore been extensively analyzed in the literature, usually with the intent of being able to extract and remove these unwanted artifacts [10][11][12][13]. Several plug-ins to EEG Lab [14] like Eye Catch[1] automatically locate eye related ICA independent components.

Two classes of voltage generators contribute to the EEG: dipole and eyelid movements [13]. The actual movement of the eyelid is the very essence of a blink; thus all blinks contain a strong signature in the EEG caused by this voltage generator. Particularly the frontal central EEG electrode sites, such as the FPz, FP1 and FP2 and to some extent in AFz, AF7 and AF8, will frequently pick up this strong signature. Average blink amplitudes can be above 200 uV at supraorbital electrodes, and a strong average blink related signal may be back-projected to FPz [15]. Blink related signatures collected from the electrooculogram (EOG) electrodes placed near the eyes peak above 20 uV, and are thus significantly stronger than EEG voltage differences typically in the range of a few uV [13]. This signature of the blink is likely caused by the eyelids acting as a sliding potential source, but muscle electromyographic activity (EMG) can also be expected to be superimposed. See Fig. 1

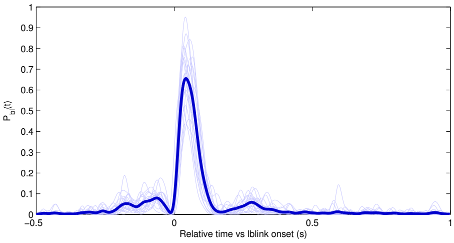


Fig. 1. Typical EEG blink signatures shown as normalized power in the electrodes FPz and AFz for 21 blinks in [16] epoched to when the eye tracker sees the left eye closing. The mean value is shown darker. A faster closing time is seen when compared to the reopening time. The eyelid movement signal is prominently present, but additional signals are visible before and after, most likely caused by EMG muscle activity.

Generally, the function $P_{mod,ev}(t)$ is introduced as an estimate proportional to the probability that event ev has occurred at time t in modality mod. This function is derived from each modality independently, taking into account the archetypical signature of the causal event ev seen in that modality mod. This need not imply specific training on a dataset, although the estimate might be improved by wavelet-identification or supervised machine learning techniques.

Removal of EEG artifacts may often be facilitated by Independent Component Analysis (ICA), however, less gain is found when applying ICA for *detection* of eye related muscle activity [12]. Further, as the eyelid movement signature is strong, *the normalized electrical power summed over selected key electrodes will henceforth serve as a rough approximation of the blink probability function*, $P_{EEG,bl}(t)$. To further improve detection, we will consider only power within specific frequency bands, and apply a band-pass filter between 2 Hz and 20 Hz, corresponding to blink durations in the range of $50 - 500$ ms. No prior extraction of archetypical signatures is done here.

Events such as eye blinks are typically of a certain duration. The corresponding archetypical *signatures* must be aligned according to a common *event* reference point across modalities. If the duration of an event is of varying length,

[1] http://sccn.ucsd.edu/wiki/EyeCatch (accessed: 2014-02-10)

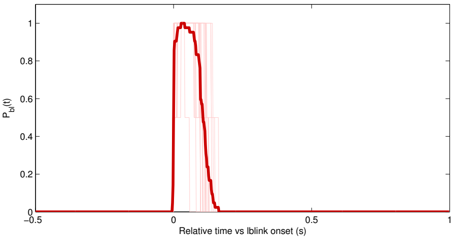


Fig. 2. The epoched blink probability function from the same 21 blinks as in Fig. 1. The darker curve denotes the mean value.

it is possible to introduce two or more probability functions in the related modalities, $P_{mod,ev0}(t)$ and $P_{mod,ev1}(t)$ signifying the probability of the beginning and the end of the event respectively, if it is more feasible than trying to identify a common mid reference-point for the events; synchronization can then be based on either or both.

Even though blinks can be considered as having three phases: 1) closing of the eyelid, 2) the closed phase, and 3) reopening of the eyelid, longer blinks primarily have a longer opening phase [6]. This might imply that the closing phase is the most well-defined, and it could alternatively be suggested to base the synchronization on the probability function of the closing phase $P_{EEG,bl0}(t)$ instead of the entire blink, $P_{EEG,bl}(t)$.

3.2. Blink signatures in Eye Tracking

Visual eye tracking equipment relies on tracing the movements of the pupil relative to reflections of the cornea. Additionally the dynamically changing size of the pupil is measured, and it is registered when the gaze no longer can be traced due to either a blink or the user diverts his attention outside the eye trackers field of view. Most modern visual eye trackers, explicitly record these events in the output data. This can be used directly for estimating the appearance of a blink.

Loss of pupil tracking in the interval of $50 - 500\ ms$ is thus taken to be indicative of a spontaneous eye blink. *It is suggested to base the blink probability function on whether the pupils are visible or not*, filtering out appearances that are longer or shorter than the indicated interval. Furthermore, using the initial, most well-defined part of the blink (the eye closing) as key to synchronization might mitigate any potential detrimental effects of equipment related reacquisition timing once the pupil becomes visible again. See also Fig. 2.

It is assumed that the stated timing constraint will filter out many false positives. If needed, blink detection could be further improved by also detecting and filtering out obvious loss of pupil tracking cases that arise from the user gazing outside of the field of view, due to the preceding gaze tracking coordinates.

3.3. Correlating Different Signatures

In the naïve case, all probability functions $P_{mod,bl}(t)$ are first converted to the same sample rate, and a cross-correlation of the signals can then derive the inter-signal delay that will make them appear in sync.

This is done by calculating the cross correlation between the two probability function time series of length M and N, where typically $M = N$ after sample rate conversion:

$$\phi_{EEG,ET}(d) = \sum_{j=max(0,d)}^{min(M-1+d,N-1)} P_{EEG,bl}(j-d)P_{ET,bl}(j)$$

and determining where $\phi(d)$ has its global maximum, at which the corresponding d value is an estimate of the delay between the time series representing the EEG and ET values, respectively. $\phi(d)$ need not be normalized for this purpose.

For longer time series, and for real time use, a sliding window should be applied to filter out older values. This makes it possible to estimate the current inter-signal delay, and by relating that to previous delay values, it can be used to compensate adaptively for data loss and drift. In the simplest possible case, at time t, the current delay, $d(t)$, can be estimated by cross correlating the functions $P_{EEG,bl}(t)$ and $P_{ET,bl}(t)$ in the window $t - w$ to t and deriving the actual delay as above. By maintaining previously calculated delay values $d(t_1), d(t_2), d(t_3), ...$ in overlapping windows, a linear regression can be used to estimate the drift rate α and the stability of the fit (the regression coefficient).

$$d(t) = \alpha t + \beta$$

(β is the initial delta at time $t = 0$). It is advisable to use a piecewise linear regression to maintain estimates of the inter-signal delay, to compensate for data loss.

The cross correlation between the probability functions will reveal information about the goodness of the fit: The width of the global maximum peak depends both on how well the estimated signatures match and on stochastic variations. Further, once drift rates have been estimated, the variance of current delay estimates vs. the predicted value gives an indication of how well the synchronization can be maintained. Any discrepancy vs. "ground truth" is however not immediately observable.

4. RESULTS

Fig. 3 to 9 details an analysis of the "Natural Reading" Dataset 3 [16]. This dataset was synchronized by an external trigger, so the "ground truth" is known. Data was recorded at

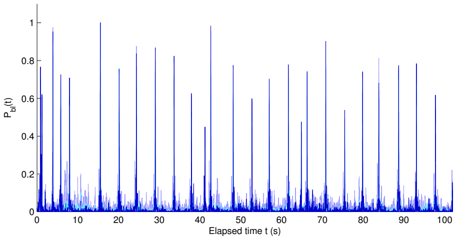


Fig. 3. Blink probability, $P_{EEG,bl}(t)$, as estimated by normalized power in selected EEG channels. Blue denotes the FPz and AFz electrodes, cyan the FPz, FP1, FP2, AFz, AF7 and AF8, whereas light blue are all 72 electrodes. Fewer electrodes gives cleaner signals but might be less robust.

512 Hz sample rate for the EEG and at 1000 Hz for the Eye Tracker and has been synchronized in "Eye EEG" in the MatLab "EEG Lab" toolbox to an actual delay in the resampled signals of 0 ms.

Initially, the full blink signature was used for synchronization, i.e. the EEG probability function $P_{EEG,bl}$ was directly based on the power of the FPz and AFz electrodes, as this provided the cleanest signal (Fig. 3). The delay was estimated by correlating this with the corresponding ET probability function (Fig. 4 [Top]). The estimated delay can be derived from Fig. 5. The stability of the estimate is illustrated by Fig. 6.

As an alternative, the blink onset was used for synchronization. Here, the EEG probability function $P_{EEG,bl0}$ was based on the results of cross correlating the first $\approx 50\ ms$ of the mean EEG blink signature with the power of the FPz and AP electrodes (Fig. 7); the ET probability function $P_{ET,bl0}$ is found similarly (Fig. 8). The delay was again estimated by doing another cross correlation between the corresponding EEG and ET probability functions. A comparison of the two methods is shown in Fig. 9.

5. DISCUSSION AND FURTHER WORK

Based on the sample data analyzed above [16] the proposed algorithmic method works satisfactorily. For the full blink synchronization, the determined delay was $3.9\ ms$ vs. the correct value of $0\ ms$. For this dataset, synchronization could thus have been achieved with an apparent error equivalent to only two samples at $512\ Hz$, based on the full blink signature, without any machine learning. After the first blink, at $0.47\ s$ relative to the shown period, the estimated delay fluctuates between only 3 different sample values at $512\ Hz$ as more blinks are analyzed (Fig. 6). This may hint that the stability of the estimated delay is not strongly impacted by the length of the period analyzed, as long as there are blinks to include, although the actual estimate may improve with multiple blinks in the presence of more noise than here. However, as Figures 5, 6 and 9 shows, the peak of the correlated EEG and ET probability functions is $\approx 32\ ms$ wide. This might indicate that the uncertainty of the synchronization here is larger than the $\approx 4\ ms$ apparent precision seen.

Synchronizing at blink onset did apparently not improve the accuracy. As it causes a further computational load it cannot be recommended based on the present analyzed dataset. However, additional datasets with more stochastic variation and noise than the laboratory sample data used in this study

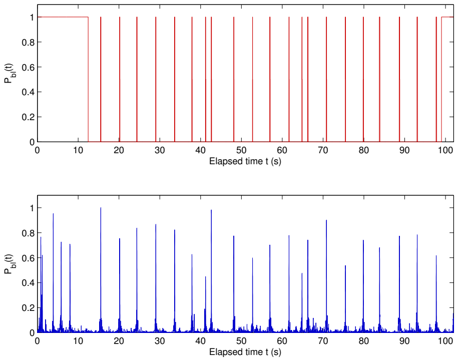


Fig. 4. [Top] The blink probability function, $P_{ET,bl}(t)$ in red, as determined by data from the eye tracker: When the pupil is lost for either eye, the probability is set to 0.5, and when both pupils are lost, it becomes 1. To the same time scale [bottom] is shown the same blue curve as in Fig. 3. Note that the initial and final eye tracking data have been "blanked" during the manual synchronization.

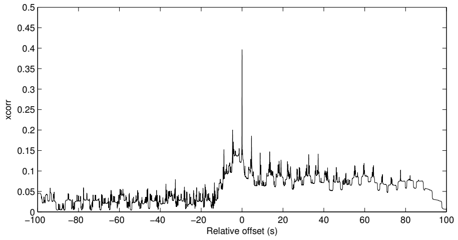


Fig. 5. A naïve cross correlation between the FPz/AFz-based EEG blink probability function $P_{EEG,bl}(t)$ and the ET blink probability function $P_{ET,bl}(t)$. A clear peak denotes the estimated delta time between the two at $3.9\ ms$ vs a "ground truth" of $0\ ms$. The width of the peak at 95% is $\approx 32\ ms$.

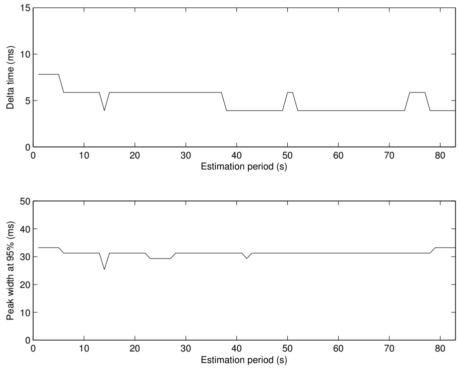


Fig. 6. [Top] shows the estimated delta between the EEG and ET time series as a function of how long a period has been analyzed. [Bottom] shows the width of the cross correlation peak at 95% of the peak value. It remains at similar levels throughout the interval.

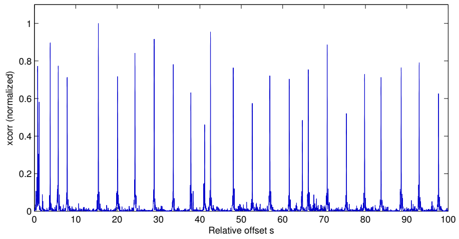


Fig. 7. Blink onset probability, $P_{EEG,bl0}(t)$, as estimated by cross correlating the power in the FPz and AFz electrodes with the blink onset signature (sampled from the mean EEG blink signature in Fig. 1 in the inverval $-15\ ms$ to $50\ ms$). A similar time series will appear from cross correlating the ET time series with the start of the ET blink onset signature from Fig. 2 or with a similar looking heaviside step function.

need to be analyzed in order to verify the stability and accuracy of the method, when based on data collected in a mobile context. The uncertainty across larger sample sets needs to be better determined, and more accurate derivations of blink probability functions (possibly based on the suggested techniques) might be required to improve upon the results.

The above outlined approach to synchronize EEG and eye tracking could potentially allow us to move outside artificial laboratory settings and instead monitor brain activity over extended periods of time under realistic conditions in an everyday context [17]. Rather than analyzing event related responses elicited a few hundred milliseconds after images or text appear on a computer screen, the ability to combine eye tracking with neuroimaging on mobile devices, might provide a foundation for continuous EEG recordings based on fixation related potentials (FRP) [3]. That is, the sequences of rapid saccades captured by the eye tracker when shifting our gaze in the direction of something that capture our attention, may be driven bottom-up by elements of visual perception as well as top-down aspects of intended action. The ensuing fixations, typically lasting a few hundred milliseconds, might be treated as virtual time stamps, corresponding to the onset of visual processing when looking at objects [18]. This enables capturing EEG event related responses in real life scenarios [19], synchronized based on eye tracking derived patterns of saccades and fixations [3]. Combining additional sensors like heart rate correlated with eye tracking measurement of pupil dilation, and EEG time series components might likewise improve the classification of emotional responses [20]. The pro-

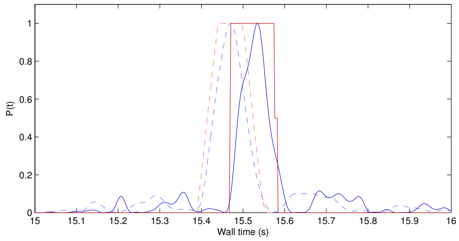


Fig. 8. All the probability functions for a sample blink (at $15.47\ s$), EEG in blue and ET in red. Blink onsets are dashed, full blinks solid. The EEG blink onset $P_{EEG,bl0}(t)$ (blue dashed) is centered on the beginning of the blink as reported by the eye tracker $P_{EEG,bl}(t)$ (red solid). Note also that the blue EEG functions are centered within the red ET functions.

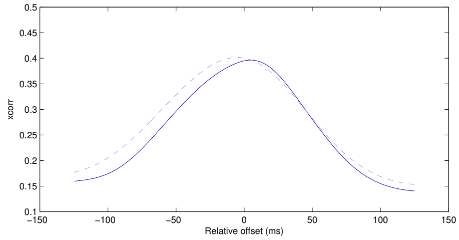


Fig. 9. Comparing correlated EEG and ET probability functions using full blinks (solid) vs. blink onset (dashed). The blink onset peak is at $-3.9\ ms$ vs. the full blink at $3.9\ ms$. Both have similar width at 95% peak value.

posed framework for aligning EEG and eye tracking based on blink signatures might therefore in a larger context enhance interpretation of inputs from multiple complementary sensors, and thereby facilitate designing next generation cognitive interfaces that dynamically reflect the state of the user.

6. REFERENCES

[1] Arkadiusz Stopczynski, Carsten Stahlhut, Michael Kai Petersen, Jakob Eg Larsen, Camilla Falk Jensen, Marieta Georgieva Ivanova, Tobias S. Andersen, and Lars Kai Hansen, "Smartphones as pocketable labs: Visions for mobile brain imaging and neurofeedback," *Int. J. of Psychophysiology*, vol. 91, no. 1, pp. 54 – 66, 2014, Towards mobile {EEG}.

[2] Arkadiusz Stopczynski, Carsten Stahlhut, Jakob Eg Larsen, Michael Kai Petersen, and Lars Kai Hansen, "The smartphone brain scanner: A portable real-time neuroimaging system," *PLOS ONE*, vol. 9, no. 2, pp. e86733, 2014.

[3] Olaf Dimigen, Werner Sommer, Annette Hohlfeld, Arthur M Jacobs, and Reinhold Kliegl, "Coregistration of eye movements and eeg in natural reading: analyses and review.," *J. of Experimental Psychology: General*, vol. 140, no. 4, pp. 552, 2011.

[4] Klaus Gramann, Joseph T Gwin, Daniel P Ferris, Kelvin Oie, Tzyy-Ping Jung, Chin-Teng Lin, Lun-De Liao, and Scott Makeig, "Cognition in action: imaging brain/body dynamics in mobile humans," *Reviews in the Neurosciences*, vol. 22, no. 6, pp. 593–608, 2011.

[5] Prarthana Shrstha, Mauro Barbieri, and Hans Weda, "Synchronization of multi-camera video recordings based on audio," *ACM Multimedia*, pp. 545–548, 2007.

[6] Philipp P. Caffier, Udo Erdmann, and Peter Ullsperger, "Experimental evaluation of eye-blink parameters as a drowsiness measure," *European J. of Applied Physiology*, vol. 89, no. 3-4, pp. 319–325, 2013.

[7] J.H. Skotte, J.K. Nøjgaard, L.V. Jørgensen, K.B. Christensen, and G. Sjøgaard, "Eye blink frequency during different computer tasks quantified by electrooculography," *European J. of Applied Physiology*, vol. 99, no. 2, pp. 113–119, 2007.

[8] John A Stern, Larry C Walrath, and Robert Goldstein, "The endogenous eyeblink," *Psychophysiology*, vol. 21, no. 1, pp. 22–33, 1984.

[9] John A Stern, Donna J Boyer, David J Schroeder, RM Touchstone, and N Stoliarov, "Blinks, saccades; and fixation pauses during vigilance task performance: 2," Tech. Rep., DTIC Document, 1996.

[10] Tzyy-Ping Jung, Scott Makeig, Marissa Westerfield, Jeanne Townsend, Eric Courchesne, and Terrence J. Sejnowski, "Removal of eye activity artifacts from visual event-related potentials in normal and clinical subjects," *Clinical Neurophysiology*, vol. 111, no. 10, pp. 1745 – 1758, 2000.

[11] Ricardo Nuno Vig rio, "Extraction of ocular artefacts from {EEG} using independent component analysis," *Electroencephalography and Clinical Neurophysiology*, vol. 103, no. 3, pp. 395 – 404, 1997.

[12] Arnaud Delorme, Terrence Sejnowski, and Scott Makeig, "Enhanced detection of artifacts in {EEG} data using higher-order statistics and independent component analysis," *NeuroImage*, vol. 34, no. 4, pp. 1443 – 1449, 2007.

[13] R.J. Croft and R.J. Barry, "Removal of ocular artifact from the eeg: a review," *Neurophysiologie Clinique/Clinical Neurophysiology*, vol. 30, no. 1, pp. 5 – 19, 2000.

[14] Arnaud Delorme and Scott Makeig, "Eeglab: an open source toolbox for analysis of single-trial {EEG} dynamics including independent component analysis," *J. of Neuroscience Methods*, vol. 134, no. 1, pp. 9 – 21, 2004.

[15] Sven Hoffmann and Michael Falkenstein, "The correction of eye blink artefacts in the eeg: A comparison of two prominent methods," *PLoS ONE*, vol. 3, no. 8, pp. e3004, 08 2008.

[16] Humboldt-Universität, "Natural reading, dataset 3," http://www2.hu-berlin.de/eyelab/testdata.html, 2013.

[17] Scott Makeig, Klaus Gramann, Tzyy-Ping Jung, Terrence J. Sejnowski, and Howard Poizner, "Linking brain, mind and behavior," *Int. J. of Psychophysiology*, vol. 73, no. 2, pp. 95 – 100, 2009, Neural Processes in Clinical Psychophysiology.

[18] Aga Bojko, *Eye Tracking the User Experience - a practical guide to research*, Rosenfeld Media, 2013.

[19] Maarten De Vos and Stefan Debener, "Mobile eeg: Towards brain activity monitoring during natural action and cognition," *Int. J. of Psychophysiology*, vol. 91, no. 1, pp. 1 – 2, 2014, Towards mobile {EEG}.

[20] Peter J Lang and Margaret M Bradley, "Emotion and the motivational brain," *Biological psychology*, vol. 84, no. 3, pp. 437–450, 2010.

Paper B

Thinking Outside of the Box or Enjoying Your 2 Seconds of Frame

This paper was presented by the author at the HCII2015 conference in Los Angeles USA, July 2015.

The trip was made possible thanks to Otto Mønsted Fonden.

Thinking Outside of the Box or Enjoying Your 2 Seconds of Frame?

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 Kgs. Lyngby, Denmark
{pgba,mkai,jaeg}@dtu.dk

Abstract. The emergence of low cost eye tracking devices will make QS quantified self monitoring of eye movements attainable on next generation mobile devices, potentially allowing us to infer reactions related to fatigue or emotional responses on a continuous basis when interacting with the screens of smartphones and tablets. In the current study we explore whether consumer grade eye trackers, despite their reduced spatio-temporal resolution, are able to monitor fixations as well as frequencies of saccades and blinks that may characterize aspects of attention, and identify consistent individual patterns that may be modulated by our overall level of engagement.

Keywords: Eye tracking · Fixation Density Maps · Fixation duration

1 Introduction

Although we may visually perceive whatever attracts our attention as a static entity, our eyes process information from short fixations characterized by foveal acuity interspersed by rapid saccadic eye movements. Essentially the oculomotor plant can be modeled as a dynamic system which through contraction of muscles and elastic tendons continuously pulls and rotates the eye globe when directing our gaze towards an area of interest. These muscles are controlled by neural signals modulated by areas in the brain responsible for constantly engaging and relocating our visual attention [12], which at the same time reflect our cognitive state as fatigue and demanding tasks have been found to lower saccadic velocity, whereas higher peak saccadic velocity indicates increasing arousal [7]. It is also known that the frequency and duration of blinks can be indicative of fatigue or time-on-task [8,9,16,17]. However, fixation density patterns and saccadic movements differ highly across individuals [13,15] yet remain stable over a variety of viewing conditions due to systematic endogenous factors [3,10]. We would therefore expect that our current eye tracking study might reflect both unique individual traits as well as variations in fixation patterns due to varying levels of engagement. Applying a low cost eye tracker running at a low resolution [6], we explore whether we are able to distinguish stable individual characteristics when viewing suddenly appearing contrasting visual stimuli with uniform

M. Antona and C. Stephanidis (Eds.): UAHCI 2015, Part II, LNCS 9176, pp. 186–195, 2015.
DOI: 10.1007/978-3-319-20681-3_17

baseline metrics [5]. The experiment is repeated over a week in order to analyze how eye tracking parameters related to fixation density maps may vary due to changing levels of perceived fatigue.

2 Experiment

Being amongst the first eye-self-trackers, this experiment explores whether we can identify individual signatures reflecting levels of attention in eye tracking

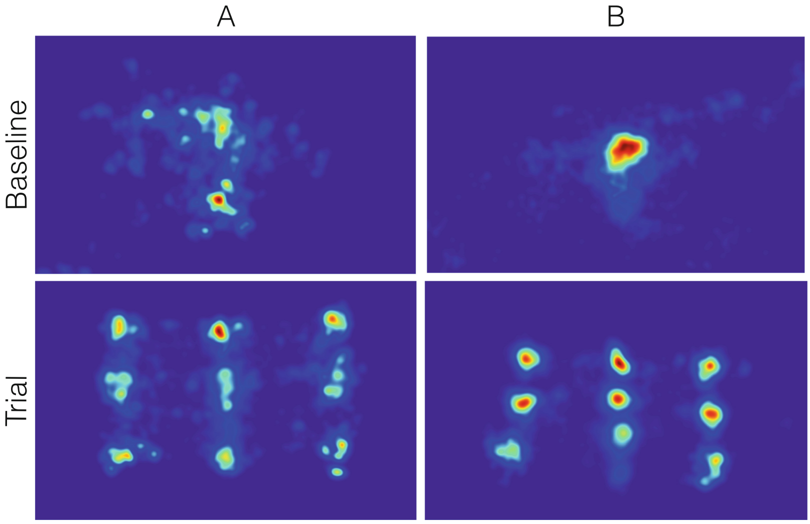


Fig. 1. Typical fixation density Maps, for A (left column) and B (right column), of fixations in the baseline examples (upper row) when observing solid colors only, and in trial (lower row) when the colored squares are presented against the complementary color background.

The Trial Fixation Density Maps (lower row) reflect the position of the visual stimuli, nevertheless there are clear differences between the test persons; B has a higher tendency to maintain focus within the frame of the squares. A appear less focused on the frame and is rather *thinking outside of the box*, while overall fixations appear less dense in the middle horizontal versus the lower and upper horizontal rows. Likewise for B the central square in the lowest horizontal row shows a larger spread and overall this row reflects a less dense focus, although we cannot rule out the possibility of calibration errors for the eye tracker in the lower screen area.

The Baseline Fixation Density Maps (upper row) depict a higher degree of difference between the subjects. Again, B has a higher tendency to maintain focus towards the center of the screen whereas A shows a tendency to focus at the middle vertical, with fixations skewed towards the left side of the screen. We speculate that this consistent offset for A, rather than being an artifact, could potentially be related to gaze direction rooted in right hemisphere dominance when processing spatial information [2] (Color figure online).

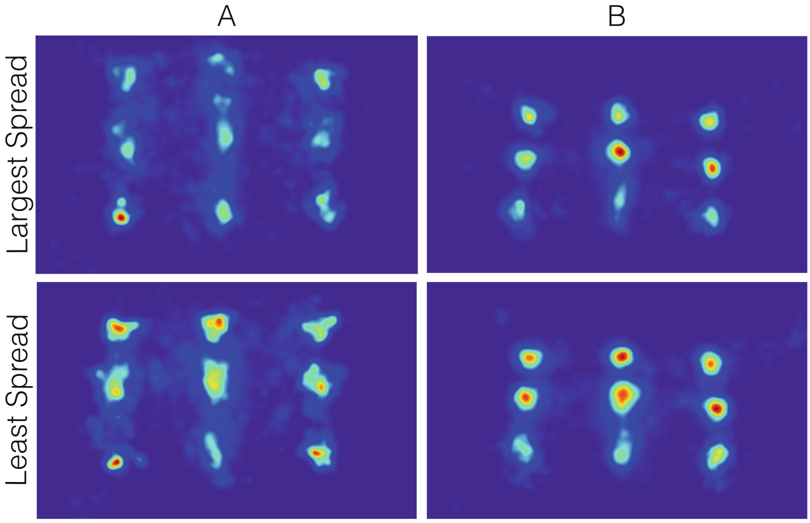


Fig. 2. Variations in fixation density maps (baseline/trial combined). Although there are variations in power within the fixation density maps for A and B over the week, individual differences are discernible, where the upper row shows the largest spread of fixations while the lower row represents more narrowly focused fixations (Color figure online).

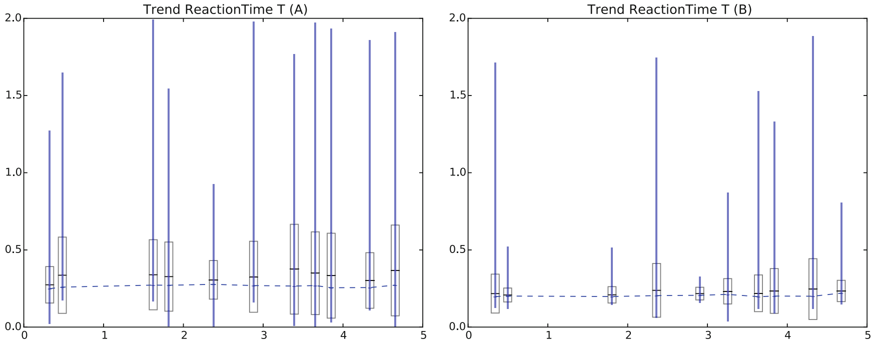


Fig. 3. Differences in *time to target* reaction time when fixating on the presented visual stimuli in trials throughout the week for A (left) and B (right); minimum, mean standard deviation, mean, mean+standard deviation, and maximum. The dashed line indicates the median.

This reaction time is measured from the presentation of the stimuli to the first fixation starts at, or close to the presented, square. This includes the reactive saccade between points. The saccade time cannot be accurately determined due to the 60 Hz sampling frequency of Eye Tribe tracker, but is estimated to be 30–50 ms. Fixations typically jump to adjacent positions in space, so the variation in distance is not large, as can be seen. The reaction time median, which best filters out any noise and accidental mis-calibrations, remains remarkably consistent throughout the entire week, and clearly differs between the test persons at around ~269 ms vs ~201 ms.

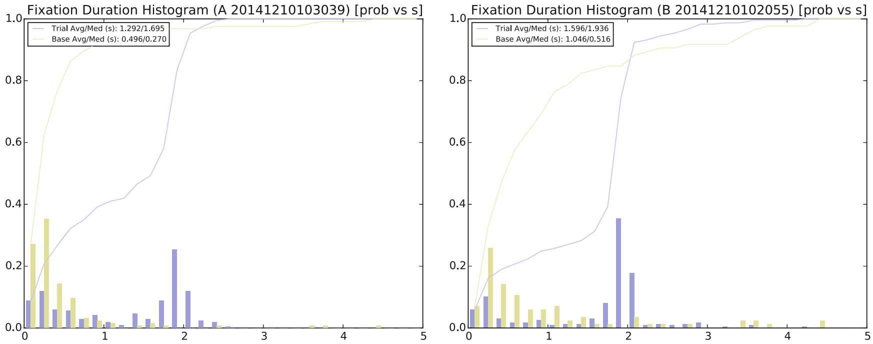


Fig. 4. Fixation duration histograms (bars) and cummulative histograms (lines), for both A (left) and B (right) during an experiment.

Fixation duration appears to be stimuli dependent with, in this case, a median time of 1.695 s vs 0.270 s for person A in Trials (blue) vs Baseline (yellow) and 1.936 s vs 0.516 s for person B. This indicates consistent differences in A and B's fixation durations. This stimuli-dependent difference when attending to the presented squares versus the solid color backgrounds is not only observed in fixation durations, but also to some extent in e.g. saccade frequencies and fixation patterns. No dependency on color of the presented squares were observed, despite the large self reported perceived differences related to the extreme complementary color contrasts such as green squares on top of a red background or yellow squares presented against a blue background.

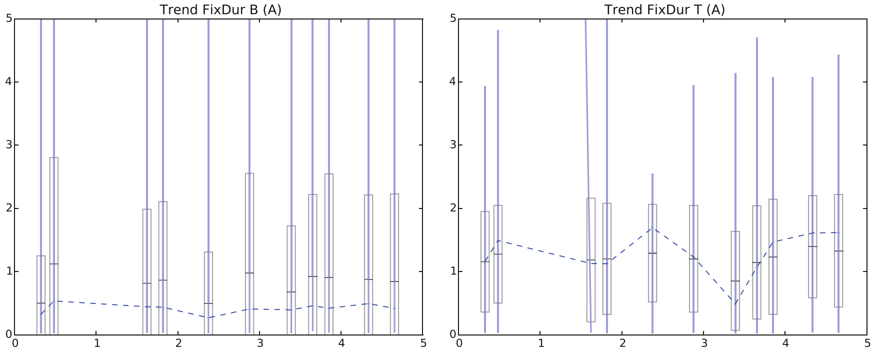


Fig. 5. Variations of the Fixation Duration for A across all experiments in the entire week. The Baseline fixation length when observing solid background colors shows less variation than when attending to the presented complementary colored squares.

data. During the experiment, eye tracking data has typically been collected twice a day over a week, each consisting of 24 trials where 8 colored squares (∼3 degrees wide) are sequentially presented on the screen, alternating between the

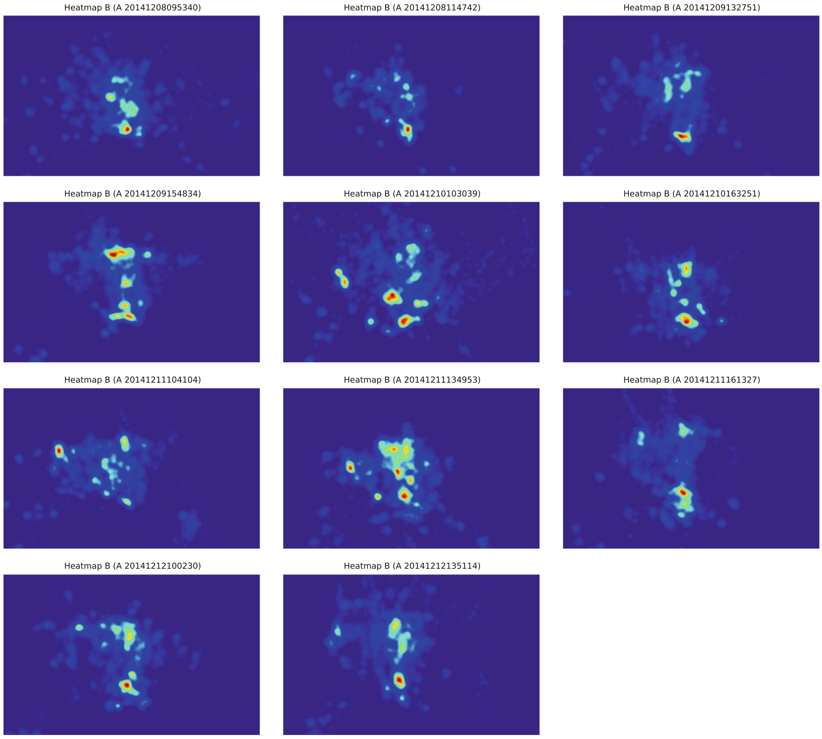


Fig. 6. Variations of the baseline fixation density maps for A plotted over the entire week (beginning of the week at the top left corner; read left-right first).

Compared to Fig. 7, there are visible differences between A and B: B's Baseline fixations are less spread out and appear mainly towards the middle of the screen, whereas A's Baseline fixations have a tendency to be skewed towards the left side of the screen and with a significantly larger spread.

Variations over the week are also visible although the general pattern for each individual appear consistent from experiment to experiment.

colors blue, yellow, green, yellow, white and black. Each presented square appeared for 2 s against their complementary color as screen background (referred to as *Trial* conditions), followed by 4 s of solid complementary color (referred to as *Baseline* conditions). In total, this constitutes 480 secs of visual stimuli for each of 11 experiments performed over a week.

After an initial calibration at the beginning of each experiment, stimuli was presented on a conventional MacBook Pro 13" in an ordinary office environment, running PsychoPy software [14]. The Eye Tribe mobile eye tracking device, connected via USB, retrieved the eye tracking data through the associated API [18],

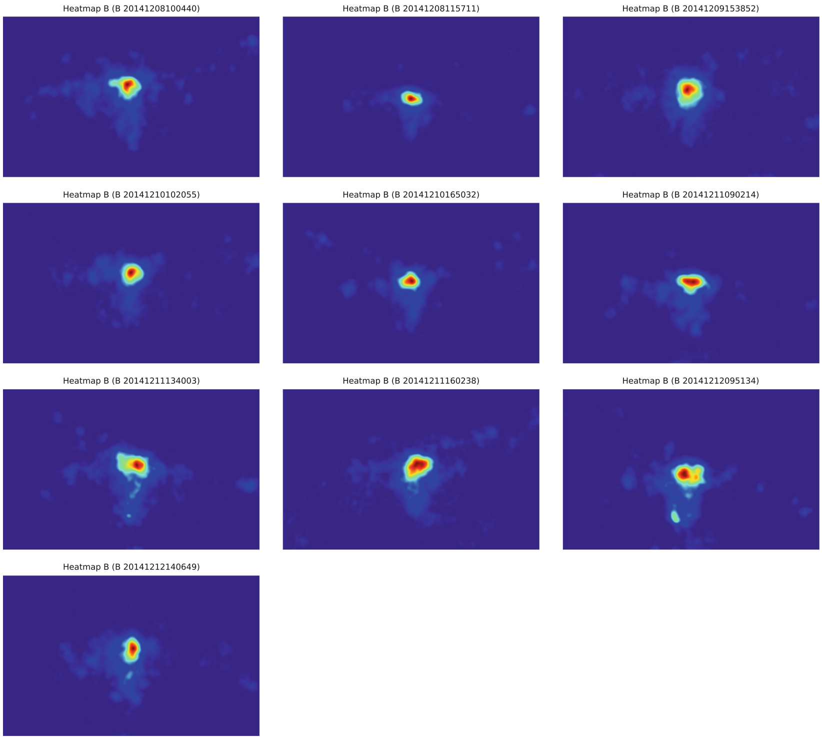


Fig. 7. Variations of the baseline fixation density maps for B plotted over the entire week (beginning of the week at the top left corner; read left-right first).

Compared to Fig. 6, there are visible differences between A and B: B's Baseline fixations are less spread out and appear mainly towards the middle of the screen, whereas A's Baseline fixations have a tendency to be skewed towards the left side of the screen and with a significantly larger spread.

Variations over the week are also visible although the general pattern for each individual appear consistent from experiment to experiment.

using PeyeTribe [1]. Subsequently a density based clustering approach to define fixations was applied. Two right-handed subjects (males, average age 55) participated in the experiments and were not instructed to follow any specific viewing patterns.

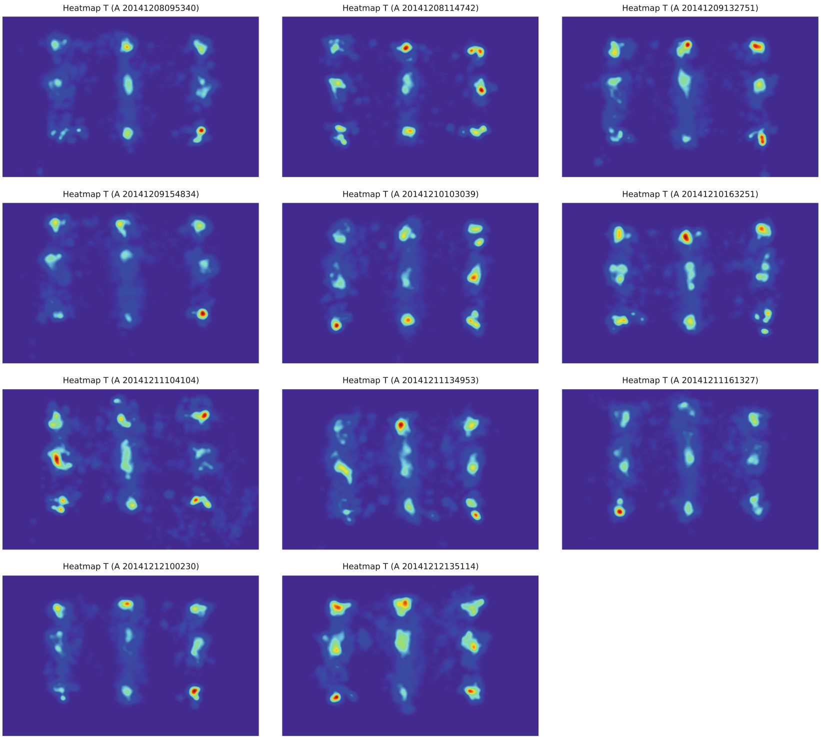


Fig. 8. Variations of the combined trial fixation density maps for A plotted over the entire week (beginning of the week at the top left corner; read left-right first).

Compared to Fig. 9 there are similar differences in these stimuli-driven Trial Fixation Density Maps as there are for the Baseline ones, although they are less pronounced. B has again less spread-out fixations compared to A, and it's likely that any random A and B Fixation Density Map could be compared and classified as belonging to either A or B based on their looks.

3 Results

Below, Figs. 1 and 2 compare typical Fixation Density Maps in Trial and Baseline for A and B, and variations in spread of the combined Fixation Density Maps. Figures 3, 4 and 5 show *time to target* reaction times and fixation durations. Figures 6, 7, 8 and 9 illustrate variations in the Fixation Density Maps for the entire week of experiments.

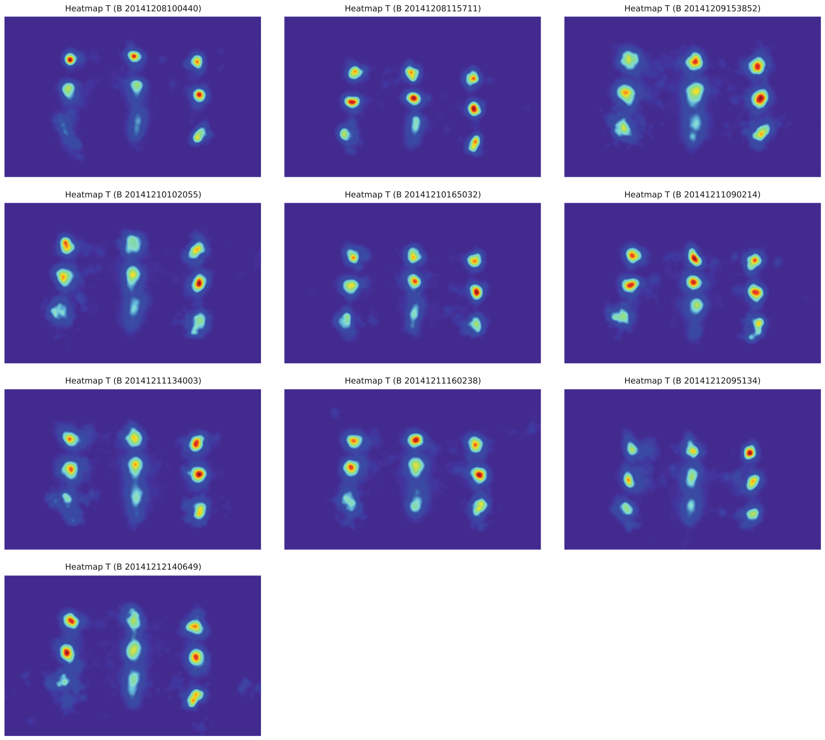


Fig. 9. Variations of the combined trial fixation density maps for B plotted over the entire week (beginning of the week at the top left corner; read left-right first).

Compared to Fig. 8 there are similar differences in these stimuli-driven Trial Fixation Density Maps as there are for the Baseline ones, although they are less pronounced. B has again less spread-out fixations compared to A, and it's likely that any random A and B Fixation Density Map could be compared and classified as belonging to either A or B based on their looks.

4 Conclusion

While the *time to target* reaction time, reaching first fixation on the presented visual stimuli, differentiates subject A from B, this eye tracking measure nevertheless appears constant within the two subjects during the whole week. This is not inconsistent with the reported findings of Wang and Stern [19], as cited by [17], which indicates *reactive saccades* to be invariant of time-on-task. It thus seems to reflect a personal signature neither affected by training nor the differing complementary color contrasts of the presented stimuli in the experiments, whereas the spread and length of fixations in response to the presented colored squares varies within subjects A and B during the experiments over the week.

Recent eye tracking studies indicate extended fixation duration time in subjects reporting feeling fatigued at non-optimal periods during the day related to their circadian rhythm [4], whereas shorter gaze duration has been found in eye tracking experiments when subjects read emotionally positive versus neutral words [11].

We initially hypothesized that the fixations on the presented visual targets would likely be more focused in the morning, compared to experiments performed in the afternoon where the subjects might presumably be feeling more tired, but this seems not to be the case.

During some of the morning experiments which resulted in less focused fixations the subjects actually reported that they felt more fresh and alert. Correspondingly, some of the most dense fixations on targets were actually recorded late in the afternoon for both subjects, raising an intriguing question as to whether the wider distribution of the fixations in the Fixation Density Map is correlated with the level of engagement of the subjects, or merely reflects a less agile focus that might be inversely related to the perceived fatigue, as reported by the subjects in some of the experiments.

Although the present study is clearly limited by the number of participants and the duration of the experiments, we find that these questions merit exploration in order to enable a continuous quantified self estimation of our changing levels of attention and fatigue.

Acknowledgment. This work is supported in part by the Innovation Fund Denmark through the project Eye Tracking for Mobile Devices.

References

1. Bækgaard, P.: Simple python interface to the eye tribe eye tracker. http://github.com/baekgaard/peyetribe/ (Accessed: 17 February 2015)
2. Carlei, C., Kerzel, D.: Gaze direction affects visuo-spatial short-term memory. Brain Cogn. **90**, 63–68 (2014)
3. Castelhano, M.S., Henderson, J.M.: Stable individual differences across images in human saccadic eye movements. Can. J. Exp. Psychol./Rev. Can. Psychol. Expérimentale **62**(1), 1 (2008)
4. Cazzoli, D., Antoniades, C.A., Kennard, C., Nyffeler, T., Bassetti, C.L., Müri, R.M.: Eye movements discriminate fatigue due to chronotypical factors and time spent on task-a double dissociation. PloS ONE **9**(1), e87146 (2014)
5. Corbetta, M., Shulman, G.L.: Control of goal-directed and stimulus-driven attention in the brain. Nat. Rev. Neurosci. **3**(3), 201–215 (2002)
6. Dalmaijer, E.: Is the low-cost eyetribe eye tracker any good for research?. Technical report. PeerJ PrePrints (2014)
7. Di Stasi, L.L., Catena, A., Canas, J.J., Macknik, S.L., Martinez-Conde, S.: Saccadic velocity as an arousal index in naturalistic tasks. Neurosci. Biobehav. Rev. **37**(5), 968–975 (2013)
8. Dodge, R.: The laws of relative fatigue. Psychol. Rev. **24**(2), 89 (1917)
9. Dodge, R., Cline, T.S.: The angle velocity of eye movements. Psychol. Rev. **8**(2), 145 (1901)

10. Kasprowski, P., Ober, J.: Eye movements in biometrics. In: Maltoni, D., Jain, A.K. (eds.) BioAW 2004. LNCS, vol. 3087, pp. 248–258. Springer, Heidelberg (2004)
11. Knickerbocker, H., Johnson, R.L., Altarriba, J.: Emotion effects during reading: influence of an emotion target word on eye movements and processing. Cogn. Emot. **29**(5), 784–806 (2015). doi:10.1080/02699931.2014.938023
12. Komogortsev, O., Holland, C., Karpov, A., Price, L.R.: Biometrics via oculomotor plant characteristics: Impact of parameters in oculomotor plant model. ACM Trans. Appl. Percept. (TAP) **11**(4), 20 (2014)
13. Komogortsev, O.V., Jayarathna, S., Aragon, C.R., Mahmoud, M.: Biometric identification via an oculomotor plant mathematical model. In: Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications, pp. 57–60. ACM (2010)
14. Peirce, J.W.: Psychopy-psychophysics software in python. J. Neurosci. Methods **162**(1), 8–13 (2007)
15. Rigas, I., Komogortsev, O.V.: Biometric recognition via fixation density maps. In: International Society for Optics and Photonics. SPIE Defense+ Security, pp. 90750M–90750M (2014)
16. Schleicher, R., Galley, N., Briest, S., Galley, L.: Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired? Ergonomics **51**(7), 982–1010 (2008)
17. Sirevaag, E.J., Stern, J.A.: Ocular measures of fatigue and cognitive factors. In: Engineering psychophysiology: Issues and Applications, pp. 269–287 (2000)
18. TheEyeTribe: Api reference eyetribe-docs. http://dev.theeyetribe.com/api/ (Accessed: 17 February 2015)
19. Wang, L., Stern, J.A.: Oculometric evaluation of subjects performing a vigilance task: The bakan continuous performance task. Unpublished Manuscript (1997)

Paper C

Assessing Levels of Attention using Low Cost Eye Tracking

This paper was presented by the author at the HCII2016 conference in Toronto Canada, July 2016.

The trip was made possible thanks to Otto Mønsted Fonden.

Correction

Please note that Fig. 2 in the included official printed version for some reason has lost the error trend. The correct figure is shown below.

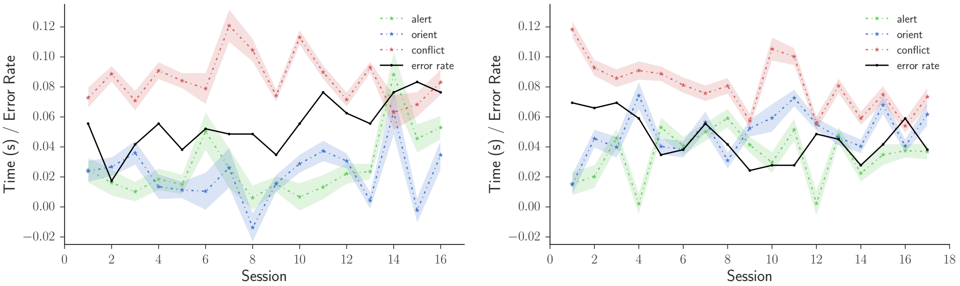


Figure C.1: Attention Network Timing over all sessions in the two week period. Conflict Resolution (Red) is slower than Alertness (Green) and Orientation (Blue). A (Left) shows an increasing error rate trend (Solid); Conflict Resolution for B gradually approaches the other +>latencies. Both A and B have large variations over time, pointing to varying levels of attention, fatigue and motivation.

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].

M. Antona and C. Stephanidis (Eds.): UAHCI 2016, Part I, LNCS 9737, pp. 409–420, 2016.
DOI: 10.1007/978-3-319-40250-5_39

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.

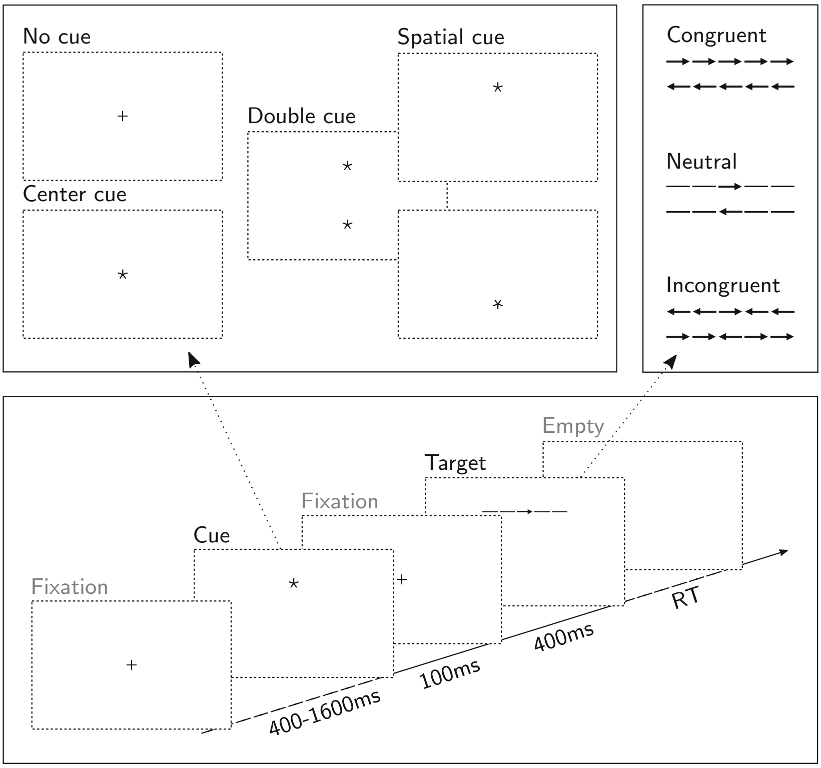


Fig. 1. This Attention Network Test procedure used here: Every 4 s, a cue (either of 4 conditions (Top, Left)) precedes a target (either of 3 congruency conditions (Top, Right)), to which the participant responds by pressing a key according to the central arrow. The reaction time differences between 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 $2 \cdot 3$ 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:

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

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.

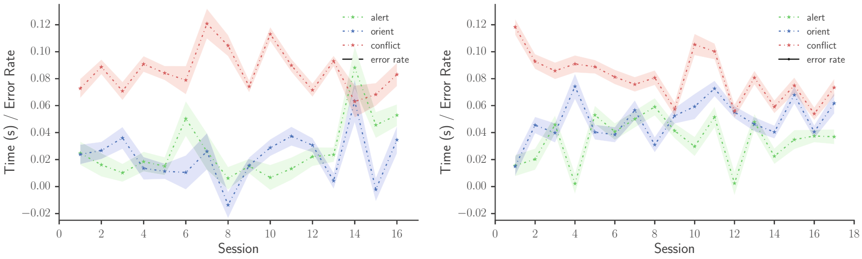


Fig. 2. Attention Network Timing over all sessions in the two week period. Conflict Resolution (Red) is slower than Alertness (Green) and Orientation (Blue). A (Left) shows an increasing error rate trend (Solid); Conflict Resolution for B gradually approaches the other latencies. Both A and B have large variations over time, pointing to varying levels of attention, fatigue and motivation. (Color figure online)

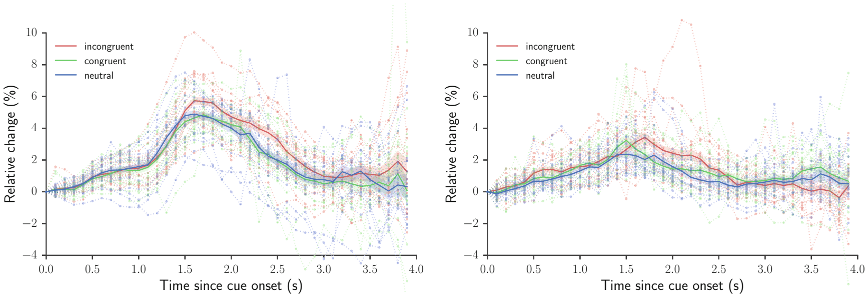


Fig. 3. Averaged left-eye pupil dilations for each session, coloured according to congruency (A (Left) and B). All-session average shown in bold, with the shaded area representing the standard error of the mean. The average incongruent (Red) pupil dilation is stronger than the others, indicating a higher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

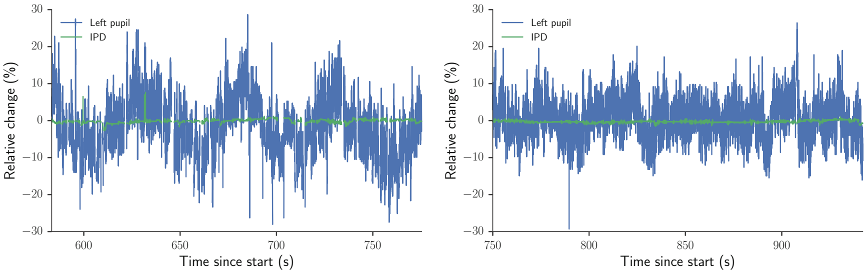


Fig. 4. Filtered pupil size plots; 48-test long sections of two experiments (B, left-eye). Relative inter-pupil distance (Green) indicates stable eye-to-screen distances. (Color figure online)

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 δ(incon) can be calculated by subtracting the congruent value from the incongruent.

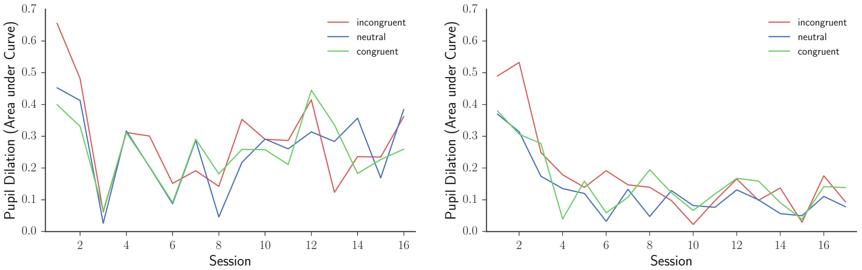


Fig. 5. Area under left-eye pupil dilation curves $[1.5, 2.5]$ s for each session, indicative of cognitive load, grouped after congruency. Both A (Left) and B show initial training effects; only A however shows an increasing trend in cognitive load for the remaining sessions. (Color figure online)

[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.

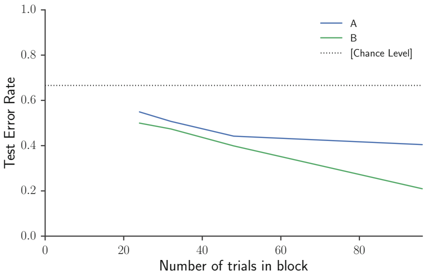


Fig. 6. Test error rates (0.9/0.1 train/test split) predicting averaged 3 s incongruent pupil dilations after cue vs number of averaged experimental tests. At 48 averaged experimental tests, the test error rate at 50 % is clearly below chance (66.6 %, dotted). (Color figure online)

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.

Acknowledgment. This work is supported in part by the Innovation Fund Denmark through the project Eye Tracking for Mobile Devices.

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

Paper D

Separating Components of Attention and Surprise

This paper is in submission to PLOS One.

Separating Components of Attention and Surprise

Per Bækgaard[1*], Michael Kai Petersen[1], Jakob Eg Larsen[1],

1 Cognitive Systems, DTU Compute, Technical University of Denmark, Kgs. Lyngby, Denmark

* pgba@dtu.dk

Abstract

Cognitive processes involved in both allocation of attention during decision making as well as surprise when making mistakes trigger release of the neurotransmitter norepinephrine, which has been shown to be correlated with an increase in pupil dilation, in turn reflecting raised levels of arousal. Extending earlier experiments based on the Attention Network Test (ANT), separating the neural components of alertness and spatial re-orientation from the attention involved in more demanding conflict resolution tasks, we demonstrate that these signatures of attention are so robust that they may be retrieved even when applying low cost eye tracking in an everyday mobile computing context. Furthermore we find that the reaction of surprise elicited when committing mistakes in a decision task, which in the neuroimaging EEG literature have been referred to as a negativity feedback error correction signal, may likewise be retrieved solely based on an increase in pupil dilation.

Introduction

The pupil provides a window into some of the processing that otherwise takes place invisibly inside the human brain. Hess and Polt [1], [2] as well as later Kahneman and Beatty [3], [4] found evidence that linked emotional and cognitive processes to pupil dilations, and Aston-Jones et al. [5], [6] and Joshi [7] have provided a framework for understanding some of the anatomical processes that take place in regulating the gain of the networks involved, and why pupillary reactions are visible:

At the core, the Locus Coeruleus-Norepinephine (LC-NE) system operates in two different modes, *tonic mode* that regulates the overall level of preparedness or arousal and *phasic mode* that is involved in responding to task-relevant stimuli. As task difficulty increases, so will tonic mode activity, modulating the gain, which in turn leads to a increased performance and a stronger phasic response to task-relevant stimuli. If, however, the arousal system and tonic activity mode increase beyond a certain peak point, the phasic responses decrease, leading to an explanation of the classical trade-off between arousal and optimal performance first analysed by Yerkes and Dodson [8].

Activity in LC-NE cells are further reflected in pupillary dilations [7], and the pupil can thus be interpreted as a marker of LC-NE activity (see also Fig. 1 and Fig. 2). Baseline pupil size varies on a large scale of 3-4mm as a response to changes in light levels [9], [10] whereas variations caused by cognitive processes are much smaller, typically on the order of 0.5mm or 15% compared to typical pupil sizes found in normal conditions [4].

The baseline pupil size is modulated by the tonic activity in LC-NE, and is never at rest; it has been known for a long time to vary. Stark et al. [11] speculated that this

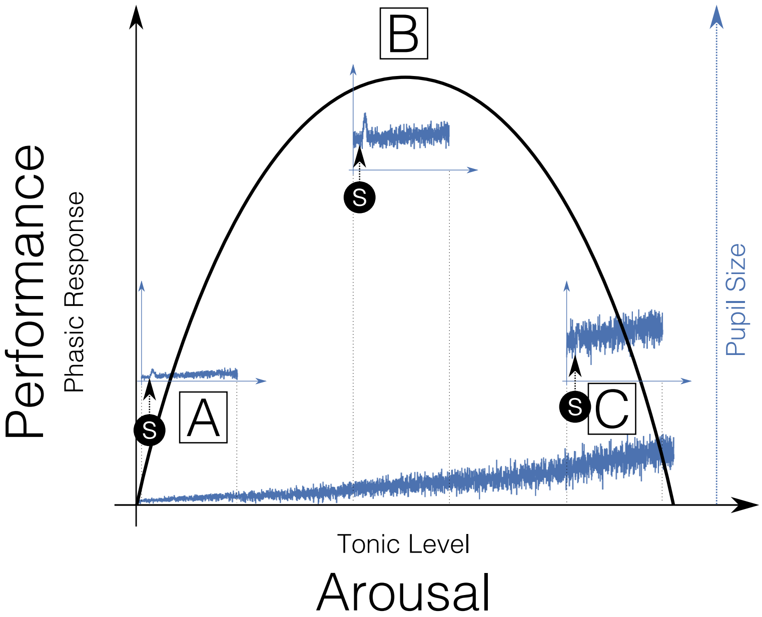


Fig 1. Performance vs Tonic Level, illustrating pupil dilations resulting from a phasic response to relevant stimuli vs the tonic baseline level, regulated by the LC-NE system. Sensitivity to task specific relevant stimuli is greatest at [B], where the largest phasic dilations are seen as compared to at [A] and [C]. Note that the graphs are not actual data to scale but is drawn for illustrative purposes. (Adapted from [5] [7], resembling the classical Yerkes-Dodson relationship [8])

could be part of an "economical" construction of the eye in the sense that there is no need for the eye to operate at a more narrow range, and in our previous study [12] we also noted slow variations of the baseline pupil size of +/-10% on a timescale of 30–60s. Task-Evoked Pupillary Responses [13], [14] (TEPR) above the current baseline are caused by phasic activity in the LC-NE system, and by averaging over many similarly conditioned tests time-locked to the stimuli, other factors can be filtered out.

Recent fMRI studies by Kuchinsky et al. have further established that activity in saliency networks triggered by attentional tasks are reflected in increased tonic pupil size, in contrast to the decreased pupil dilation typically observed when we are in a default resting state [15].

Phasic activations of the LC-NE system in the noradrenergic (NE) neurons also play a role in rapid adaptation to changing conditions, as demonstrated by Bouret et al. [16], in that it may facilitate reorganisation of the innervated areas. This allows for adaption of behaviour to changes in task conditions; real or when they deviate from anticipation.

Preuschoff et al. [17] have found that pupil dilations not only reflect decision making per se or the level of engagement, but also indicates surprise when committing mistakes in decision tasks, suggesting that NE plays a role in error signalling. This appear similar to the negativity feedback components, which is an Event Related Potential (ERP) typically observed in EEG neuroimaging experiments 250-300 ms after participants

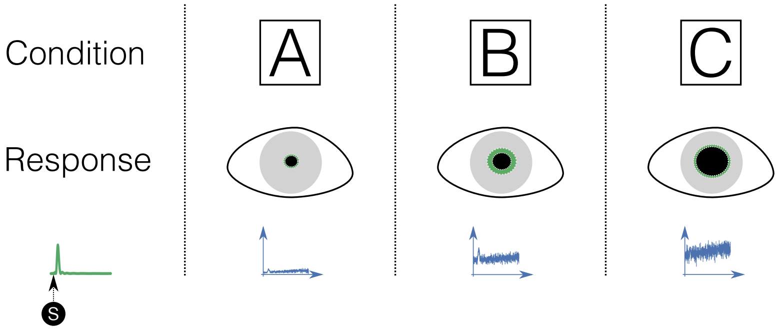


Fig 2. Pupil dilation baseline vs respose to relevant stimuli in 3 different conditions, corresponding to drowsiness [A], highly focused task-specific attention [B] and distractible, scanning attention [C]. The blue curve illustrate the level and fluctuations of the pupil size at each condition. The baseline pupil size is shown in black, with the green area denoting the size of the response present when a task-relevant stimuli appear. Note that the drawing is not to scale.

realize that an incorrect choice was made [18].

Materials and Methods

While attention can be broadly understood as “the appropriate allocation of processing resources to relevant stimuli” [19], Posner and Petersen [20], [21], [22] have shown that three systems, which regulate attention, are anatomically separate from other processing systems and carry out different cognitive roles as part of the attention networks. These are:

- Alerting,
- Orienting and
- Executive Control.

Fan et al. [23] designed a behavioral experiment, known as the Attention Network Test (ANT), to assess which of the network components are activated based on differences in reaction time when responding to visual cues.

We have in a previous experiment [12], measured Task-Evoked Pupillar Responses during the ANT test in a longitudinal study of two subjects. A stronger response was triggered by incongruent conditions in the conflict resolution decision task, likely involving the executive network.

This study expands the number of subjects, investigates the changes in mean pupil size over the experiment, and look at the relationship between the tonic level and the accuracy of the responses.

Experimental Procedure

The procedure followed and the equipment used is identical to that described in [12], and is further illustrated in Fig. 3. In this present study, in total N=18 participants (7

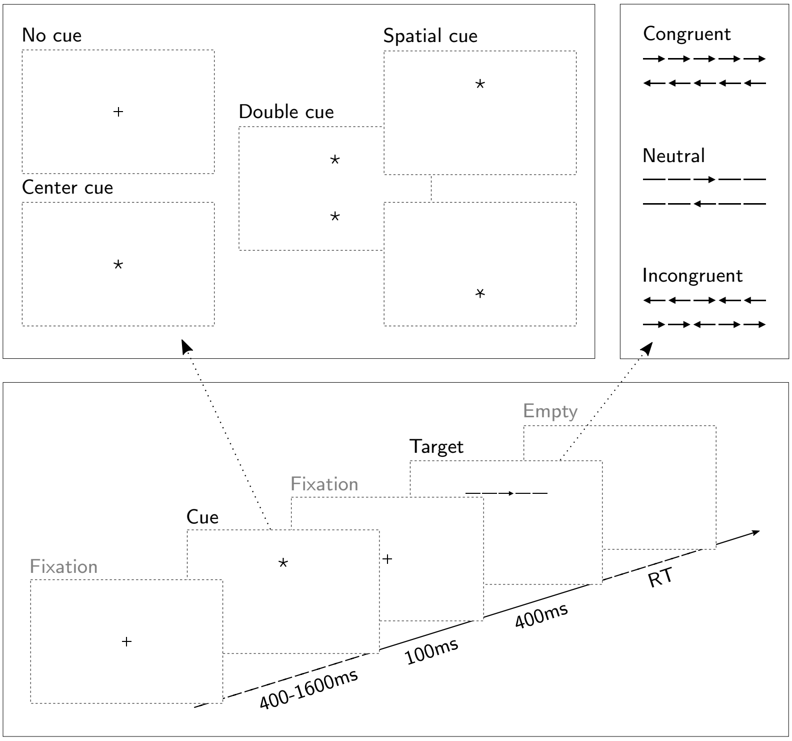


Fig 3. The Attention Network Test procedure used here: Every 4 seconds, a cue (either of 4 conditions (Top, Left)) precedes a target (either of 3 congruency conditions (Top, Right)), to which the participant responds by pressing a key according to the central arrow. The reaction time differences between cue- and congruency conditions form the basis for calculating the latencies of the attention, orientation and conflict resolution networks. *This figure and description is from [12].*

female and 11 male) with a mean age of 25.3 years were tested once. None used glasses or contact lenses, and all but one were right-handed.

The participant were all volunteers that were only allowed to complete the test if they gave consent to their data being used anonymously for research purposes. The ANT test itself is a standard paradigm in widespread use.

Analysis

Pupil size is recorded at 60 Hz, and blink-affected periods are removed. A Hampel [24] filter with a centered window of +/-83ms and a limit of 3σ is applied to remove outliers, and when data is not present in at least half the window, the center point is also removed. This later part takes care of removing any samples immediately before and after blinks, to avoid accidental pupil size changes caused by distortion of the visible part of the eye. Finally data is downsampled to 100ms resolution with a windowed averaging filter.

For the TEPR calcuation, data is epoch'ed to the cue presentation and individually scaled to the value at the start of the epoch.

For the tonic pupil size, a period of 1s immediately before target presentation is sampled to give a representative value without the phasic response, that in most conditions appear to fade away after around 2.5s after stimuli. The pupil was further corrected for variations in head-distance by means of the eye-to-eye distance reported by the eye-tracker.

The mean pupil size was calculated in each of the 4 periods (the initial trial round and the three actual blocks of reaction time tests) as the average value of the filtered pupil data corrected for head-distance variations, which means it is representative of both the tonic pupil size and the ovelaid phasic responses.

Results

Table 1. Average Reaction- and Attention Network-Times over all correct tests across all users (± Sample Standard Deviation listed in parenthesis), in seconds.

MeanRT	Alerting	Orienting	Conflict
0.505 (±0.074)	0.033 (±0.022)	0.019 (±0.016)	0.093 (±0.033)

The mean reaction time and the effect of the alerting, orienting and conflict resolution networks are summarized in table 1. The mean alerting effect was 33ms (±22ms), comparable to the 47ms (±18ms) reported by Fan et al. [23]. The mean orienting and conflict effecs were 19ms (±16ms) and 93ms (±33ms) respectively, to be compared against somewhat higher 51ms (±21ms) and more similar 84 ms (±25ms).

Phasic Pupil Dilation vs Condition and Surprise

Fig. 4 shows the average pupil dilation for all correctly replied tests at the three congruencies as an average value of all correctly replied tests for all subjects, for both left and right eye. The incongruent condition, where the executive control network is also invoked, shows a longer-lasting response in both eyes. Left and right eye responses do not appear statistically significant.

Fig. 5 shows the difference between correct and incorrect (incongruent) responses. A statistically significant stronger response is seen when an error is made, which indicate that the subjects are aware of having made an error. Similar results are seen also for the other congruency conditions, and is also seen for the original longitudinal study when analysed in the same way (not shown).

Tonic Pupil Dilation vs Accuracy and Reaction Time

Fig. 6 and Fig. 7 show scatter plots of all correct (green) and incorrect (red) responses to the incongruent condition according to the reaction time and the tonic pupil size immediately before the test. Incorrect replies are associated with shorter reaction times. The reaction times are statistically different between conditions for all participants of the present study and for A and B in the longitudinal study. The tonic pupil size does not differ in the present study between conditions. However, for the original longitudinal study, participant A shows a statistically significant difference between conditions, with the mean tonic pupil size smaller when incorrect replies are given. See also Table 2.

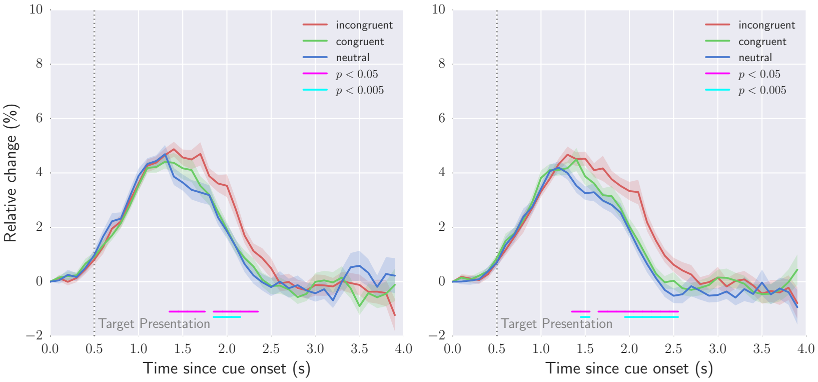


Fig 4. Average Pupil Dilation at the three congruencies for left and right eye, respectively, for the N=18 subjects of the present study (correct responses only). A slight initial reaction appear to the presentation of the cue at $t = 0$, followed by a continued and stronger response to presentation of the target at $t = 0.5$. The congruent (green) and neutral (blue) pupil dilations are similar, but the incongruent (red) response is stronger and lasts longer. The shaded areas represent one standard error of the mean (SEM) to each side. As blinks are more frequently occuring somewhere in the range of t between 1 and 2 seconds, the SEM is somewhat larger here. The magenta line shows where a Welch t-test between the incongruent and the neutral conditions are significantly different with a confidence level of $p < 0.05$; the cyan line marks the $p < 0.005$ level.

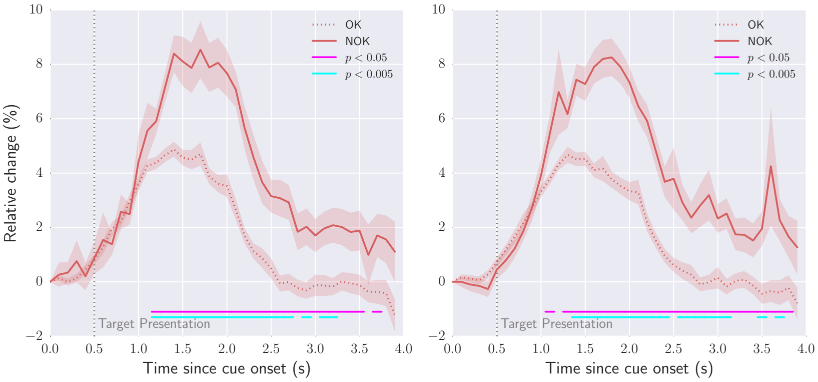


Fig 5. Average incorrect incongruent Pupil Dilations vs correct responses for left and right eye, respectively, for the N=18 subjects, graphed similarly to Fig 4. An incorrect response invokes a statistically significant stronger pupil dilation, almost twice as large as for correct responses.

Tonic Pupil Dilation over time

Fig. 8 shows an illustrative sample of how the mean pupil size (corrected for variations in head distance) varies over the course of the initial training round and the three actual trial blocks. Left and right pupil size are slightly different for this particular subject,

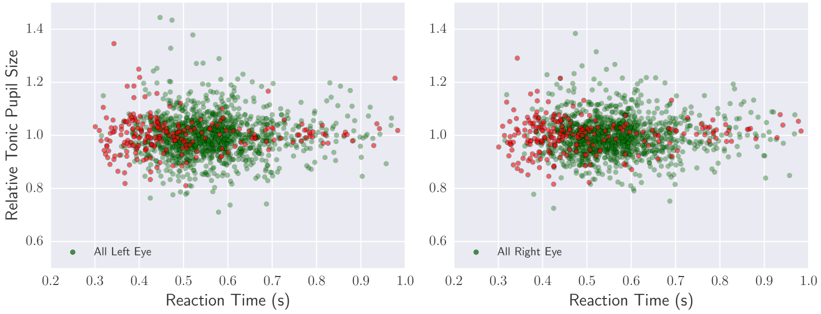


Fig 6. Scatter plot of the tonic pupil size relative to the session mean vs reaction time for incongruent conditions for all participants (left and right eye), colour coded according to correct (green) and incorrect (red) responses. The mean reaction time between correct and incorrect responses are significantly different for the incongruent condition (Welch t-test $t = 7.00$, $p < 0.000001$). The mean relative tonic pupil size between correct and incorrect responses do not significantly.

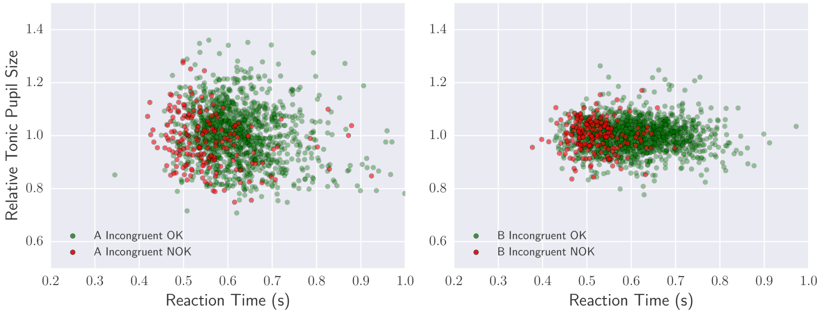


Fig 7. Scatter plot of the left eye tonic pupil size relative to the session mean vs reaction time for incongruent conditions for both participant A (left) and B (right) over all sessions of the longitudinal study, colour coded according to correct (green) and incorrect (red) responses (left eye only shown). The mean reaction time between correct and incorrect responses are significantly different for the incongruent condition for either participant; see Table 2. The mean relative tonic pupil size between correct and incorrect responses only differ significantly for participant A (Welch t-test $t = 2.47$, $p = 0.014$ left eye and $t = 2.39$, $p = 0.017$ right eye); for B they appear very similar. Also note that variations in the relative tonic pupil size appear larger for A than for B.

but there is good correlation between variations of the two (Pearson's $R = 0.948$). An regression corresponding to a low pass filter (a 2nd order polynomial fit) is shown overlaid, and can explain approximately 30-35% of variance (explained variance $R^2 = 0.345$ and $R^2 = 0.301$ respectively).

It also appears as if each block has a slightly larger tonic pupil size initially followed by a decline of approximately 10%. The means for each block also apper to differ, with the initial training round having the larger tonic pupil size.

When comparing the mean tonic pupil size between the initial training round and

Table 2. Relative tonic pupil size and reaction time

			μ	SEM	N	p()
All (N=18)	PSz	OK	0.998	0.002	1472	
		NOK	1.010	0.007	254	
		δ	0.017			0.116
	μRT	OK	0.568	0.003	1472	
		NOK	0.496	0.010	254	
		δ	-0.072			0.000
A	PSz	OK	1.003	0.003	1333	
		NOK	0.982	0.007	198	
		δ	-0.021			0.014
	μRT	OK	0.631	0.003	1333	
		NOK	0.572	0.009	198	
		δ	-0.059			0.000
B	PSz	OK	1.000	0.002	1434	
		NOK	1.001	0.004	197	
		δ	0.001			0.845
	μRT	OK	0.612	0.002	1434	
		NOK	0.519	0.004	197	
		δ	-0.093			0.000

Left eye tonic pupil size, as measured immediately before target presentation, relative to each session's mean, and the reaction times, are listed across all subjects of the present study, and for both participant A and B over all sessions of the longitudinal study, for incongruent conditions, divided into groups of correct and incorrect responses. The mean reaction time (μRT) differ between correct and incorrect responses in a significant way (Welch' t-test $t = 7.00$, $t = 5.99$ and $t = 21.29$ respectively, p<0.000001) for both A and B. The means of the tonic pupil size (PSz) differ significantly between correct and incorrect responses for A (Welch' t-test $t = 2.47$, $p = 0.014$); for B and the participants of the present study, the means between the conditions do not show a statistically significant difference. Almost identical results are found for right eye pupil sizes (not listed here).

the three actual trial blocks, there are statistically significant differences across all participants of the present study, and also for participant B of the longitudinal study. For participant A, however, there are no statistically significant differences. See Fig. 9 and Fig. 10

Fixation Density Map Differences

Average Fixation Density Maps, adjusted for accidental mis-calibrations, were built for each experiment, and were compared between conditions. We did, as expected, see recognisable differences when the target presentation was above vs below the fixation cross, but we were not able to detect any significant spatial differences between congruency conditions nor between cue conditions.

Discussion

The results of this study, with a larger population, supports our previous findings: There is a difference in the incongruent vs congruent/neutral flanker scenarios in that an incongruent condition solicit a larger pupillary response compare to the other two

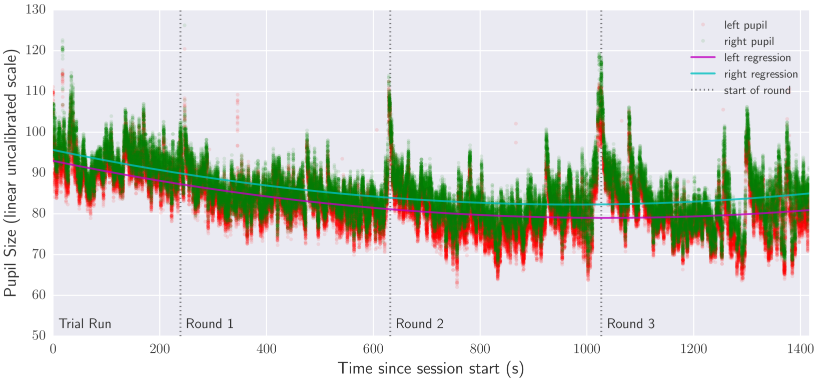


Fig 8. Pupil Size over a sample session that illustrates interesting trends clearly. This session is a first-run for a participant in the larger study. Red and Green marks the measured pupil size (compensated for changes in head distance) for left and right eye respectively. The solid magenta and cyan lines are 2nd order approximations to the pupil size; the explained variance (R^2) are 0.345 and 0.301 respectively. Initiation of each of the 3 rounds of the session are marked with dashed lines. An initial increased pupil dilation diminishes over time as entraining takes place, with a slight increase towards the end. It can also be seen that, in this case, each round starts out with an increase pupil dilation.

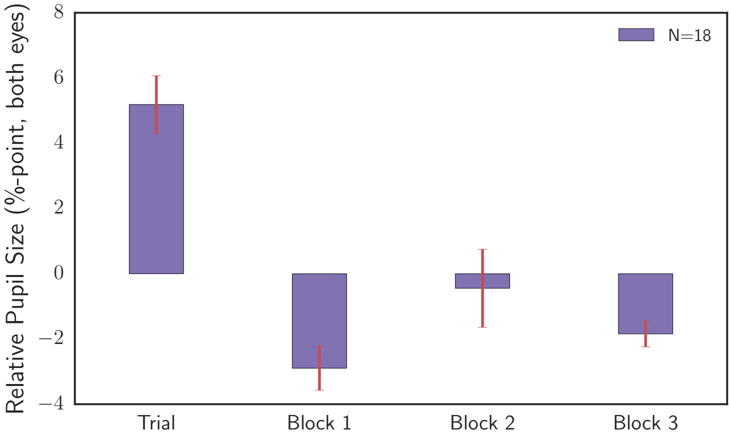


Fig 9. Mean Relative Pupil Size for all subjects in this present study divided into the initial training round and the three actual trial blocks. The red line denotes the SEM. The differences between the initial training round and any of the three other blocks are statistically significant (Wilcoxon signed-rank test $T = 12$, $T = 111$, $T = 35$, all with a confidence level $p < 0.001$). The differences between the other blocks are not statistically significant.

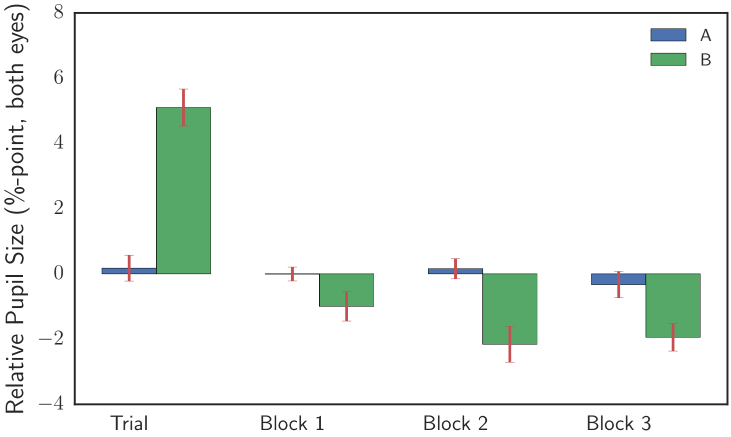


Fig 10. Mean Relative Pupil Size for all subjects in the longitudinal study divided into the initial training round and the three actual trial blocks. The red line denotes the SEM. The differences between the initial training round and any of the three other blocks are statistically significant for B (Wilcoxon signed-rank test $T = 18$, $T = 13$, $T = 18$, all with a confidence level $p < 0.001$). The differences between the other blocks are not statistically significant. However, for A there are no statistically significant differences between the blocks; the variations between the 4 blocks are comparatively much smaller than than what is seen for other participants.

conditions. As the age group is different compared to the previous study, there are indications that the results may be robust and can translate to different settings..

In most cases we see a high correlation (R values from 0.8–0.95) between left and right pupil size, although a few have what may be less than optimal tracking. We cannot conclude any significant difference in the pupil dilation responses between the two eyes, but we notice that the significance level of the difference between the incongruent and the neutral condition is higher and lasts slightly longer for the right eye.

In addition, we also found a significantly different response when subjects replied incorrectly, which happens much more frequently for the incongruent condition. This response may be related to the adaptation and required reorganization reported by Bouret et al. [16] and/or to the surprise elements reported by Preuschoff et al. [17].

Thus, the phasic response reported here as well as in our previous study can be divided into two components that apparently cause a higher level of LC-NE activations: one related to the incongruent condition and one to the incorrect reply.

Comparing the mean relative pupil size over the 4 parts of the experiment (training round and 3 blocks of tests) we found that for the subjects of the present study, as well as for subject B of the longitudinal study, the training round had a statistically significant higher level, around 5%, compared to the other three blocks that averaged around -2%. The subject A of the longitudinal study, however, did not show any such variation between the blocks. We hypothesize that this may point to differences in individual characteristics, behavoiur or preferences.

Further, comparing relative normalized tonic pupil sizes (excluding the phasic responses) showed a statistically significant difference between the level immediately before an incorrect reply compared to the level before a correct reply for subject A of the longitudinal study but not for subject B nor for the participants of the present study. However, while for subject B the levels are almost identical, there is a larger difference even if not statistically significant for the participants of this study, and it therefore cannot be ruled out that participants could fall in different groups that, with more data, would reveal more individual variation.

We also point out that the possible familiarity effects of higher pupillary responses mainly in the two first complete experiments were not tested for in the present experiment, since it was performed only once for each participant. We do, however, see hints at an overall adaptation, as the average (tonic) level decreases as initial entraining to the tasks take place, with a flat or in some cases slightly increased tonic levels towards task completion. This appear similar to the *familiarity effect* reported by Hyönä et al. [25]

We were not able to find any spatial differences in eye movements, at the resolution we worked with, that was related to the conditions of the test, apart from the up-down position of the target.

Acknowledgments

This work is supported in part by the Innovation Fund Denmark through the project Eye Tracking for Mobile Devices.

References

1. Hess EH, Polt JM. Pupil size as related to interest value of visual stimuli. Science. 1960;132:349–50. Available from: `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve{&}db=PubMed{&}dopt=Citation{&}list{_}uids=14401489`.

2. Hess EH, Polt JM. Pupil Size in Relation to Mental Activity during Simple Problem-Solving. Science. 1964;143(3611):1190–1192. Available from: `http://www.jstor.org.proxy.findit.dtu.dk/stable/1712692`.

3. Kahneman D, Beatty J. Pupil Diameter and Load on Memory. Science. 1966;154(3756):1583–1585. Available from: `http://www.jstor.org.proxy.findit.dtu.dk/stable/1720478`.

4. Beatty J, Lucero-Wagoner B. The Pupillary System. Cacioppo JT, Tassinary L, Berntson G, editors. Cambridge University Press; 2000.

5. Aston-Jones G, Rajkowski J, Cohen J. Role of locus coeruleus in attention and behavioral flexibility. Biological Psychiatry. 1999;46(9):1309–1320.

6. Aston-Jones G, Cohen JD. An Integrative Theory of Locus Coeruleus-Norepinephrine Function: Adaptive Gain and Optimal Performance. Annual Review of Neuroscience. 2005;28(1):403–450. Available from: `http://www.annualreviews.org/doi/abs/10.1146/annurev.neuro.28.061604.135709`.

7. Joshi S, Li Y, Kalwani RM, Gold JI. Relationships between Pupil Diameter and Neuronal Activity in the Locus Coeruleus, Colliculi, and Cingulate Cortex. Neuron. 2016;89(1):221–234. Available from: `http://dx.doi.org/10.1016/j.neuron.2015.11.028`.

8. Yerkes RM, Dodson JD. The relation of strength of stimulus to rapidity of habit-formation. Journal of Comparative Neurology and Psychology. 1908;18(5):459–482.

9. Walker HK, Hall WD, Hurst JW. Clinical Methods. Butterworths; 1990. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21250045.

10. Ellis CJ. The pupillary light reflex in normal subjects. The British journal of ophthalmology. 1981;65(11):754–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1039657{&}tool=pmcentrez{&}rendertype=abstract.

11. Stark L, Campbell FW, Atwood J. Pupil Unrest: An Example of Noise in a Biological Servomechanism. Nature. 1958 sep;182(4639):857–858. Available from: http://www.nature.com/doifinder/10.1038/182857a0.

12. Bækgaard P, Petersen MK, Larsen JE. In: Antona M, Stephanidis C, editors. Assessing Levels of Attention Using Low Cost Eye Tracking. Cham: Springer International Publishing; 2016. p. 409–420. Available from: http://dx.doi.org/10.1007/978-3-319-40250-5{_}39.

13. Beatty J. Task-evoked pupillary responses, processing load, and the structure of processing resources; 1982.

14. Ahern S, Beatty J. Pupillary responses during information processing vary with Scholastic Aptitude Test scores. Science. 1979;205(4412):1289–1292. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.472746.

15. Kuchinsky SE, Pandža NB, Haarmann HJ. In: Schmorrow DD, Fidopiastis MC, editors. Linking Indices of Tonic Alertness: Resting-State Pupil Dilation and Cingulo-Opercular Neural Activity. Cham: Springer International Publishing; 2016. p. 218–230. Available from: http://dx.doi.org/10.1007/978-3-319-39955-3{_}21.

16. Bouret S, Sara SJ. Network reset: A simplified overarching theory of locus coeruleus noradrenaline function. Trends in Neurosciences. 2005;28(11):574–582.

17. Preuschoff K, 't Hart BM, Einh??user W. Pupil dilation signals surprise: Evidence for noradrenaline's role in decision making. Frontiers in Neuroscience. 2011;5(SEP):1–12.

18. Sato A, Yasuda A, Ohira H, Miyawaki K, Nishikawa M, Kumano H, et al. Effects of value and reward magnitude on feedback negativity and P300. Neuroreport. 2005;16(4):407–411.

19. Coull JT. Neural correlates of attention and arousal: Insights from electrophysiology, functional neuroimaging and psychopharmacology; 1998.

20. Posner MI, Petersen SE. The attention system of the human brain. Annual review of neuroscience. 1990;13:25–42. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3413263{&}tool=pmcentrez{&}rendertype=abstract$\delimiter"026E30F$nhttp://www.ncbi.nlm.nih.gov/pubmed/2183676.

21. Posner MI. Attentional networks and consciousness. Frontiers in Psychology. 2012;3(MAR):1–4. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3298960/.

22. Petersen SE, Posner M. The Attention System of the Human Brain: 20 Years After. Annual review of neuroscience. 2012;21(35):73–89.

23. Fan J, McCandliss BD, Sommer T, Raz A, Posner MI. Testing the Efficiency and Independence of Attentional Networks. Journal of Cognitive Neuroscience. 2002;14(3):340–347. Available from: http://www.mitpressjournals.org/doi/abs/10.1162/089892902317361886.

24. Hampel FR. The Influence Curve and its Role in Robust Estimation. Journal of the American Statistical Association. 1974;69(346):383–393. Available from: http://www.tandfonline.com/doi/abs/10.1080/01621459.1974.10482962.

25. Hyönä J, Tommola J, Alaja AM. Pupil Dilation as a Measure of Processing Load in Simultaneous Interpretation and Other Language Tasks. The Quarterly Journal of Experimental Psychology Section A. 1995;48(3):598–612. Available from: http://www.tandfonline.com/doi/abs/10.1080/14640749508401407.

Paper E

Pupillary Measurements During an Assembly Task

This paper is in submission to IEEE Transactions on Human-Computer Interactions.

TRANSACTIONS ON HUMAN-MACHINE SYSTEMS 1

Pupillary Measurement During an Assembly Task

Per Bækgaard[*], Shahram Jalaliniya[‡], and John Paulin Hansen[*]

Abstract—**This paper seeks to clarify whether pupillary measures can be used to analyse user experiences in-the-wild. We conducted an empirical study of 57 children using a printed Booklet and a digital Tablet instruction for LEGO® construction while they wore a head-mounted gaze tracker. There were significant differences in the children's pupil dilation under the two task conditions, with Booklets causing a particularly strong dilation when encountered as the first media. Overall, the relative pupil size dropped 3% for those who started with the Booklet and finished with the Tablet, in contrast to relative smaller pupil size changes for the remaining participants. The children who were least productive and asked for assistance more often had a significant different pupil pattern than the rest. Subjective responses confirmed the Tablet to be more fun, faster and easier to use. We suggest how a possible link between pupillary measurement and user experience might be explained in terms of the Yerkes-Dodson [1] relationship between performance and arousal. Finally, we envision how these measurements may be collected automatically by digital devices with gaze tracking built into them.**

I. Introduction

Children are introduced to pictorial instructions for construction toy sets. Instructions remain important throughout adulthood for do-it-yourself assembly of machinery, household appliances, furniture, etc. In the near future, successful joining of homemade 3D-printed components may depend on easy-to-follow instruction material. While 3D-printing offers in-situ production, the pieces fabricated need to be assembled, and without clear instructions on how to do this, the maker will be lost.

Unclear instructions pose a risk that the assembled products will not work or will become unsafe. Safety is of particular concern when people do maintenance or repair on machinery from drawings not fully understood. Emergency procedures, e.g. a defibrillator or a fire extinguisher, require that step-by-step instructions can be read under extremely stressful conditions, emphasizing the need for careful design and evaluation of the instructions.

An assembly task itself is part of the product experience. If drawings are difficult to interpret, assembly may become very frustrating. On the other hand, successful construction guided by well-designed material may provide feelings of achievement and joy [2].

Consequently, when pictorial instructions are a crucial part of a task, it becomes important to consider the support they provide. Is it easy to understand, safe and maybe even fun to use? Ignoring these questions could make people break the product or even injure someone.

Construction tasks include manual assembly of the pieces and handling the instructions, visual perception of the displayed information, and cognitive processing of the information. Video analysis of manual activities and logging of input to digital devices are common. Eye movement recordings are used in a substantial amount of research on naturalistic tasks, e.g. exploring daily activities [3] and web page use [4]. Measurements of mental workload are commonly done by subjective ratings [5], which risk being rationalizations made in hindsight, or using dual-task paradigms [6] that interfere with execution of the primary task. EEG provides another source of data for studying continuous cognitive load [7], but placing electrodes on a participant's head can be a challenge, and subjects wearing them may feel awkward. Likewise, Galvanic Skin Response (GSR) and heart rate measurements require sensors to be attached to the subject, which may hinder mobility.

Pupillometry holds potential as an unobtrusive way to measure the cognitive effort associated with a given task. Data may be collected remotely in real-time by a gaze tracking camera. Gaze tracking is becoming low-cost and is likely to be integrated with future displays. However, in real task situations, one of the main challenges is to distinguish the rather large pupil reactions caused by unknown and uncontrollable changes in the ambient light from the minuscule dilations that reflect changes in cognitive effort. Another challenge is the large individual differences in pupil size and in people's reactions to task difficulties. The pupil size may change over time during execution of a task as people become more – or less – engaged in it. Can noisy pupil data provide information about the user experience and task engagement for a given population? That is the main research question addressed in this paper.

II. Pictorial assembly instructions

Since the pioneering work of Yarbus [8], it has been well known that high-level cognitive goals will govern eye movements in a top-down fashion (e.g. [9]). The task will determine what parts of an assembly instruction get attention at a particular time of the building process. As an example, an initial question could be to establish what tools are needed, or how many people to get involved (cf Fig. 1, left and right respectively) whereas later the Areas of Interest (AOI's)will be determined by the current state of the building process.

The simplest form of assembly instruction is an image of the final state, for instance on the package or a leaflet. More advanced drawings may depict part-whole relationships by slightly separating the pieces in an exploded 3D view. Often, each piece has a number to indicate the order by which it

[*] Technical University of Denmark
[‡] Malmö University, Sweden

should be added, and frequently a line shows exactly where to mount it.

Exploded drawings (cf Fig. 2) can include numerous objects in one perspective view, as seen in repair manuals for cars. Complex assemblages may be hard to read for non-experts because of difficulties identifying the individual parts [10] and because the actual assemblage will change its appearance from the original exploded picture when sub-parts are joined together.

A common way to address complexity is by breaking the building process into steps, with one picture for each step. How best to decompose and depict an assembly process has been extensively studied by e.g. Agrawala et al. [11] and Schumacher [12]. They recommend that each step should cover a significant action in a clear and explicit order, with particular emphasis on new parts added. The angle of view should be kept constant, similar to how it would appear on the floor or workbench, and the orientation should minimise obstruction of significant pieces. Novick and Morse [13] found that the advantage of step-by-step breakdowns became more evident when the number of steps increased (i.e. from 7 to 42 steps).

The preferred medium for step-by-step instructions has been print on paper. However, paper manuals have serious shortcomings. They tend to be very large, heavy and less portable if systems become complex [14]. Locating a particular item in volumes of maintenance documentation may be extremely difficult. Correcting an error after printing a manual is costly, and the drawings will be out-dated when just a single piece of the assemblage changes. Printed drawings offers one fixed point of view and one zoom level only. Experts and novices are served equally with no adaption to the changes in skill level that comes with practise. Finally, printed instructions are impractical because they take up (work-)space and because (at least) one hand is needed when turning a page or unfolding a leaflet. This is particularly inconvenient when people are holding tools or assembly parts in their hands, or wearing gloves.

In recent years, new digital forms of instruction material have emerged. Video instructions are extensively shared on websites like YouTube. They show how to do the assembling task, most often guided by a speaker and sometimes with graphics, text and animation overlays. Research has found interactive video to be an efficient learning tool (e.g. [15]). A common difficulty with video instructions, however, is to locate specific information, for instance finding the sequence where a particular item gets used [16].

Augmented reality (AR) projects an image next to or on top of the items manipulated. Projectors may be mounted at the workspace or embedded in head-worn displays. Several studies have shown potential for AR assembly instructions. Provided the AR systems are able to track the location of items correctly, they may significantly improve work efficiency and reduce cognitive load [17], [18], [19], [20], [21]. To overcome the challenge of keeping the projections synchronized with the assembly items, it has been suggested to use 3D motion sensors and object recognition [22] or even to build sensors into furniture pieces [23].

Apps with interactive building instructions have been provided by LEGO® for some years. When downloaded to a PC, Tablet or smartphone they offer step-by-step instructions on how to build a model. Sequential navigation is done by touching a forward or a backward button located in the corners. Details can be examined by zoom and rotation; animations show how to place a component and a forward/rewind slider helps find a particular event in the construction sequence (cf Fig. 3). Digital instructions often have more steps than the Booklet that comes with the box, for instance model #42031 has 99 steps in the digital version and 84 on paper.

Paper manuals are still dominant because they are universally accessible, robust and easy to produce. Digital manuals require the end-user to possess a display device (e.g. a Tablet or a head-mounted display), an on-line connection, as well as electrical power to charge the device. The device itself is costly and may be vulnerable in rough work environments; it may be difficult to read in sunlight, and handheld devices are almost impossible to operate with gloves or greasy fingers. While it only takes an illustrator to make the drawings needed for a printed manual, a digital manual requires an additional team of software developers.

However, we expect digital instructions to become more common in the future because they offer easy updating, extended explanations (e.g. more steps and animations) and because a large amount of products are routinely modelled in 3D for design, production, and marketing, which makes it less costly to produce digital instructions for end-users. New display forms, like Microsoft HoloLens or virtual reality (VR) headsets, equipped with so-called see-through capabilities, may soon bring augmented reality to industrial and medical settings – and even the general population. If this happens, there are several reasons to believe that gaze tracking will eventually become part of a head-mounted set-up:

- *Substantial amounts of processing-power may be saved by only displaying full image detail at the current fixation point*; so-called gaze contingent resolution [24], [25].
- *Gaze may serve as a hands-free pointer for effortless interaction with the AR-display* [26].
- *Content providers may collect rich information on what users look at, what they ignore, and what causes significant task-related changes in their pupil size.* [27], [28].
- *Adding an eye camera to the set-up is technically uncomplicated; several providers of gaze tracking equipment offer to build binocular cameras into commercial VR gear.*
- *The eye camera will be completely unobtrusive for the user, unlike e.g. EEG or Galvanic Skin Response (GSR) that require additional sensors to be mounted.*

Consequently, it is timely to explore whether pupil measurement could provide information about the user experience when executing an instructed task. This may benefit designers of digital instructions, and it might also provide feedback to students and instructors on learning progress [29].

III. The pupil as an indicator of cognitive effort

The size of the pupil varies from 2mm and 8mm across different subjects and light conditions [30]. Large variations

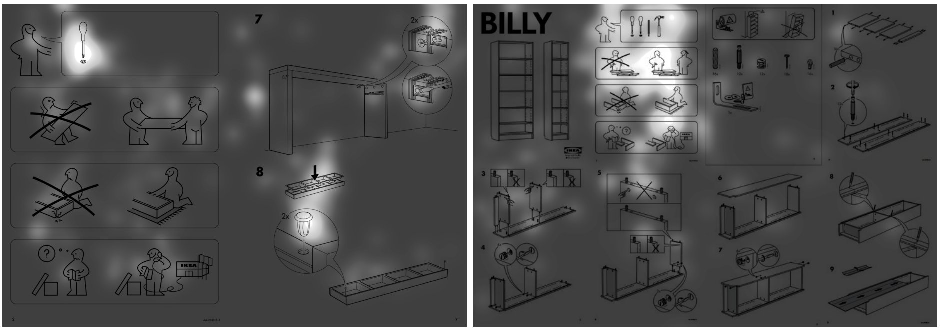


Fig. 1. Burn-through focus maps of 14 people (mean age 30.6 years, 8 males and 6 females) looking at Ikea instructions to determine whether a hammer is needed (right) and to determine whether construction requires one or two people (left). The brightest areas show the areas looked at for the relative longest time.

(3mm) within subjects are caused by changes in light levels controlled by the pupillary light reflex, as reported by e.g. Ellis [31].

However, in 1964 Hess and Polt [32], in a now famous experiment, reported that pupil dilation could also be used as an index of mental activity during multiplication problems. This was based on earlier work [33] relating pupil size to emotional interest in presented material, under control of the sympathetic division of the autonomous nervous system and independent of the light reflex, which is controlled by the parasympathetic division. Kahneman and Beatty [34] subsequently confirmed this finding in a separate study, further suggesting that pupil responses are indicative of memory and processing load. This effectively re-introduced pupillometry as a discipline [35], [36], [37].

Pupillary dilations in response to cognitive processes are usually small [35], up to 0.5mm, which in typical conditions may amount to around 15% of the baseline pupil size. It is therefore common to record several responses from repeated presentations of the same stimuli and conditions, and calculate the average of these in order to filter out noise from other processes. These are called Task-Evoked Pupillary Responses (TEPR), a term probably coined by Beatty [38] and Ahern [39].

Responding to external stimuli relevant to current tasks requires a measured response, to which resources need to be allocated. This process of allocating resources is commonly seen as a key part of the brain's arousal and attentional systems [40]. Aston-Jones et al. investigated the relationship between the Locus Coeruleus-Norepinephrine (LC-NE) system and attention [41], and suggest that LC cells have two overlapping modes of *tonic* and *phasic* activity respectively, modulating attention between focused, or selective, attention (phasic mode, used in exploitation) and scanning attention or high behavioural flexibility (tonic mode, used in exploration). The phasic system is responsible for the TEPR. The balance between the modes is driven by a desire to optimize utility over both short and longer time-scales [42] through an adaptive gain system. The dominant mode impacts the (task-specific) performance, and this system can help explain the classical relationship between arousal and performance originally demonstrated by Yerkes and Dodson in 1908 [1].

Aston-Jones and Cohen further correlate the LC-NE activity to pupil diameter, as also demonstrated later by Murphy et al. [43] who reported a positive correlation between functional neuroimaging BOLD activity of specific parts of the LC-NE system and pupil diameter, and in a very recent work Joshi et al. [44] show that LC-NE-mediated activity may reflect in changes in the pupil diameter, based on (invasive) tests on monkeys. Activation in LC-NE typically leads to pupillary responses on a timescale of seconds, with fast (250ms) response onsets [45].

Other neural correlates classically known to correspond to global arousal levels are seen in the distribution of power in the Electroencephalograpic (EEG) spectrum, whereas components of the Event Related Potentials (ERP) such as P100/N100 and P300 can be modulated by attentional processes [40]. However, EEG recordings are currently not necessarily a suitable match for "in-the-wild" experiments, due to the often invasive nature of the equipment needed and the setup time usually required. For this reason, pupillometry suggests itself as a non-invasive and contact-less alternative, as the arousal and attention systems also appear to correlate with changes in pupil size.

Fig. 4 illustrates a synthesis of these ideas, borrowing especially from Aston-Jones et al. [41], [44]: As the arousal level (x-axis) increases from drowsiness [A], via exploitation, focused attention and peak task-specific performance [B], to a state of highly distractible exploratory behaviour [C], the LC-NE tonic activations increase and the pupil dilates, as illustrated by the blue curve. The task-related performance (y-axis) in response to relevant stimuli increases from [A] to [B], but declines again hereafter as tonic level increases. At the peak [B], relevant stimuli result in the largest phasic pupil response. At [A] and [C], task-specific pupil responses are very small or absent. The blue graph illustrates the resulting visible

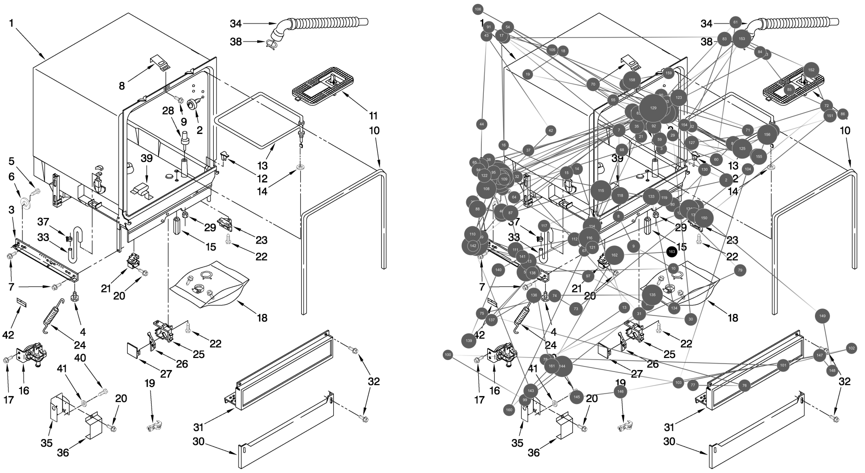


Fig. 2. Exploded 3D-view drawing of an oven (left). An adult subject trying to count the numbers of screws needed to assemble the oven made 163 fixations in 65 seconds to do this (right). Each circle shows a fixation of more than 125 ms, with the diameter proportionate to the duration.

pupil dilation that combines the output from both phasic and tonic processes.

Fig. 5 illustrates the pupil size at each of the conditions [A], [B] and [C] (same as in Fig. 4), without task-relevant stimuli (left) and with task-relevant stimuli (right). The green area (right side only) marks the dilation, i.e. the change from baseline (left, driven by tonic processes) to response maximum when the phasic response is added to the current baseline. At [A], corresponding to a state of drowsiness, and at [C], corresponding to a highly distractible state, phasic pupil dilations are absent or very small. At [B] they are most clearly present.

In unconstrained tasks, the timing of events that trigger phasic responses is typically not known. Marshall [46], [47] has proposed a (patented) technology, the Index of Cognitive Activity, based on filtering abrupt pupil dilations using wavelet transforms to index and identify cognitively induced phasic responses from constrictions and dilations caused by the pupillary light reflex. In the present paper, we instead focus on the simpler concept of averaged means of the pupil size; a concept which reflects not only phasic responses but rather a combination of the current tonic level (arousal) and any phasic activations that take place. It may therefore serve as an index into a combined *level of cognitive effort*[1], which is a result of the neural gain level modulated by activity in the LC-NE system, even though it is also prone to be impacted by changes in light level and therefore may require a balanced experimental design.

Consistent with the Adaptive Gain theory proposed by Aston-Jones and Cohen [42], it can be hypothesized that the brain tries to economize the available resources, and thus picks a tonic level that is motivated by solving a given task and which leads to an adequate performance for the anticipated tasks. Changes in this level may show as changes in the baseline pupil size, e.g. with sustained processing or increasing difficulties leading to an increase in the tonic level [37], [51], [38]. When including the effect of phasic responses in unconstrained tasks in the range from [A] to [B], it can be expected that the mean pupil size will correlate with both the tonic level and the phasic response. From [B] to [C], however, the phasic response correlates inversely with the tonic level and it is less clear what the mean pupil size indicates, as it depends on the amount of task-specific stimuli occurring. It can be assumed, however, that the pupil size continuously reflects activity in LC-NE and as such still

[1] The concepts here are likely to be similar to what Hyönä [48] refers to as *global processing load* or the *pupillometric estimate of mental load*. Similar ideas have also been proposed elsewhere, e.g. Iqbal et al. [49] *percentage change in pupil size* or Palinko et al. [50] *mean pupil diameter change*.

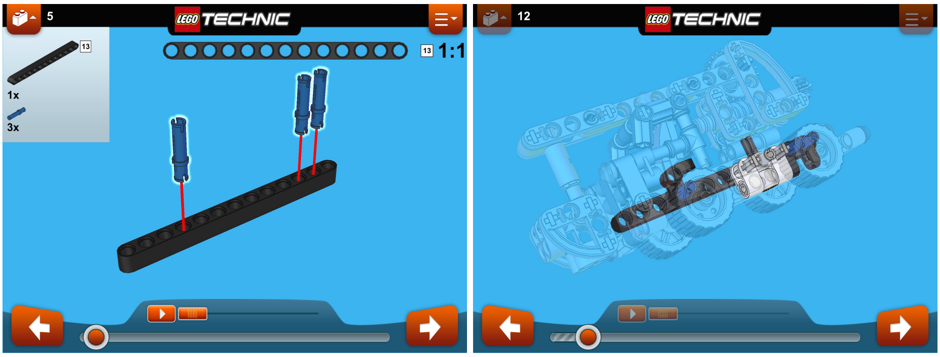


Fig. 3. Left: LEGO® TECHNIC digital building instruction. Arrows in the corners step backward (left corner) and forward (right corner). The play button (middle section) offers a short animation of where to place the two pieces. The slider next to the play button allows the user to control the animation frame-by-frame. The bottom slider with the circle serves as an interactive progress bar. Two-finger pinch on the touch-screen will zoom, and moving the finger on the Tablet will rotate the model. Right: A slightly rotated view of the current build within a ghosted view of the final model.

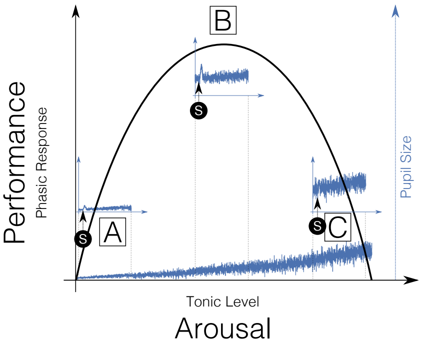


Fig. 4. Illustration of the pupil dilation (right y-axis) resulting from a phasic response to relevant stimuli and the tonic baseline level (x-axis) regulated by the LC-NE system (adapted from [41] [44], resembling the classical Yerkes-Dodson relationship [1]). As the system gains and thereby tonic level increases, and the dominant mode changes from drowsiness [A], via highly focused task-specific attention (exploitation mode) [B] to distractible, scanning attention (exploratory mode) [C], so does the baseline pupil size. Sensitivity to task-specific relevant stimuli is greatest at [B], where the largest phasic dilations on top of the baseline are seen. At [A] and [C] no or only small phasic pupil dilations are seen. The mean measurable pupil size, averaged over time, is a result of combining the phasic and tonic processes. The LC-NE activity generally reflects in the mean pupil size, and even with phasic activity, it can be expected that the mean pupil size increases as attention to task-specific stimuli increases from [A] to [B]. The effort required at a given level of performance manifests itself in the size of the phasic response. Note that the graphs are not actual data to scale but are drawn for illustrative purposes.

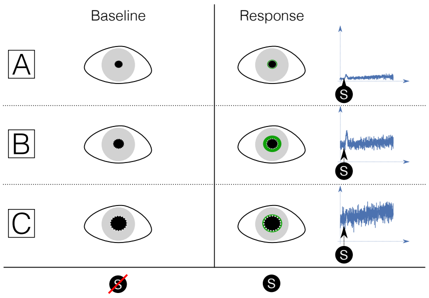


Fig. 5. Pupil dilation compared to baseline as a response to relevant stimuli in three different conditions, corresponding to drowsiness [A], highly focused task-specific attention [B] and distractible, scanning attention [C]. The blue graphs illustrate the pupil size, with fluctuations at each condition. The green area in the pupil to the right illustrates the change from baseline. Note that the drawing is not to scale; pupil variations caused by cognitive processes are typically on the order of 0.5mm, corresponding to a 17% change on top of e.g. a 3mm pupil. Adaptation to light levels causes a significantly larger variation.

represents a level of cognitive effort invested in the current scenario. Thus pupil size may contain important clues on how subjects respond to a given situation. It is additionally worth noting that the settings of the adaptive gain system also have an influence on learning [52].

IV. Previous work

A number of human factor studies of building instructions has been reported (Table I). Common interest was the impact of different media and display forms ([20], [17], [16], [18], [19]) and how to best present the pictorial instructions ([53], [54], [55], [56]). A few studies ([56], [22]) looked at the influence of individual factors like age, gender, education and building experience. The measurements most often taken were task time and errors; while five studies ([20], [55], [56], [16], [18]) included subjective ratings of experienced difficulty, mental effort, fatigue and/or satisfaction and two ([17], [16]) asked for the participants' preferences. Interestingly, three studies ([53], [55], [56] included a manual record of visual behaviour (in terms of "number of looks", "study time", "viewing time" and "gaze time") but without the use of gaze tracking equipment.

The potential benefit of using eye activity to measure cognitive workload in tasks has been addressed in several previous studies. For instance, Van Orden et al. [57] presented a display with target density as the workload variable. Blink frequencies, fixation frequency and pupil diameter showed strong correlation with the density of targets. The changes in pupil size from 1 to 9 targets were highly significant, even though the average change in actual size (calculated as moving estimates of means over a 2 second window) was less than 1 mm. Ahlstrom and Friedman-Berg [58] found no significant effects of task conditions on subjective workload ratings, but significant effects on blink durations, which became shorter when conditions were difficult. The mean pupil diameter increased from 2.4 mm to 3.9 mm for the least responsive subject, while the most responsive had an increase from 2.62 mm to 4.39 mm. The correlation between number of objects (i.e. aircraft) to be supervised (in a simulated air traffic controller operation) and the pupil dilations was rather high (i.e. r^2=0.7). They concluded that measurements of eye activity provide a more sensitive measure of workload over task time than subjective ratings, and suggest this to be particularly relevant when trying to identify display components that cause workload changes which are not reported by operators themselves.

Dehais et al. [59] conducted a study with six pilots flying under low-light (nightfall) conditions with a remote gaze tracker mounted in the cockpit. They observed a large dilation after a simulated engine failure compared to the pilots' pupils during normal flight. Performance data recorded in a driving simulator (i.e. variances of steering wheel angle and lane position) showed a high correspondence to changes in pupil size, even without explicit control of the lighting conditions, in a study presented by Palinko et al. [50]. Finally, in a recent experiment [60] the variations of operators' pupil size were found to be a reliable indicator of the perceived mental workload during a simulated plant emergency; subjects with low, moderate and high task loads showed distinguishably different dilation patterns across the incident.

In summary, building instructions have been addressed in human-factor studies for 20 years; some of them included manual recording of visual behaviour. Human-factor studies in several disciplines have found measurement of pupil size to correspond with performance data and task complexity, but none of these studies dealt with building instructions. Digital building instructions with logging facilities could potentially provide data for further analysis on e.g. task time (in total or on individual steps). Automatic error detection from logs would be more difficult to deal with. Regressions (i.e. "stepping back") might indicate that something went wrong and the step had to be re-built, but regressions due to errors are inseparable from those instances where the user just made forward steps to see what is coming next, and then went back to finish the current part. If gaze tracking capabilities are built into the display of digital instructions, additional visual behaviour may be recorded which could provide an objective, physiological source of data on e.g. fatigue by detection of blink rates [61] and, as is the focus of this paper, an objective measurement of cognitive effort from changes in pupil diameter. Research indicates that detailed analysis of fixation patterns in real time may catch particular types of errors (e.g. [62]). Preferences and subjective impressions of pleasure would still have to be gathered by post-hoc interviews, though.

V. Experimental Setup and Results

A. Participants

57 children (54 boys and 3 girls) between 8 and 10 years-old (average 8.3 years) were recruited in groups of four for the experiment when they were standing in the entrance queue for a LEGO® exhibition. Two were Swedish, the remaining 55 were Danish. Only one participant used glasses; none used contact lenses. All except five were frequent users of smartphones or Tablets, having used one within the last week; 38 had used one the very same day. All participants had built a LEGO® model before, half within the last month. A majority of 51 had never used digital building instructions before; 6 had used them once or a couple of times. The parents signed a consent form, and after the experiment the child was given a LEGO® figure as appreciation for their participation.

B. Equipment

LEGO® instructions served as our study material because they are sufficiently feature-rich and visually compelling to run naturalistic, engaging tests. The digital LEGO® application we used for our study showed model #42031, c.f Fig. 3, which has a total of 99 steps. Details can be examined by zoom and rotation; animations indicate how to place a component, and a forward/rewind slider helps find a particular event in the construction sequence. The Booklet used for comparison was the insert from the box, featuring 84 steps on 31 pages. This LEGO® model has a total of 163 pieces. Since there was a difference in steps between the Booklet and the Tablet we converted all Booklet steps into corresponding Tablet steps for the analysis of productivity (see below).

An Apple Ipad®, model Air 2, with a "LEGO® Building Instructions" app turned on, provided the digital instructions.

A head-mounted tracker (Tobii model Glasses2 RU, 60 Hz, binocular) was used to record eye tracking data, and the Tobii Analyzer software, version 1.36.1430 (x64) was used to extract fixation data with corresponding pupil sizes. Visit times on manuals were scored manually, while the visit times on the Tablet were mapped automatically by the analyzer software. For automatic annotation of the videos, a snapshot image of the Areas of Interest (AOIs) are defined in the analyzer software (See Figure 8). The automatic annotation did not work well for the Booklet so we used manual annotation in the Analyzer software. We reviewed the output of automatic annotation by the Analyzer software and fixed manually a few missing annotations. For both manual and automatic annotation, gaze

[2] HMD = Head Mounted Display

TABLE I
EXPERIMENTAL STUDIES OF ASSEMBLY TASKS WITH PICTORIAL INSTRUCTIONS

Author	Year	N	Independent variables	Dependent variables
Pillay, H. [53]	1997	40	Orthographic drawing; isometric drawing; physical model	Task time; errors; correct completed models; number of looks; study time.
Rodriguez, M.A. [54]	2002	40	Diagrammatic instructions; text and diagrams	Task time; errors
Tang, A. et al. [20]	2003	75	Printed manual; computer,; HMD[2]; AR	Errors; mental effort (NASA TLX rating)
Richardson, M. et al. [55]	2004	64	Number of items, symmetrical planes, fastening points, fastenings, novel assemblies and assembly steps	Subjective rating of assembly complexity (from just reading the instructions); viewing time
Martin, C. V. et al. [56]	2008	24	Toy instructions (LEGO®, K'NEX®, Lincoln Logs®); age; gender; building experience	Task time; errors; usability problems found; gaze time; subjective rating of ease and fun
Henderson, S. et al. [17]	2011	6	Computer; AR; HMD[2]	Task time; head movements; subjective preferences
Gupta, A. et al. [22]	2012	16	Spatial competence; gender; educational level, building experience; feedback guidance (tracking); figure-based guidance	Task time; errors per step
Alexander, K. P. [16]	2013	28	Print; video	Errors, comprehension, subjective rating of satisfaction; preferences
Wille, M. et al. [18]	2014	20	HMD[2] (Google Glass); Tablet PC	Task time (per step); dual task performance; subjective rating of visual and mental fatigue
Funk, M. et al. [19]	2015	15	In-situ projections; print	Task time; errors

fixations were detected automatically by the Analyzer software using the Tobii I-VT (attention) gaze filter (with a minimum fixation length of 60 ms).

C. Procedure

Upon arrival, the children were seated in groups of four around one big table. Each child was randomly assigned to start building the LEGO® model with either a digital instruction on an Apple Ipad (N=28) or the Booklet that comes with the model used for the experiment (N=27). We told the child that the task was to build a model of LEGO® wearing a pair of glasses that would record what he or she was looking at. We emphasized that this was not any kind of competition – the gift would be given no matter how much of the model was completed by the end – and the child was told to ask for assistance if there were any problems.

A pair of Tobii Pro gaze tracking glasses was put on, and a one-point calibration procedure performed. Those starting with a Tablet version were given a short introduction to the main functions of it. Those starting with a Booklet were told to open the box and take out the building instructions. For both media we stressed that the child could go forwards and backwards as they pleased.

After 5 minutes of building, the adult assistant would note how far the child had progressed. When building using the Tablet, the child would at this point be shown how the slider feature for animation worked. Under the Booklet, the assistant would just ask the child to continue. After 10 minutes, the child was asked to pause for a short while for us to change the instructions media; a child starting with a Booklet would now try the Tablet version (given a short introduction) and a child starting with the Tablet would switch to the Booklet. The assistant made sure that the new instructions showed the step currently in progress. If assistance was provided at any time during the building process, the assistant made a note of this.

At the end of the session, the participants were asked which of the two instructions they preferred and why. Also, they were asked whether they thought each of these were easy or difficult, slow or fast, and fun or boring (presented as forced choices).

D. Experimental design

Three independent variables were considered in the experiment:

- *Media*: [Booklet / Tablet]
- *Time*: [First 10 minutes / Last 10 minutes]
- *Novelty*: [New: First five minutes with a given media / Known: Last five minutes with the same media presentation]

This creates in total 8 blocks.

When comparing data between identical sets of participants, within-group comparisons are used (e.g. related samples t-test or Wilcoxon signed-rank test), otherwise between-group comparisons are used (independent samples t-test or Mann-Whitney rank test).

Four dependent variables were examined:

- *Productivity*: Number of building steps produced
- *Assistance*: Number of times an adult assistant was called
- *Gaze time*: The proportionate time (in %) of the build time
- *Pupil data*: The mean pupil size (left and right eyes)

E. Results

We will first present some general results regarding performance and preferences and then a detailed analysis of gaze and pupil data.

1) Performance: Children were not significantly different in their productivity (steps progressed) with the Booklet (mean=18.4; SD=9.4) and the Tablet (mean=18.6 ; SD=10.1); p=0.47, Wilcoxon signed-rank Test. There were large individual differences, indicated by the high standard deviations. The fastest child built 71 steps in total; the slowest only 13 steps. Productivity did not increase from the first 10 minutes (mean=18.6, SD=9.1) to the last, (mean=18.3, SD=10.4); p=0.30, Wilcoxon signed-rank Test.

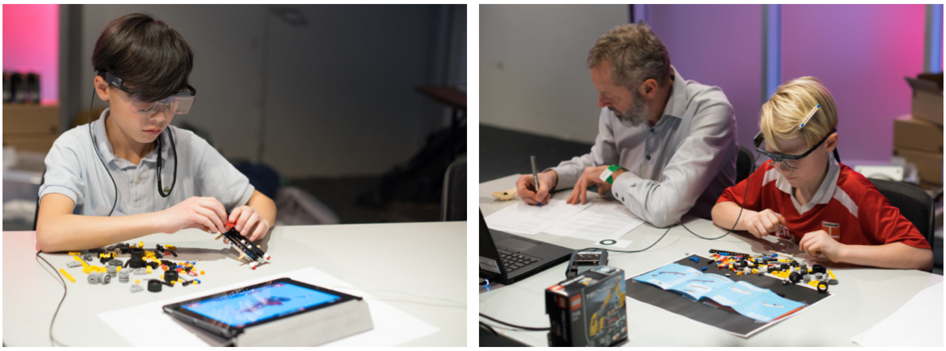


Fig. 6. A child wearing gaze tracking glasses while building LEGO® with Booklet instructions (right) and with Tablet digital instructions (left).

There were no significant differences in the number of times children were assisted with the Tablet (mean=0.7, SD=0.9) and the Booklet (mean=0.9); p=0.25, Wilcoxon signed-rank Test). Assistance was not requested more often during the first 10 minutes (mean=0.9, SD=1.1) compared to the last 10 minutes (mean=0.7, SD=0.9); p=0.35, Wilcoxon signed-rank Test. There was a moderate statistically significant correlation between the number of times a child was assisted by an adult and their productivity, R=-0.65, p<0.001.

2) Preferences: When asked during the post session interview which of the two instruction forms the child preferred, 38 (66%) favoured the Tablet and 19 (33%) the Booklet. Table II shows how the children rated entertainment, ease and efficiency.

TABLE II
TWO-ALTERNATIVE FORCED CHOICE SUBJECTIVE EVALUATION OF TABLET AND BOOKLET, N = 57

Was it...	Tablet	Booklet
fun	55 (97 %)	36 (63 %)
or boring ?	2 (3 %)	21 (37 %)
easy	53 (93 %)	44 (77 %)
or difficult ?	4 (7 %)	13 (23 %)
fast	41 (72 %)	27 (47 %)
or slow	16 (28 %)	31 (53 %)

3) Gaze fixations: We recorded reliable gaze recordings of 48 of the 57 children for both the Booklet and the Tablet sessions, of whom 22 started with the Tablet and 26 started with the Booklet. Nine recordings failed due to extreme looking angles, hair dropping down and covering the eyes, or participants moving the glasses. Only full data sets are included in the analysis.

The children spent on average 52% of the time (SD=12%) studying the Booklet and 64% (SD=12%) studying Tablet instructions. The difference is significant (p<0.0001, Wilcoxon signed-rank Test). The child who studied the Booklet least spent 18% of his time; the child who looked most spent 81%; the corresponding range for the Tablet was 34% to 86%. Comparing visit time per step produced showed a similar pattern, with Booklet being studied for 23.2s in each step (SD=18.3s) and the Tablet studied 27.7s (SD=19.9s); again this difference was significant; p=0.02, Wilcoxon signed-rank Test.

There was a low but statistically significant correlation between productivity and the time spent per building step looking at the Booklet (R=-0.37, p=0.01), and a high and statistically significant correlation for the Tablet (R=-0.71; p<0.0001, Wilcoxon signed-rank Test), meaning that the more the child had to look at the pictures for each step to decode them, the less was built by the end of that session.

Participants spent relatively more time looking at the Booklet instructions during the first 5 minutes than during the last 5 minutes (mean 54%, SD=15% versus mean 49%, SD=14%); p=0.02, Wilcoxon signed-rank Test. This difference, however, was not found for the Tablet condition (mean 64%, SD=15% versus mean 63%, SD=13%); p=0.71, Wilcoxon signed-rank Test. For the last 5 minutes, the difference between Booklet and Tablet was still highly significant, p<0.0001, Wilcoxon signed-rank Test).

We conducted an analysis on the fixation distribution for the Tablet condition, with a predefined set of areas of interest as shown in Fig. 8. The main attractor was the model with 66% of the proportional visit time during the first 5 minutes (SD=16%) and 64% (SD=16%) during the last five minutes. The slider area was the next most attended, with 7% (SD=8%) in the first period and 9% (SD=9%) in the last. The small rise in attention to the slider area between the first and the second part was insignificant; p=0.06, Wilcoxon signed-rank Test. Table III summarizes the distribution.

4) Pupil data analysis: When the eye tracker reported fixations on the relevant instruction media, pupil data was collected for left and right eye and averaged over blocks of 5 minutes, corresponding to New (first 5 min) and Known (last 5 min) conditions for each of the two media, Tablet and Booklet. Subjects were divided into two groups according to the order of presentation, Tablet First or Booklet First. This creates in total of 8 data blocks, with each subject contributing

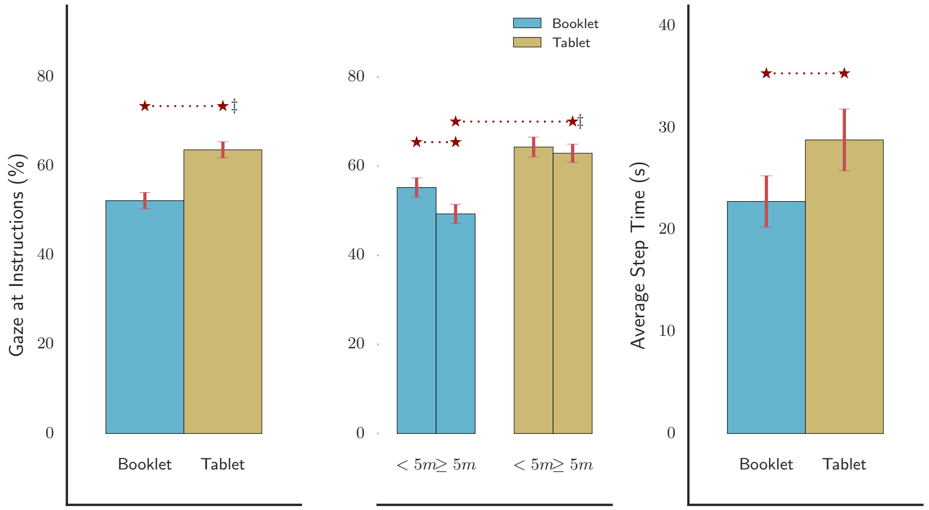


Fig. 7. Differences between fraction of time (in percent) with gaze at instructions between media (Booklet and Tablet; far left) and further split into first and second block with that particular media (middle). There is a statistically significant difference (Wilcoxon signed-rank test, p<0.001) between the media types, as indicated by the ‡, and also between the first and second block for the Booklet (p<0.05), but not for the Tablet. There is also a significant difference (p<0.001) between the second block of Booklet vs Tablet. Also shown (right) is the average step time (seconds) for Booklet and Tablet. The difference is statistically significant (p<0.05).

TABLE III
THE DISTRIBUTION OF AVERAGE PROPORTIONAL VISIT TIME ON AOI'S ON THE TABLET INTERFACE; N=48, NUMBERS IN PARENTHESES SHOW THE STANDARD DEVIATION.

	1:1	Backward	Forward	Menu List	Model	Slider
Tablet New First 5 min	1% (2%)	1% (2%)	2% (1%)	1% (1%)	66% (16%)	7% (8%)
Tablet Known Last 5 min	1% (2%)	1% (1%)	2% (1%)	1% (2%)	64% (16%)	9% (9%)

to 4 of these. The correlation between left and right eye was calculated for each subject, and subjects who had very low correlation, R<0.3, in any of the blocks were completely discarded (the resulting median correlation was R=0.85 across all remaining subjects). In addition, any subjects for whom we did not have reliable gaze data were discarded. This resulted in 43 remaining subjects, 23 having Booklet first and 20 having Tablet first. The pupil data was finally made comparable between groups by normalizing to the relative deviation from each subject's overall median across all 4 blocks.

A three-way ANOVA analysis of the effect on the mean pupil size for each 5-minute block of the three independent variables Time, Media and Novelty (with interactions) was performed. There were statistically significant (p<0.05) effects of Time [F(1,339)=37.3, p<0.001] and Media [F(1,339)=6.8, p<0.009], with the effect of Media:Novelty not being statistically significant [F(2,339)=2.7, p=0.07]. Table IV summarizes the results (Model A).[3]

The slowest group of the 43 participants, which had achieved less than 25 steps and had received assistance at least twice, were picked out and marked as slow performers. Similarly, those that built more than 50 steps and received assistance at most once were marked as fast performers. This resulted in 11 fast performers and 12 slow performers, leaving 20 as normal performers.

A further four-way ANOVA analysis was performed that also included the effects of being slower to build on the mean pupil size. There were statistically significant effects of Time:Performance [F(2,336)=29.3, p<0.001] (the effect of Time differed between the performance groups), Media:Performance [F(2,336)=10.2, p<0.001] (the effect of Media

[3] It was tested whether individual preferences for either Booklet or Tablet stated after the experiment had an impact on the mean pupil size differences, but no effect was found [F(1,338)=0.228, p=0.633].

TABLE IV
ANOVA model A and model B for Experiment 1, with the dependent variable being the mean pupil size during each of the 4 · 5 minute blocks for each participant. Model A includes as independent variables *time* (first or last 10 minutes of the experiment), *media* (Booklet or Tablet) and interaction between media and *novelty* (first or last 5 minutes with the media). Both *time* and *media* have strong statistically significant correlations to the pupil mean size. The total explained variance of Model A, R^2, is 0.133 (adjusted 0.123). Model B includes additionally interaction with the *performance*. The statistical significances are similar, however the explained variance, R^2, increases to 0.211 (adjusted 0.195). The Akaike Information Criterion (AIC) is 1635.0 for Model A and 1609.0 for Model B. Hence, B may be the better model despite the additional parameters needed to estimate the fit.

		SS	df	F	p	η^2	ω^2
A	Time	250.0	1	F(1,339)=37.3	0.000	0.096	0.093
	Media	45.6	1	F(1,339)=6.8	0.009	0.018	0.015
	Media:Novelty	35.9	2	F(2,339)=2.7	0.070	0.014	0.009
	Residual	2270	339				
B	Time:Performance	359.9	2	F(2,336)=29.3	0.000	0.136	0.131
	Media:Performance	187.4	3	F(3,336)=10.2	0.000	0.071	0.064
	Media:Novelty	35.9	2	F(2,336)=2.9	0.055	0.014	0.009
	Residual	2065.8	336				

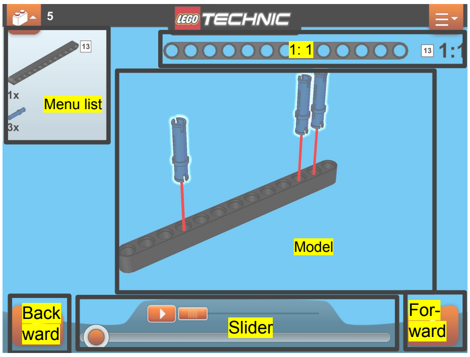


Fig. 8. The Areas of Interest defined for analysis of the visit time on the Tablet instructions.

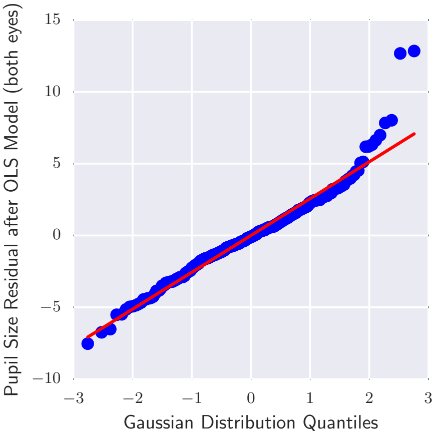


Fig. 9. Quantile-quantile plot of residuals after fitting Model A. The residuals are close to a normal distribution, as can be seen, although with some outliers and a slightly higher kurtosis than with the large amount of data points causes a Shapiro-Wilk test to reject the (null) hypothesis that the distribution is Gaussian. For that reason, we have not only relied on ANOVA and t-tests, but we have also verified (with similar results and significance levels) with Kruskal-Wallis and Wilcoxon signed-rank/Mann-Whitney rank test).

differed between the performance groups). Media:Novelty did not have a statistically significant effect [F(2,336)=2.9, p=0.055]. Table IV summarizes the results (Model B).

Comparing the two Models A (the simpler model) and B (which also includes slow performance) further reveals differences in the effect η^2 caused by the independent variables. The total explained variance, R^2 of Model A is 0.133 (adjusted R^2 0.123) whereas Model B has an explained variance of 0.211 (0.195). The AICs for the two models are 1635 and 1609, indicating Model B may be the better fit even if it has less degrees of freedom.

A quantile-quantile plot of the residuals after fitting Model A is shown in table 9. The residuals resemble a normal distribution, although with some outliers and a slightly higher kurtosis. For this reason, the more conservative non-parametric tests will be used for the Post-Hoc tests[4].

Non-parametric Post-Hoc tests were performed between the blocks where it was meaningful. The results are shown in Fig. 10, where the predictions of Model A are also shown. When the two blocks contain the same set of subjects, within-group (Wilcoxon signed-rank) tests were used, otherwise between-group (Mann-Whitney rank) tests were used.

Looking at the underlying regression model further illustrates the effects of Time, Novelty and Media on the mean pupil size for Model A and B (divided into slow and not-slow performers). See Table V.

For Model A, the mean pupil dilation is 1.71 percentage-points lower in the second half of the experiment (the probability $p(H_0)$ of the effect being 0 is p<0.001) compared to the first half. The last 5 minutes (Known) of using the Booklet has a mean pupil dilation of 0.89 (p=0.025) less than the first 5

[4] The statistical significance levels are however very comparable if using parametric t-tests.

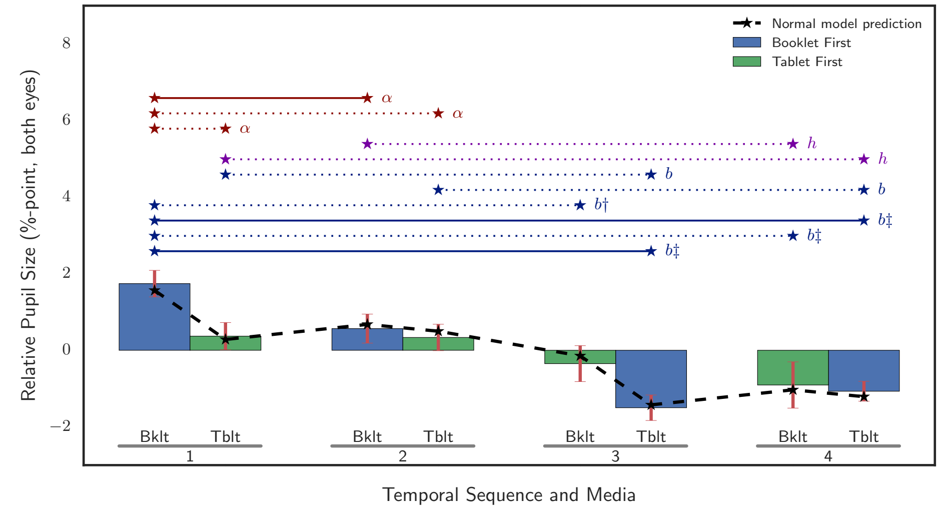


Fig. 10. Post-Hoc between the means of the data from each of the blocks of the ANOVA regression, colour coded according to the two groups of participants that had Booklet first (blue) and Tablet first (green), with corresponding significance levels marked as well as the predicted value (dashed line). The red lines indicate one standard error of the means to each side. The lines above the bars indicate where there are statistically significant differences, with solid lines indicating within-group and dotted lines between-groups. The markers α indicate uncorrected significance, h corrected with Holm-Bonferroni factor, and b the most conservative with Bonferroni correction. The levels are at p<0.05 unless marked with a † or ‡, that indicates 0.01 and 0.001 respectively.

minutes (New); there is a smaller effect of 0.22 the other way round for the Tablet, but this is not statistically significant. Finally, comparing the use of a Booklet to the Tablet shows a mean pupil dilation of 1.28 (p=0.001) less for the Tablet.

For Model B, that includes the effect of being a slow performer, there are differences between how the mean pupil dilation changes over conditions: For the slow performers, there is a small and statistically insignificant difference between the first 10 minutes and the last 10 minutes of the experiment, but for the rest (normal and fast performers) there is a decrease in mean pupil dilation of 2.40 (p<0.001) percentage-points. The Novelty effect is similar to the above (-0.89 (p=0.019) and insignificantly 0.22 respectively), but the differences between the use of a Booklet and Tablet are different: For the slowest performers, the Tablet has a 3.03 (p<0.000) percentage-point smaller mean pupil dilation compared to the Booklet, whereas the rest of the subjects have a smaller (and insignificant) difference of only 0.71 (p=0.085).

Comparing the slowest performers to the rest of the group can also be done by comparing the mean pupil size of the first 10 min to the last 10 min of the experiment, as illustrated in Fig. 11. There are statistically significant differences for the Tablet-first slowest performers compared to the normal group.

Fig. 12 further illustrates the overall trends of the mean pupil size across subjects for the two groups Booklet-first and Tablet-first. The graph has been created by fitting a 2nd order polynomial to each subject's pupil size and averaging across subjects. The Booklet-first group starts around 3% higher than the median value, and declines down to around 1% less than the media. The Tablet-first group shows a more flat curve.

VI. Discussion

We found statistically significant differences in pupil dilation between the eight blocks of the experiment (Table IV) and different trends depending on which media was used first (Fig. 12). Post-hoc analysis (Fig. 10) and modelling (Table V) showed modest but statistically significant effects caused by the independent variables Time, Media and interaction between Media:Novelty. This was further modulated by being a slow performer.

Separating the effects, we found that the pupil constricts close to 2% from the first to the second half of the experiment, although the slowest performers did not show such an effect. We also saw a difference between the two media, Booklet and Tablet, with the Tablet leading to an approximately 1% smaller pupil size, although interestingly we found that the slowest performers had a larger difference of close to 3% between the two media.

With a careful balanced experimental design, this suggest that even in "in-the-wild" conditions like here, where it is not possible to control for variations in light levels, it may be possible to use pupil size as a measurement of cognitive effort: Even if it cannot be ruled out that different luminosity levels impact the pupil size, the identification of novelty effects

TABLE V
CORRECTION TERMS (GENERATED BY THE REGRESSION) FOR MODEL A AND MODEL B. THE NUMBERS ARE PERCENTAGE-POINTS THAT CAN BE SUMMED WHEN ESTIMATING THE MODEL MEAN PUPIL SIZE FOR A GIVEN BLOCK OF THE EXPERIMENT. THE * INDICATES THAT THE TERM DIFFERS FROM 0 AT A SIGNIFICANCE LEVEL OF P<0.05. THE ‡ INDICATES A P<0.001 SIGNIFICANCE LEVEL. NOTE THAT MODEL B, WHEN COMPARED TO MODEL A, SHOULD ALSO INCLUDE AN ADDITIONAL FACTOR OF +0.21 TO BE ADDED TO THE BASELINE FOR THE SLOWEST PERFORMERS (NOT SHOWN IN THE TABLE).

Legend			*Model A: All*			*Model B / All but Slowest*			*Model B / Slowest*		
Time	Novelty	Media	Time	Novelty	Media	Time	Novelty	Media	Time	Novelty	Media
First 10 min	New	Booklet									
		Tablet			-1.28‡			-0.71			-3.03‡
	Known	Booklet		-0.89*			-0.89*			-0.89*	
		Tablet		0.22	-1.28‡		0.22	-0.71		0.22	-3.03‡
Last 10 min	New	Booklet	-1.71‡			-2.40‡			0.33		
		Tablet			-1.28‡			-0.71			-3.03‡
	Known	Booklet		-0.89*			-0.89*			-0.89*	
		Tablet		0.22	-1.28‡		0.22	-0.71		0.22	-3.03‡

(changes between first and second block of using the Booklet but not significantly so for the Tablet) and the differences in effect size of Time and Media between slow performers and the rest of the group suggests that the simpler explanation is that these differences are linked to levels of cognitive effort: 1) Those that experience a slow progress keep investing a higher level of effort and attention compared to the rest of the group, and 2) they also appear to benefit more from using the Tablet instead of the Booklet.

Our main motivation for conducting the analysis of the children's pupillary response during the experiment was to investigate whether this could yield a deeper insight to the experiences they had while building with the Booklet and the Tablet. The majority of the children (i.e. 67%) preferred the Tablet to the Booklet, and it was generally considered more fun, easy and fast to use (cf. Table II). The verbal explanations they gave for their preference for the Tablet were that they liked the rotations and animations, that it was something new, and that it was so easy to get to the next step with no need to turn pages. Those who liked the Booklet explained that it was a format familiar to them and the Tablet was annoying to use because it required them to press buttons all the time. Performance data gave no indications of this difference in preferences, as productivity and the number of times the children were assisted by an adult were not significantly different for the two types of instructions.

Preferences for one of the two forms of instructions did not reveal itself in different pupil patterns, c.f. section V-E4 on page 8. However, the pupil trend curve Fig. 12 suggests that children starting with the Booklet had to show more effort in the beginning of the experiment than those who started with the Tablet. This is further confirmed when comparing the first 5 minutes and last 5 minutes of each block Fig. 10. Across the full session, when participants became more acquainted with the building task, this difference diminished. The group of children who had started with the Tablet did not show any noteworthy increase in pupil dillation when they later encountered the Booklet. Being deeply involved in the building task at this time, they would just continue at the same level of effort as with the Booklet. One group, however, did not follow this general pattern, namely the slow builders starting on a Tablet Fig. 11. They began the experiment with a comparatively smaller relative pupil size (Fig. 11, top) , but then, when the Booklet was encountered, they showed an increase in pupil size that was significantly different from the other groups. Slow builders starting with a Booklet and then trying the Tablet, did not show this increase in pupil size. This finding suggests that monitoring changes in pupil size may be used in the evaluation of a learning proces, c.f. [29] and may help identifying tools for learning that support challenged students best.

The participants looked for a relatively longer time in total at the Tablet instructions (i.e. 64%) than the Booklet (i.e. 52%). This difference may have be explained by the fact that the Tablet version showed slightly more individual building steps (i.e. 99) than the Booklet (i.e. 84), but the difference was also significant when analysing the relative attention time per normalized step. Another explanation might have been that children had to look for the forward button on the Tablet whenever they wanted to go to the next step. Folding a page every now and then in a Booklet, which showed 4 steps per page-turn, would require less attention to doing this. However, the analysis of visit time on the Tablet (c.f. Table III on page 9) showed that only 3 % of time was spent looking at the forward and backward buttons. Manual inspections of gaze recordings revealed that most of the button clicks were performed when the child was looking at the model area, i.e. in the fringe of the vision. Presumably, the main reason for them to look more at the Tablet was the time it took for it to play animations and the time children spent rotating, zooming and sliding the model. The time spent on doing this was not counter-productive, and it may have contributed to the positive impression of the Tablet.

If future displays of digital instructions get eye tracking capabilities, we imagine that this would make it possible for designers at companies like LEGO® or IKEA® to have a test panel of e.g. 100 people, who would try out various versions of instructions and various new models at home. Logging their use of the instructions would allow for measurements of time spent on each step of a task, regressions, visit time on AOIs in the instructions, time-to-first-fixation on AOIs, dwell time, saccade lengths and fixation patterns. In addition, the experiment reported in this paper suggests, that pupil trend curves may reveal patterns in mental effort across the task. When recording pupil variations from a large population, a normal model prediction could be built for every product, identifying e.g. those that were unusually demanding. LEGO®, for instance, could consolidate its recommendation of the age group suitable for that particular model. Furthermore, if future displays become able to record the single areas looked at for every image shown, the pupil analysis may become even more precise, taking into account the brightness of the fixated area. Most importantly, this would allow for an analysis of the within-task variation in task demands that single pieces of information might cause.

VII. Conclusion

We compared digital and paper instructions for LEGO® and found significant differences in children's preferences and gaze behaviour for the two formats, but no difference in their task performance between the two media. Pupil measurements were in accordance with the participants' impression of the digital version being easy to use, indicated by a smaller dilation in pupil size when encountering the Tablet as the first media, compared to the dilation found for those encountering the Booklet first. Children with low performance had a different pupil pattern associated with the the Booklet than the rest, while all children showed a decrease in pupil size across the experiment.

Acknowledgment

The authors would like to thank the Danish Innovation Foundation, who provided part of the funding, and Olav Gjerulfsen and Thomas Visby Snitker, LEGO®, Henning Boje Andersen, Diako Mardanbegi, Lars Yndal Soerensen, Emil Andersen, DTU and Henrik Skovsgaard, The Eye Tribe.

References

[1] R. M. Yerkes and J. D. Dodson, "The relation of strength of stimulus to rapidity of habit-formation," *Journal of Comparative Neurology and Psychology*, vol. 18, no. 5, pp. 459–482, 1908.

[2] T. Tseng and M. Resnick, "Building examples: media and learning affordances," in *Proceedings of the 11th International Conference on Interaction Design and Children*. ACM, 2012, pp. 176–179.

[3] M. Land and B. Tatler, *Looking and acting: vision and eye movements in natural behaviour*. Oxford University Press, 2009.

[4] C. Ehmke and S. Wilson, "Identifying web usability problems from eye-tracking data," in *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it-Volume 1*. British Computer Society, 2007, pp. 119–128.

[5] G. B. Reid and T. E. Nygren, "The subjective workload assessment technique: A scaling procedure for measuring mental workload," *Advances in psychology*, vol. 52, pp. 185–218, 1988.

[6] C. D. Wickens, "Processing resources and attention," *Multiple-task performance*, vol. 1991, pp. 3–34, 1991.

[7] W. Klimesch, "Eeg alpha and theta oscillations reflect cognitive and memory performance: a review and analysis," *Brain research reviews*, vol. 29, no. 2, pp. 169–195, 1999.

[8] A. L. Yarbus, "Eye movement and vision, trans. b. haigh," *ed: Plenum Press, New York*, 1967.

[9] T. T. Brunyé and H. A. Taylor, "When goals constrain: Eye movements and memory for goal-oriented map study," *Applied Cognitive Psychology*, vol. 23, no. 6, pp. 772–787, 2009.

[10] R. E. Mayer and J. K. Gallini, "When is an illustration worth ten thousand words?" *Journal of educational psychology*, vol. 82, no. 4, p. 715, 1990.

[11] M. Agrawala, D. Phan, J. Heiser, J. Haymaker, J. Klingner, P. Hanrahan, and B. Tversky, "Designing effective step-by-step assembly instructions," in *ACM Transactions on Graphics (TOG)*, vol. 22, no. 3. ACM, 2003, pp. 828–837.

[12] P. Schumacher, "Creating effective illustrations for pictorial assembly instructions," *Information Design Journal*, vol. 15, no. 2, pp. 97–109, 2007.

[13] L. R. Novick and D. L. Morse, "Folding a fish, making a mushroom: The role of diagrams in executing assembly procedures," *Memory & Cognition*, vol. 28, no. 7, pp. 1242–1256, 2000.

[14] C. A. Ventura, "Why switch from paper to electronic manuals," in *Proceedings of the ACM conference on Document processing systems*. ACM, 2000, pp. 111–116.

[15] D. Zhang, L. Zhou, R. O. Briggs, and J. F. Nunamaker, "Instructional video in e-learning: Assessing the impact of interactive video on learning effectiveness," *Information & management*, vol. 43, no. 1, pp. 15–27, 2006.

[16] K. P. Alexander, "The usability of print and online video instructions," *Technical Communication Quarterly*, vol. 22, no. 3, pp. 237–259, 2013.

[17] S. Henderson and S. Feiner, "Exploring the benefits of augmented reality documentation for maintenance and repair," *IEEE transactions on visualization and computer graphics*, vol. 17, no. 10, pp. 1355–1368, 2011.

[18] M. Wille, P. M. Scholl, S. Wischniewski, and K. Van Laerhoven, "Comparing google glass with tablet-pc as guidance system for assembling tasks," in *Wearable and Implantable Body Sensor Networks Workshops (BSN Workshops), 2014 11th International Conference on*. IEEE, 2014, pp. 38–41.

[19] M. Funk, S. Mayer, and A. Schmidt, "Using in-situ projection to support cognitively impaired workers at the workplace," in *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*. ACM, 2015, pp. 185–192.

[20] A. Tang, C. Owen, F. Biocca, and W. Mou, "Comparative effectiveness of augmented reality in object assembly," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2003, pp. 73–80.

[21] S. Büttner, O. Sand, and C. Röcker, "Extending the design space in industrial manufacturing through mobile projection," in *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. ACM, 2015, pp. 1130–1133.

[22] A. Gupta, D. Fox, B. Curless, and M. Cohen, "Duplotrack: a real-time system for authoring and guiding duplo block assembly," in *Proceedings of the 25th annual ACM symposium on User interface software and technology*. ACM, 2012, pp. 389–402.

[23] S. Antifakos, F. Michahelles, and B. Schiele, "Proactive instructions for furniture assembly," in *International Conference on Ubiquitous Computing*. Springer, 2002, pp. 351–360.

[24] H. Murphy and A. T. Duchowski, "Gaze-contingent level of detail rendering," *EuroGraphics 2001*, 2001.

[25] E. M. Reingold, L. C. Loschky, G. W. McConkie, and D. M. Stampe, "Gaze-contingent multiresolutional displays: An integrative review," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 45, no. 2, pp. 307–328, 2003.

[26] S. Jalaliniya, D. Mardanbegi, and T. Pederson, "Magic pointing for eyewear computers," in *Proceedings of the 2015 ACM International Symposium on Wearable Computers*. ACM, 2015, pp. 155–158.

[27] A. Bojko, *Eye tracking the user experience*. Rosenfeld Media, 2013.

[28] K. Holmqvist, *Eye Tracking: a comprehensive guide to methods and measures*. Oxford University Press, 2011.

[29] A. Kaklauskas, A. Vlasenko, V. Raudonis, E. K. Zavadskas, R. Gudauskas, M. Seniut, A. Juozapaitis, I. Jackute, L. Kanapeckiene, S. Rimkuviene *et al.*, "Student progress assessment with the help of an intelligent pupil analysis system," *Engineering applications of artificial intelligence*, vol. 26, no. 1, pp. 35–50, 2013.

[30] H. K. Walker, W. D. Hall, and J. W. Hurst, *Clinical Methods*. Butterworths, 1990. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21250045

[31] C. J. Ellis, "The pupillary light reflex in normal subjects." *The British journal of ophthalmology*, vol. 65, no. 11, pp. 754–9, 1981. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1039657{\&}tool=pmcentrez{\&}rendertype=abstract

[32] E. H. Hess and J. M. Polt, "Pupil Size in Relation to Mental Activity during Simple Problem-Solving," *Science*, vol. 143, no. 3611, pp. 1190–1192, 1964. [Online]. Available: http://www.jstor.org.proxy.findit.dtu.dk/stable/1712692

[33] ——, "Pupil size as related to interest value of visual stimuli," *Science*, vol. 132, pp. 349–50, 1960. [Online]. Available: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve{\&}db=PubMed{\&}dopt=Citation{\&}list{_}uids=14401489

[34] D. Kahneman and J. Beatty, "Pupil Diameter and Load on Memory," *Science*, vol. 154, no. 3756, pp. 1583–1585, 1966. [Online]. Available: http://www.jstor.org.proxy.findit.dtu.dk/stable/1720478

[35] J. Beatty and B. Lucero-Wagoner, "The pupillary system," *Handbook of psychophysiology*, vol. 2, pp. 142–162, 2000.

[36] R. F. Stanners, M. Coulter, A. W. Sweet, and P. Murphy, "The pupillary response as an indicator of arousal and cognition," *Motivation and Emotion*, vol. 3, no. 4, pp. 319–340, 1979. [Online]. Available: http://dx.doi.org/10.1007/BF00994048

[37] B. Laeng, S. Sirois, and G. Gredeback, "Pupillometry: A Window to the Preconscious?" *Perspectives on Psychological Science*, vol. 7, no. 1, pp. 18–27, 2012.

[38] J. Beatty, "Task-evoked pupillary responses, processing load, and the structure of processing resources." *Psychological bulletin*, vol. 91, no. 2, pp. 276–292, 1982.

[39] S. Ahern and J. Beatty, "Pupillary responses during information processing vary with Scholastic Aptitude Test scores," *Science*, vol. 205, no. 4412, pp. 1289–1292, 1979. [Online]. Available: http://www.sciencemag.org/cgi/doi/10.1126/science.472746

[40] J. T. Coull, "Neural correlates of attention and arousal: Insights from electrophysiology, functional neuroimaging and psychopharmacology," pp. 343–361, 1998.

[41] G. Aston-Jones, J. Rajkowski, and J. Cohen, "Role of locus coeruleus in attention and behavioral flexibility," *Biological Psychiatry*, vol. 46, no. 9, pp. 1309–1320, 1999.

[42] G. Aston-Jones and J. D. Cohen, "An Integrative Theory of Locus Coeruleus-Norepinephrine Function: Adaptive Gain and Optimal Performance," *Annual Review of Neuroscience*, vol. 28, no. 1, pp. 403–450, 2005. [Online]. Available: http://www.annualreviews.org/doi/abs/10.1146/annurev.neuro.28.061604.135709

[43] P. R. Murphy, R. G. O'Connell, M. O'Sullivan, I. H. Robertson, and J. H. Balsters, "Pupil diameter covaries with BOLD activity in human locus coeruleus," *Human Brain Mapping*, vol. 35, no. 8, pp. 4140–4154, 2014.

[44] S. Joshi, Y. Li, R. M. Kalwani, and J. I. Gold, "Relationships between Pupil Diameter and Neuronal Activity in the Locus Coeruleus, Colliculi, and Cingulate Cortex," *Neuron*, vol. 89, no. 1, pp. 221–234, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.neuron.2015.11.028

[45] V. D. Costa and P. H. Rudebeck, "More than Meets the Eye: The Relationship between Pupil Size and Locus Coeruleus Activity," *Neuron*, vol. 89, no. 1, pp. 8–10, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.neuron.2015.12.031

[46] S. P. Marshall, "The Index of Cognitive Activity: measuring cognitive workload," in *Human factors and power plants, 2002. proceedings of the 2002 IEEE 7th conference on*, 2002, pp. 5–9.

[47] ——, "Identifying cognitive state from eye metrics," *Aviation Space and Environmental Medicine*, vol. 78, no. 5 II, pp. 165–175, 2007. [Online]. Available: http://www.ingentaconnect.com/content/asma/asem/2007/00000078/a00105s1/art00025

[48] J. Hyönä, J. Tommola, and A.-M. Alaja, "Pupil Dilation as a Measure of Processing Load in Simultaneous Interpretation and Other Language Tasks," *The Quarterly Journal of Experimental Psychology Section A*, vol. 48, no. 3, pp. 598–612, 1995. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/14640749508401407

[49] S. T. Iqbal, X. S. Zheng, and B. P. Bailey, "Task-evoked pupillary response to mental workload in human-computer interaction," *Extended abstracts of the 2004 conference on Human factors and computing systems CHI 04*, p. 1477, 2004. [Online]. Available: http://portal.acm.org/citation.cfm?doid=985921.986094

[50] O. Palinko, A. L. Kun, A. Shyrokov, and P. Heeman, "Estimating cognitive load using remote eye tracking in a driving simulator," in *Proceedings of the 2010 symposium on eye-tracking research & applications*. ACM, 2010, pp. 141–144.

[51] G. J. Siegle, N. Ichikawa, and S. Steinhauer, "Blink before and after you think: Blinks occur prior to and following cognitive load indexed by pupillary responses," *Psychophysiology*, vol. 45, pp. 679–687, 2008.

[52] E. Eldar, J. D. Cohen, and Y. Niv, "The effects of neural gain on attention and learning." *Nature neuroscience*, vol. 16, no. 8, pp. 1146–53, 2013. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3725201{\&}tool=pmcentrez{\&}rendertype=abstract

[53] H. Pillay, "Cognitive processes and strategies employed by children to learn spatial representations," *Learning and Instruction*, vol. 8, no. 1, pp. 1–18, 1998.

[54] M. A. Rodriguez, "Development of diagrammatic procedural instructions for performing complex one-time tasks," *International Journal of Human-Computer Interaction*, vol. 14, no. 3-4, pp. 405–422, 2002.

[55] M. Richardson, G. Jones, and M. Torrance, "Identifying the task variables that influence perceived object assembly complexity," *Ergonomics*, vol. 47, no. 9, pp. 945–964, 2004.

[56] C. V. Martin and T. L. Smith-Jackson, "Evaluation of pictorial assembly instructions for young children," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 50, no. 4, pp. 652–662, 2008.

[57] K. F. Van Orden, W. Limbert, S. Makeig, and T.-P. Jung, "Eye activity correlates of workload during a visuospatial memory task," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 43, no. 1, pp. 111–121, 2001.

[58] U. Ahlstrom and F. J. Friedman-Berg, "Using eye movement activity as a correlate of cognitive workload," *International Journal of Industrial Ergonomics*, vol. 36, no. 7, pp. 623–636, 2006.

[59] F. Dehais, M. Causse, and J. Pastor, "Embedded eye tracker in a real aircraft: new perspectives on pilot/aircraft interaction monitoring," in *Proceedings from The 3rd International Conference on Research in Air Transportation. Fairfax, USA: Federal Aviation Administration*, 2008.

[60] P. Bhavsar, B. Srinivasan, and R. Srinivasan, "Pupillometry based real-time monitoring of operatorâĂŹs cognitive workload to prevent human error during abnormal situations," *Industrial & Engineering Chemistry Research*, vol. 55, no. 12, pp. 3372–3382, 2015.

[61] J. a. Stern, D. Boyer, and D. Schroeder, "Blink rate: a possible measure of fatigue." *Human factors*, vol. 36, no. 2, pp. 285–297, 1994.

[62] R. M. Ratwani and J. G. Trafton, "A real-time eye tracking system for predicting and preventing postcompletion errors," *Human–Computer Interaction*, vol. 26, no. 3, pp. 205–245, 2011.

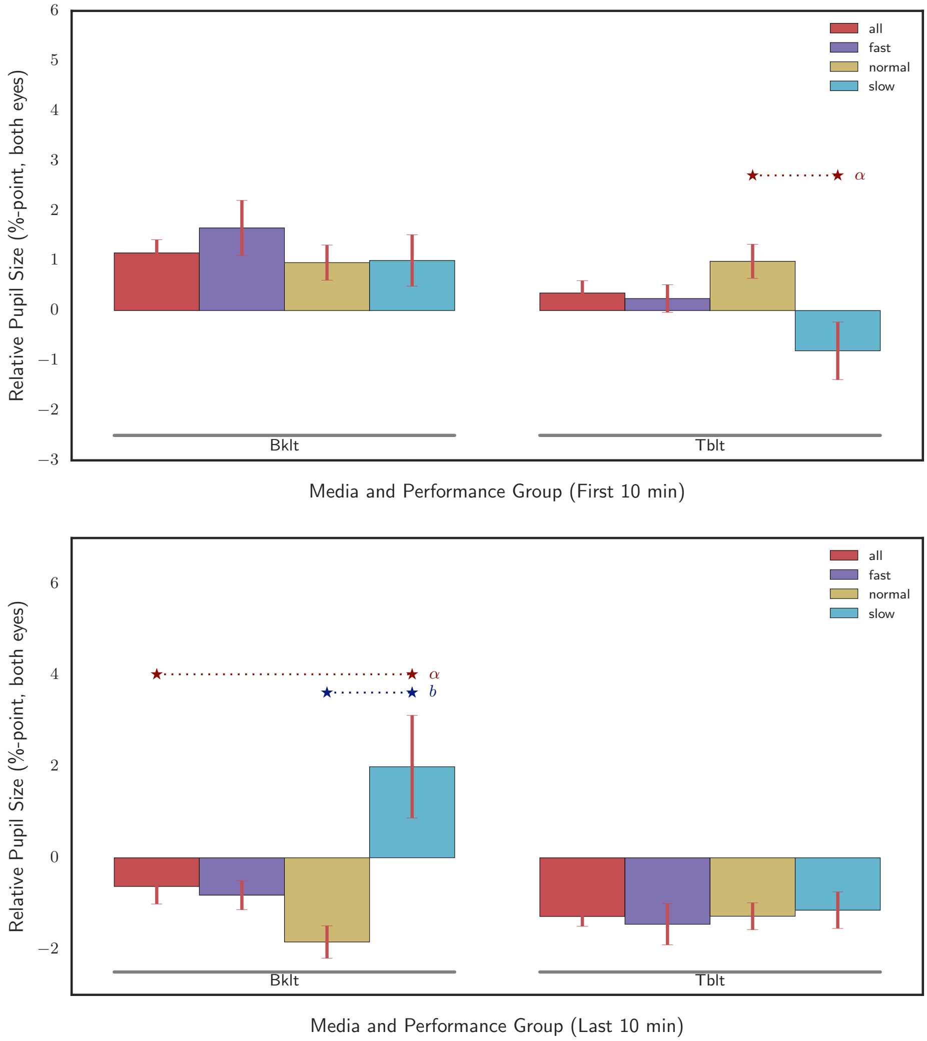


Fig. 11. Differences between the slowest and fastest performing groups vs all or all others, here shown over the two 10-minute periods of the experiment (first 10 minutes at the top, last 10 minutes at the bottom). There are statistically significant differences during the last 10 minutes when continuing on the Booklet (after 10 minutes of Tablet), as marked: The slowest performers have significantly larger pupil size compared to the normal performers and to the overall mean (bottom). There is also a statistically significant difference between the slowest and the normal performers for the group that start with Tablets during the first 10 minutes (top). An α denotes significance at the p<0.05 level; a b denotes significance at the same level after Bonferroni correction. Only comparisons within each media (Booklet or Tablet) and time (first or last 10 min) between the 4 groups of performers have been tested and are shown.

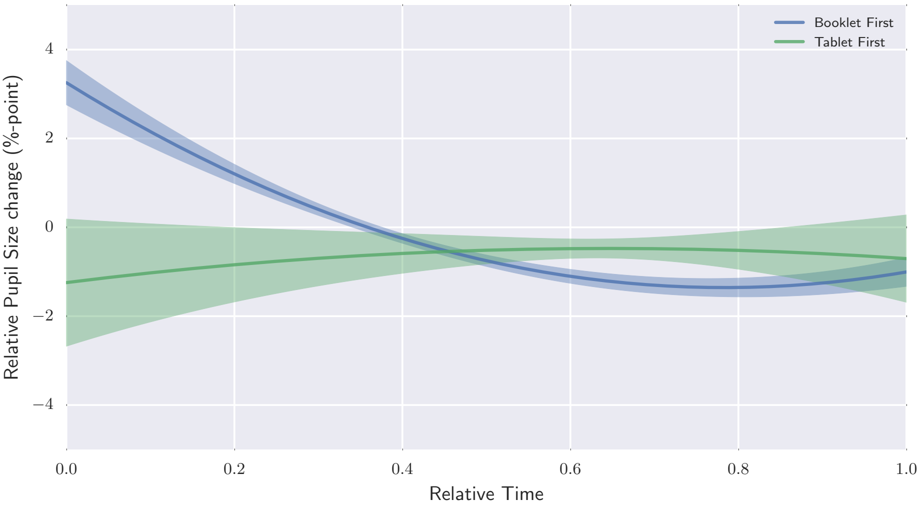


Fig. 12. Filtered trend of how the mean pupil size develops over time for the two experimental groups of participants that had Booklet first (blue) resp. Tablet first (green). The graph is the result of filtering the data for each participant using a 2nd order polynomial (i.e. a heavy low-pass filtering) and then calculating the means over all of these. The shaded area represents one standard error of the mean to each side.

Paper F

SensibleSleep: A Bayesian Model for Learning Sleep Patterns from Smartphone Events

This paper is in submission to PLOS One.

SensibleSleep: A Bayesian Model for Learning Sleep Patterns from Smartphone Events

Andrea Cuttone[1◑*], Per Bækgaard[1◑], Vedran Sekara[1,3], Håkan Jonsson[3], Jakob Eg Larsen[1], Sune Lehmann[1,2]

1 DTU Compute, Technical University of Denmark, Kgs. Lyngby, Denmark
2 The Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark
3 Sony Mobile, Nya Vattentornet, Mobilvägen, 221 88 Lund, Sweden.

◑These authors contributed equally to this work.
* ancu@dtu.dk

Abstract

We propose a Bayesian model for extracting sleep patterns from smartphone events. Our method is able to identify individuals' daily sleep periods and their evolution over time, and provides an estimation of the probability of sleep and wake transitions. The model is fitted to more than 400 participants from two different datasets, and we verify the results against ground truth from dedicated armband sleep trackers. We show that the model is able to produce reliable sleep estimates with an accuracy of 0.89, both at the individual and at the collective level. Moreover the Bayesian model is able to quantify uncertainty and encode prior knowledge about sleep patterns. Compared with existing smartphone-based systems, our method requires only screen on/off events, and is therefore much less intrusive in terms of privacy and more battery-efficient.

Introduction

Sleep is an important part of life, and quality of sleep has a significant impact on individual well-being and performance. This calls for methods to analyze sleep patterns in large populations, preferably without laborious or invasive consequences, as people typically disapprove of the use of intrusive technologies [1].

Large scale studies of human sleep patterns are typically carried out using questionnaires, a method that is known to be unreliable. It is possible to perform more accurate studies, but these are currently carried out within small controlled environments, such as sleep labs. In order to perform accurate measurements of sleep in large populations—consisting of thousands of individuals—without dramatically increasing costs, alternative methods are needed.

Smartphones have become excellent proxies for studies of human behavior [2,3], as they are able to automatically log data from built-in sensors (GPS, Bluetooth, WiFi) and on usage patterns (phone calls, SMS and screen interaction), from which underlying user behavioral patterns can be derived.

Smartphone data has been used to infer facets of human behavior such as social interactions [4], communication [5], mobility [6], depression [7] and also sleep patterns [8]. Either paired with additional sensors or on their own, mobile app solutions are able – sometimes very ingeniously – to track individual sleep patterns and visualize them. We cite as examples *Smart Alarm Clock* [9], *Sleep Cycle* [10], *SleepBot* [11], and *Sleep as Android* [12].

Using mobile phone data to derive sleep patterns has thus already been demonstrated and verified, and offers advantages (i.e. reduced cost) as an alternative to dedicated sleep monitoring devices. In this paper we suggest extending previous approaches, using a Bayesian model to infer rest and wake periods based on smartphone screen activity information. The advantages of our proposed Bayesian approach *SensibleSleep*, as compared to previous work, are that it:

- is less sensitive to "noisy" data, for instance infrequent phone usage during sleep interruptions (such as checking the phone at night)
- is able to quantify not only specific rest and wake times but also characterize their distributions and thus uncertainty
- can encode specific prior beliefs, for instance on expected rest periods (when desirable)
- can capture complex dependencies between model variables, and possibly even detect and relate patterns that are common to a group of people with diverging individual patterns (when using one of the proposed hierarchical models), such as detecting how available daylight may modulate sleep patterns across an otherwise heterogeneous group of users

Our method, moreover, only needs screen on/off events and is thus *non-intrusive*, *privacy-preserving*, and has *lower battery cost* than microphone or accelerometer based ones.

We start by providing an overview of the related work. We then describe the collected data, and introduce the Bayesian model. We compare the model results with ground truth obtained by sleep trackers, and show how the model is able to infer the sleep patterns with high accuracy. Finally we describe the individual and collective sleep patterns inferred from the data.

Related Work

A key finding by Zhang et al. [13] shows a global prevalence of sleep deprivation in a group of students, partly linked to heavy media usage. In this study sleep patterns are largely deduced from the teachers' perception or based on individual self-reports, lacking more direct measurements.

Corroborating this finding, Orzech et al. [14] report that digital media usage before bedtime is common among university students, and negatively impacts sleep. The findings are based on studies involving self-reports through (online) sleep diaries and digital media surveys, and also lacks more direct measurements of sleep patterns. Additionally, this would make it possible to increase the scale of the experiment and enable the study of larger populations.

Abdullah et al. [8] have previously demonstrated using 9 subjects how a simple rule-based algorithm is able to infer sleep onset, duration and midpoint based on a (filtered) list of screen on-off patterns with the help of previously learned individual corrective terms, and further analyzed behavioral traits of the inferred *circadian rhythm* [15, 16]. The algorithm uses an initial two weeks of data with journal self-reported sleep for learning key corrective terms in order to improve the accuracy and compensate for differences between *actual* sleep and *inferred* nightly rest period. The method has been verified against a daily online sleep journal and results in differences less than 45 minutes of average sleep duration over the entire analysed period. While our proposed Bayesian model, which has been applied to more than 400 users, may be more complex, it increases the robustness and allows us to better quantify the uncertainties of the inferred resting periods as well as offer the possibility of building more advanced models across heterogeneous groups of users. In particular, our model may better be able to handle short midnight interruptions, which appear to be not uncommon, without any additional filtering.

In contrast to Abdullah et al. using (only) screen on-off events, a fine-grained sleep monitoring by “hearing” and analyzing breathing through the earphone of a smartphone is suggested by Ren et al. [17]. Here six users tested the system over a period of 6 months, demonstrating the feasibility of using smartphones for the purpose of analysing breathing patterns, using a Respiration Monitor Logger as ground truth. Sleep estimates are not directly inferred in this paper, however. This technology is also non-invasive, although it does requires capturing and analyzing large samples of audio data.

iSleep [18] proposes detecting sleep patterns by means of a decision tree model, also based on audio features. The system was evaluated with 7 users for a total of 51 days, and shows high accuracy in detecting snoring and coughing as well as sleep periods, but report drops in performance due to ambient noise.

Increasing the number of features, the *Best Effort Sleep model* [19] is based on a linear combination of phone usage, accelerometer, audio, light, and time features using a self-reporting sleep journal, and subsequently achieved a 42 minutes mean error on 8 subjects in a test period of 7 days.

Other work also tries to estimate sleep quality, for example *Intelligent Sleep Stage Mining Service with Smartphones* [20], which uses Conditional Random Fields on a similar set of features trained on 45 subjects over 2 nights, and reports over 65% accuracy of detection of sleep phases, compared to EEG ground truth on 15 test subjects over 2 nights.

Candy Crushing Your Sleep [21] uses the longest period of phone usage inactivity as heuristic for sleep, with some ad-hoc rules for merging multiple periods, and proceeds to quantify the sleep quality and to identify aspects of daily life that may affect sleep. The inferred sleep period was however not validated against any ground truth.

The *Sleep Well* framework [22] deploys a Bayesian probabilistic change-point detection, in parallel with an unsupervised classification, of features extracted from accelerometer data, in order to identify fine-grained sleep state transitions. It then uses an active learning process to allow users to incrementally label sleep states, improving accuracy over time. It was evaluated both on existing datasets with clinical ground truth, and on 17 users for 8-10 days with user diary data as ground truth, reaching an average sleep stage classification accuracy approaching 79%.

In comparison, even though sleep quality is not estimated, our non-intrusive model only needs screen on/off events and has been tested on a large user-base, and can suitable for very large-scale deployment.

Methods

Data Collection

We have analyzed two datasets in this work.

The first dataset (A) was provided by Sony Mobile, and contains smartphone app launches coupled with sleep tracking data from the SWR10 and SWR30 fitness tracking armbands [23]. For each user we have a set of records containing an anonymized unique user identifier, a timestamp and the unique app package name. Note that the model only uses the app launch timestamp and completely ignores the app identifier, therefore no privacy risks related to app names are present. The sleep tracking data indicates when each user is detected asleep or awake with a granularity of one minute, serving as ground truth that we will compare our results against. From this dataset we select 126 users that have at least 3 hours of tracked sleep per day, and have between 2 and 4 weeks of contiguously tracked sleep.

The second dataset (B) originates from the *SensibleDTU* project [24], which collected smartphone sensor data for more than 800 students at the Technical University of Denmark. In this dataset we focus on the screen interaction sensor that records whenever the smartphone screen is turned on or off, either by user interaction or by notifications. Each record contains a unique user identifier, a timestamp, and the event type (on or off). From this dataset we select 324 users in November 2013 that have at least 10 events per day, thus filtering out users with gaps in the collected data or with very sparse data. There is on average ≈ 76 screen-on activations pr. day pr. user in this period.

Data collection for the SensibleDTU dataset was approved by the Danish Data Protection Agency, and informed consent has been obtained for all study all participants. Data collection for the Sony dataset has been approved by the Sony Mobile Logging Board and informed consent has been obtained for all study participants according to the Sony Mobile Application Terms of Service and the Sony Mobile Privacy Policy.

Model Assumptions

The underlying assumptions of the model are (1) that the user is in one of two modes: being *awake* or *sleeping*, and (2) that mobile phone usage differs between the two modes. In particular a user will have many screen interactions when awake, and very few or even no interactions when sleeping.

Sleeping is here considered as an extended *resting* period that typically takes place once every 24 hours at roughly similar times, as governed by the users circadian rhythm and influenced by socio-dynamic structures, during which the owner physically rests and/or sleeps. Resting periods, however, might be interrupted by short periods of activity, such as checking the time on the phone or responding to urgent messages. This behavior leads to two different activity levels, which we label λ_{awake} and λ_{sleep}, one for each mode.

If we can deduce when the switchpoint between the two distributions occur during each 24 hour period, we can also infer the time during which the owner is *resting* for the night, and thereby also the period within which sleeping takes place.

Short of using the more invasive EEG or polysomnographic methods, properly differentiating the resting period and actual sleep is difficult; even sleep diaries may easily contain reporting bias or be somewhat inaccurate. To remove self-reporting bias and to study a larger population we have therefore decided on using a motion-based detector (Sony fitness tracking armbands) as ground truth.

If higher accuracy would be required, applying individual corrective terms (i.e. average sleep/rest time differences) learned from an initial period by more accurate means (polysomnography, external observer or possibly a careful user diary) might be possible, similar to what as demonstrated by Abdullah et al. [8].

Model Structure

Each user is considered independently. We divide time into 24−hour periods starting at 16:00 and ending at 15:59 on the next calendar day, so that the night period and the expected sleep midpoint is in the middle, for convenience. Each day is divided into $n = 24 * 4 = 96$ time bins of size 15 minutes. We count the number of events that start within each time bin, where an event is an app launch for dataset A and a screen-on for dataset B. Information about the duration of the events is purposely discarded, as phone usage typically takes place in short bursts. This is supported by the median duration of screen events in dataset B, which is ≈ 26.5 seconds. It is reasonable to assume that the count of events k in each time bin follows a Poisson distribution:

$$P(k) = \text{Poisson}(k, \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

with $\lambda = \lambda_{awake}$ or $\lambda = \lambda_{sleep}$, depending on the mode of the user. It is, furthermore, assumed that the user mode, and consequently the value for λ, is determined by two switchpoint variables t_{sleep} and t_{awake}, both assuming values from 0 to n:

$$\lambda = \begin{cases} \lambda_{sleep} & \text{if } t_{sleep} \leq t < t_{awake} \\ \lambda_{awake} & \text{if } t < t_{sleep} \vee t \geq t_{awake} \end{cases}$$

For simplicity, all models assume that λ_{sleep} is identical for all days of a given user. It can be expected that users have a very low number of screen events during sleep mode, which is encoded in this prior belief:

$$\lambda_{sleep} \sim \text{Exponential}(10^4)$$

Here Exponential represents the exponential distribution:

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

The rate parameter is set to a very large value to encode our prior belief that almost no events should happen during the sleep time.

Fig. 1 shows an illustration of the model idea.

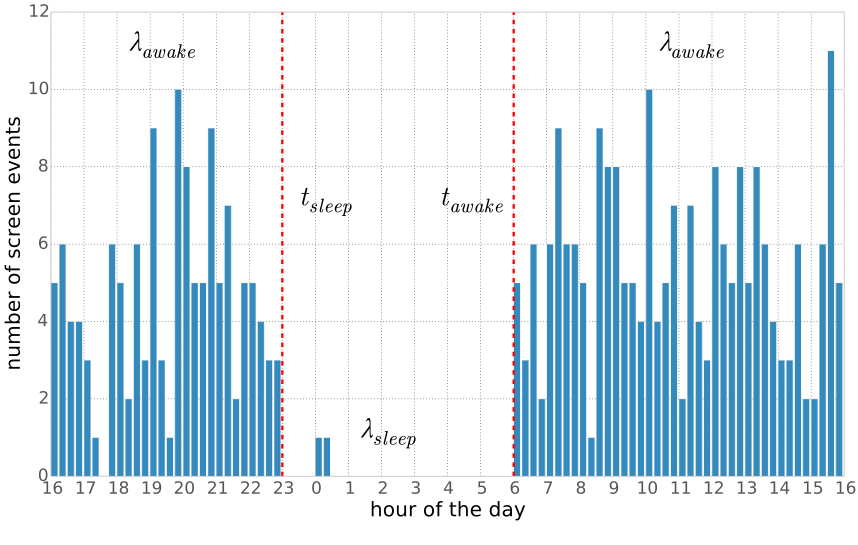


Fig 1: Conceptual illustration of the model. We assume that for each day the event counts follow two different Poisson distributions: one for sleep periods (rate λ_{sleep}) and one for awake periods (rate λ_{awake}). Furthermore we assume that two switchpoints t_{sleep} and t_{awake} determine the rate (i.e. the Poisson distribution) that generates the events.

We now propose four different models, which differ in the assumptions made on the relation of the rate and sleep/awake time parameters for different days.

Pooled-Pooled Model: Pooled Times and Rates

The simplest model assumes that for a given user there is a single λ_{awake}; i.e. the user has very similar phone interaction patterns each day. Also t_{sleep} and t_{awake} are each identical for all days, that is: the user goes to sleep, and wakes up, at the same times each day:

$$\begin{aligned} t_{sleep} &\sim \text{DiscreteUniform}(0, n) \\ t_{wake} &\sim \text{DiscreteUniform}(0, n) \\ \lambda_{awake} &\sim \text{Gamma}(2.5, 1) \end{aligned}$$

Here $\text{DiscreteUniform}(0, n)$ represents a uniform probability to choose a timebin between 0 and $n = 96$. No additional prior knowledge of t_{sleep} and t_{awake} is assumed; there is equal probability of any bin value. In other words, sleep and awake time are equally probable at any time of the day. The prior for λ_{awake} is chosen to represent our prior belief of a reasonable rate of events, specifically with both mean and variance $= 2.5$ (events/bin) and a longer tail than a normal distribution.

Independent-Pooled Model: Independent Times

A somewhat more realistic model would assume that each day has independent t_{sleep} and t_{awake} times, while still sharing λ_{awake} rates. Therefore in this model there are t^i_{sleep} and t^i_{awake}, with $i = 1...m$, one for the each of the considered days:

$$\begin{aligned} t^i_{sleep} &\sim \text{DiscreteUniform}(0, n) \text{ for } i = 1...m \\ t^i_{wake} &\sim \text{DiscreteUniform}(0, n) \text{ for } i = 1...m \\ \lambda_{awake} &\sim \text{Gamma}(2.5, 1) \end{aligned}$$

The rest of the model remains as above.

Independent-Independent Model: Independent Times and Rates

It may further be assumed that each day could have its own specific activity rate. We modeled this as separate λ^i_{awake} for each of the m days, in addition to t_{sleep} and t_{awake} for each of the m days:

$$
\begin{aligned}
t^i_{sleep} &\sim \text{DiscreteUniform}(0, n) \text{ for } i = 1...m \\
t^i_{wake} &\sim \text{DiscreteUniform}(0, n) \text{ for } i = 1...m \\
\lambda^i_{awake} &\sim \text{Gamma}(2.5, 1) \text{ for } i = 1...m
\end{aligned}
$$

Independent-Hyper Model: Hierarchical Rates

The assumption that each day's interaction rate is completely independent may not be correct. It may not be unreasonable to imagine that the daily rate(s) arise from an underlying user-specific rate; i.e. the user may have certain habits that varies from day to day but share some similarities specific to that user. This is modeled by adding α_λ and β_λ hyperparameters to the Gamma priors for λ^i_{awake}:

$$
\begin{aligned}
t^i_{sleep} &\sim \text{DiscreteUniform}(0, n) \text{ for } i = 1...m \\
t^i_{wake} &\sim \text{DiscreteUniform}(0, n) \text{ for } i = 1...m \\
\alpha_\lambda &\sim \text{Exponential}(1) \\
\beta_\lambda &\sim \text{Exponential}(1) \\
\lambda^i_{awake} &\sim \text{Gamma}(\alpha_\lambda, \beta_\lambda) \text{ for } i = 1...m
\end{aligned}
$$

We do not have strong prior beliefs for α and β, so we set their prior distributions to generic exponential distribution with rate parameter = 1, Exponential(1).

Hyper-Hyper Model: Hierarchical Times and Rates

Finally we could assume that each day's sleep and awake times derive from an underlying circadian rhythm that is specific to the user, but still modulated by events that take place during the week. This can be modeled by changing the t^i_{sleep} and t^i_{awake} priors to a normal distribution, with hyperparameters α_t, β_t and τ_t as follows:

$$
\begin{aligned}
\alpha_t &\sim \text{Exponential}(1) \\
\beta_t &\sim \text{Exponential}(1) \\
\tau_t &\sim \text{Gamma}(\alpha_t, \beta_t) \\
t^i_{sleep} &\sim \text{Normal}(8 * (n/24), \tau_t) \text{ for } i = 1...m \\
t^i_{wake} &\sim \text{Normal}(15 * (n/24), \tau_t) \text{ for } i = 1...m \\
\alpha_\lambda &\sim \text{Exponential}(1) \\
\beta_\lambda &\sim \text{Exponential}(1) \\
\lambda^i_{awake} &\sim \text{Gamma}(\alpha_\lambda, \beta_\lambda) \text{ for } i = 1...m
\end{aligned}
$$

The t^i_{sleep} are here chosen to be centered at the bin corresponding to 23:00, while the t^i_{awake} are centered at the bin corresponding to 07:00. Also in this case we have no strong prior knowledge of the τ_t, α_t and β_t parameters, so we set their prior distribution to a non-informative Exponential and Gamma respectively.

Model Fitting and Selection

The models are fitted using Markov Chain Monte Carlo (MCMC) sampling [25], where the parameter values are estimated by a random walk in the parameter space guided by the log likelihood. We use the *pymc3* python library [26,27] for running the sampling, but any MCMC framework could be used to implement our model. The result of the Bayesian inference is a trace that captures the most probable values of the parameters, and also gives an indication of the uncertainty of the estimation.

It is important to note that the models are unsupervised, which means that they are fitted only to the number of events without having access to the ground truth of the actual sleep patterns. This allows the model to be fit to other datasets where we do not have ground truth of sleep patterns, which is desirable if the sleep inference has to be deployed on a large scale. For dataset A we verify the fit by comparing with the sleep patterns from sleep trackers, while for dataset B we evaluate the fit by inspecting the inferred sleep patterns.

In order to find the model that provides the best overall fit for the intended purpose without introducing too many degrees of freedom, we compare the log posterior from the traces of the models, logp, and see how they converge.

One example of a plot of logp traces for the five models is shown in Fig. 2, which shows that the hyper-hyper model (blue) has the highest (least negative) logp, followed by the independent-hyper model for dataset B. The three other models appear with lower logp. In 76% of the analyzed cases of dataset A (84% for dataset B), the hyper-hyper model has the highest logp score, followed by the independent-hyper model with the highest logp in 11% (13%) of the cases.

The logp estimation does not, however, take into account the added complexity of the more advanced models. An attempt to do so is the Deviance Information Criterion (DIC) [28], which penalizes the increased degrees of freedom (more model parameters) that usually result in a model that is easier to fit to the data. Fig. 3 shows the Relative DIC score (vs. the simplest model, pooled-pooled). The order is identical for both datasets.

Further, Table 1 compares the 5 models by ranking the calculated DIC for all 126 and 324 users. The *median* rank shows that the hyper-hyper model is the "best" model; it has a probability of being the best ranked model (*p(Best)*) in 62% of the cases for dataset A (69% for dataset B). The independent-hyper model follows as a somewhat distant 2nd best, ranking highest in 17% (19%) of the cases.

It should be noted that, in addition to their different abilities to reflect the underlying assumptions and provide varying levels of fit to the actual data, the models also differ in their runtime; the most complex model typically takes 15 times longer to execute than the simplest. In particular, the hyper-hyper model

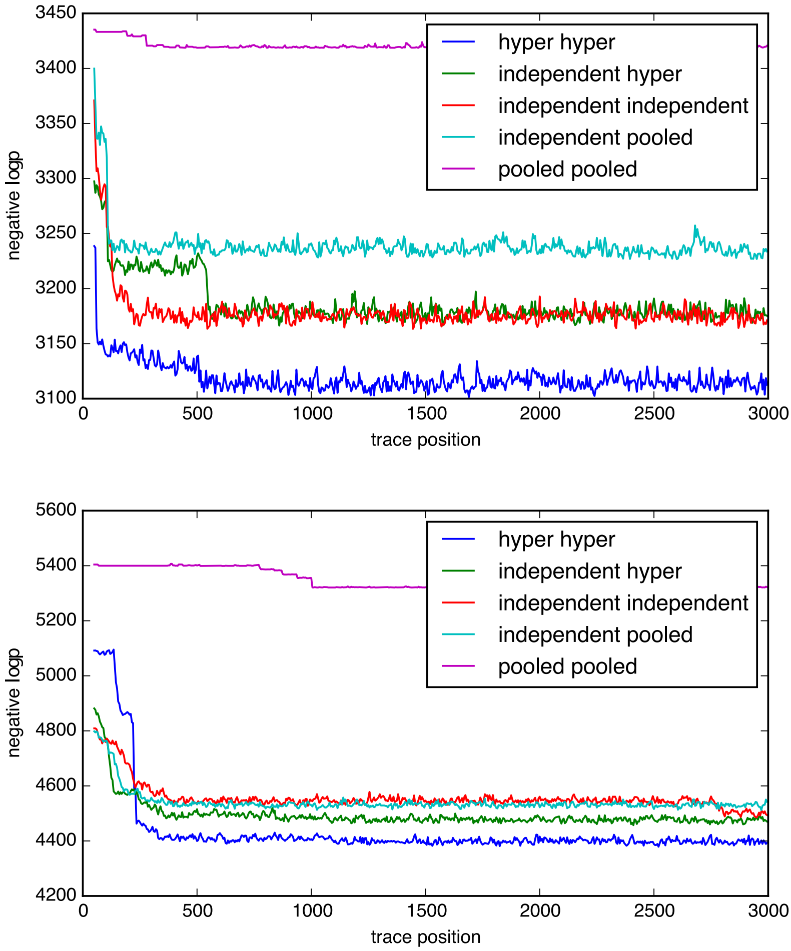


Fig 2: Typical logp traces (A top, B bottom)

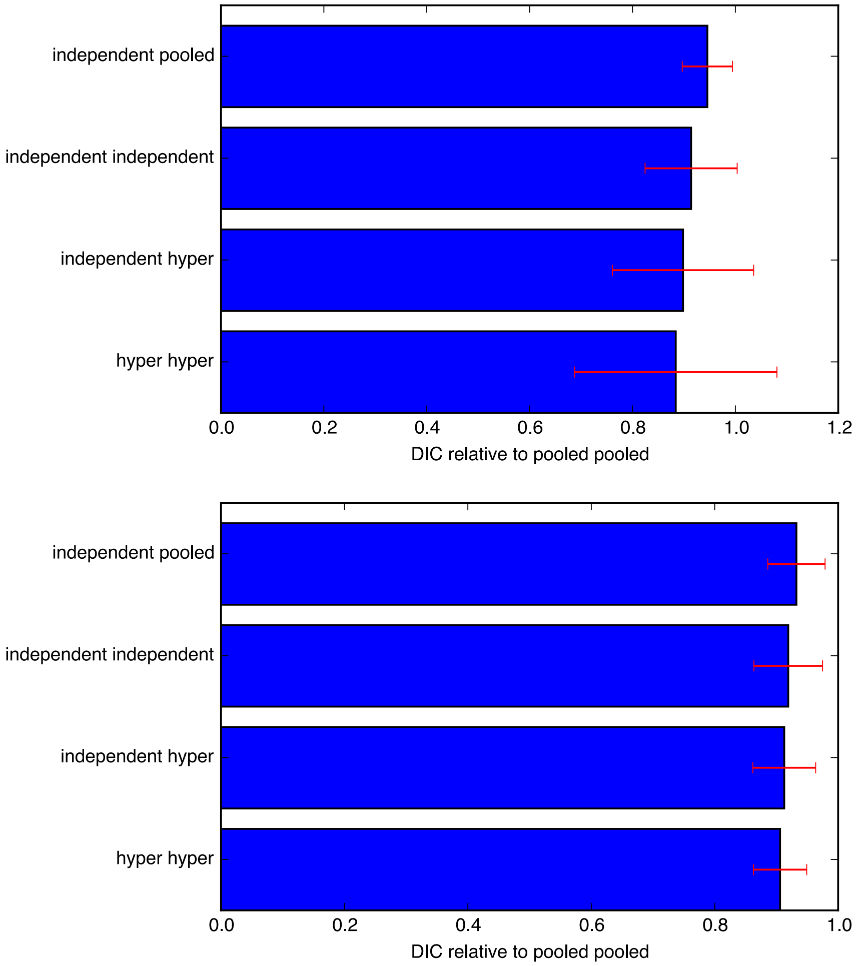


Fig 3: Relative DIC scores (A top, B bottom), sorted by their mean value (error bars represent one standard deviation). For both datasets the order is the same, with the hyper-hyper model having the lowest mean DIC.

on average had a runtime that is 60% longer than the independent-hyper model, so there may be cases where the latter would be a better model to use despite the slightly worse DIC ranking.

	Model Ranks	Median	Mean		p(Best)	Mean Relative DIC	
			Value	(StdDev)		Value	(StdDev)
A	pooled-pooled	5	4.27	(1.37)	0.10	0.96	(0.16)
	independent-pooled	4	3.82	(0.85)	0.03	0.95	(0.05)
	independent-independent	3	2.86	(1.08)	0.08	0.91	(0.09)
	independent-hyper	2	2.29	(0.83)	0.17	0.90	(0.14)
	hyper-hyper	1	1.76	(1.11)	0.62	0.88	(0.20)
B	pooled-pooled	5	4.70	(0.89)	0.02	0.99	(0.01)
	independent-pooled	4	3.75	(0.66)	0.02	0.93	(0.05)
	independent-independent	3	2.92	(1.02)	0.09	0.92	(0.06)
	independent-hyper	2	2.06	(0.69)	0.19	0.91	(0.05)
	hyper-hyper	1	1.56	(0.94)	0.69	0.91	(0.04)

Table 1: Model DIC comparisons

Results

All five models have been run on both datasets, producing an estimation of the times of sleep and wake up for each day, as well as estimates for the other hyperparameters, for each user. Moreover, we calculated logp and DIC as discussed in the previous section. We firstly verify the accuracy our method using the ground truth from the sleep trackers. We then provide a qualitative analysis of some key examples of individual sleep patterns, and a description of the aggregated sleep patterns for both datasets. For the remainder of the paper we restrict our analysis to the model with the best fit, the hyper-hyper model.

Comparison to Related Work and to Ground Truth

To assess the results, we compare the sleep periods inferred by our model and those inferred by a previously suggested rule-based method to the ground truth collected by the Sony sleep trackers.

For each day we calculate the time of sleep and time of awake inferred by our model as the mean of the t^i_{sleep} and t^i_{wake} respectively, and we consider the user asleep (Z = 1) for all time bins between t^i_{sleep} and t^i_{wake}, and awake (Z = 0) for the remaining bins.

For a representative and comparable method, we chose to implement a rule-based algorithm similar to what is proposed by Abdullah et. al. [8] to derive sleep data for dataset A. This rule-based method essentially works by finding the longest contiguous sleep period, with a prior assumption that sleep must start after 10 PM and before 7 AM next morning. Note that the original algorithm is based on screen on-off events and furthermore discards events of short duration during the night; in our case we use app launches with no available duration, and thus cannot discard events of short duration.

For the sleep trackers we can directly mark each time bin as sleep (Z = 1) if the trackers have detected at least one sleep status in that bin, and awake (Z = 0) otherwise.

We again consider one user at a time. For each user we now have three binary matrices: two inferred sleep status values per time bin from either model, and one measured sleep status value per time bin (ground truth) . We evaluate this as two binary classification problems, and calculate accuracy, precision, recall and F1 for each model and for each user according to the definitions:

$$\text{accuracy} = \frac{\text{correct predictions}}{\text{predictions}}$$

$$\text{precision} = \frac{\text{true positives}}{\text{predicted positives}}$$

$$\text{recall} = \frac{\text{true positives}}{\text{all positives}}$$

$$\text{F1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Fig. 4 shows the resulting distribution of accuracy, precision, recall and F1 scores for the proposed method. The SensibleSleep method achieves a mean accuracy of 0.89, and a mean F1 score of 0.83. The below-average scores for some users are expected, since it is likely that among the large population under study there will be people having irregular sleep schedule or noisy sleep ground truth.

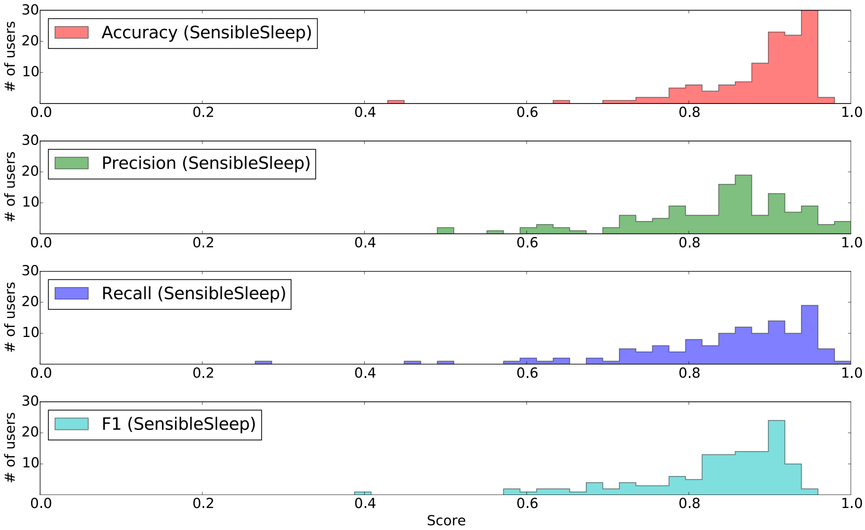


Fig 4: Histogram of the calculated accuracy, precision, recall and F1 score for users in dataset A, comparing the proposed method to the sleep tracker ground truth.

Fig. 5 shows the corresponding *complementary cumulative distributions* of the accuracy, precision, recall and F1 scores of the proposed SensibleSleep model vs that of the rule-based model [8]. The results are generally comparable between the two models, on this particular dataset. Our model has slightly better accuracy and precision whereas the previously suggested rule-based model has a slightly better recall. The F1 scores, which weights precision and recall equally, are comparable. This particular dataset has only very limited sleep interruptions during the night. For populations with more interrupted sleep, we expect our model to maintain a high score.

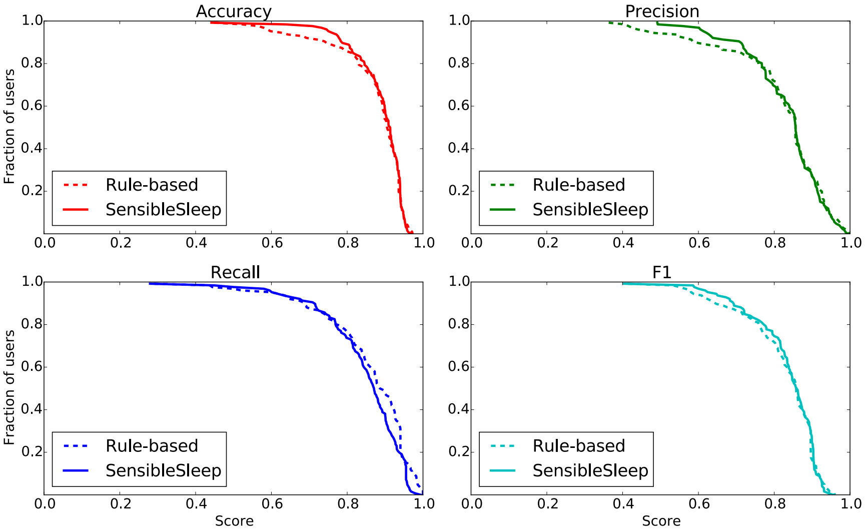


Fig 5: Complementary cumulative distribution of accuracy, precision, recall and F1 scores for users in dataset A, comparing the proposed model (solid line) to the rule-based model (dashed line), showing the proportion of users (y-axis) having a score less than or equal to a specific value (x-axis).

Individual Sleep Patterns

We now analyze individual sleep patterns to show the results of the model in details. For each user we create a visualization of sleep schedules. We call this the *sleep matrix*. Each row represents one day, and each column represents one time bin. The blue color shows the probability that sleep takes place within the interval; the darker the color the higher the probability. The red dots show activity count per bin; the larger the radius the more events are registered within that particular bin. This compact representation is able to capture at a glance the sleep patterns of individuals over time. We have created one such sleep matrix for each of the users, which allows us to inspect hundreds of sleep patterns quickly. Large individual variability both in sleep schedules (regular, irregular) and in phone activity (low, high, during day or night) are noticeable.

Still, in most cases it is evident that the model is able to capture a reasonable sleep period, even if it may have been somewhat interrupted.

Let us consider the inferred sleep patterns for two example users in Fig. 6. The top user has a pretty regular schedule, waking up around 5:30 except every few days, when he/she wakes up later – presumably due to vacation or weekends. Notice the light blue sections that indicate how the model is less confident about the probability of sleep due to events that do not follow the usual patterns. The bottom user instead has a much more unstable app usage, therefore the model infers a correspondingly more unstable sleep schedule. The bottom user has also some events in the middle of the night throughout many days (which is presumably checking the phone at night) yet the model is still able to correctly infer this being a sleep phase. Finally notice how the two users have significantly different intensity of app usage (the bottom one uses the phone much more than the top one), yet this is not a problem since the model learns individual activity rates.

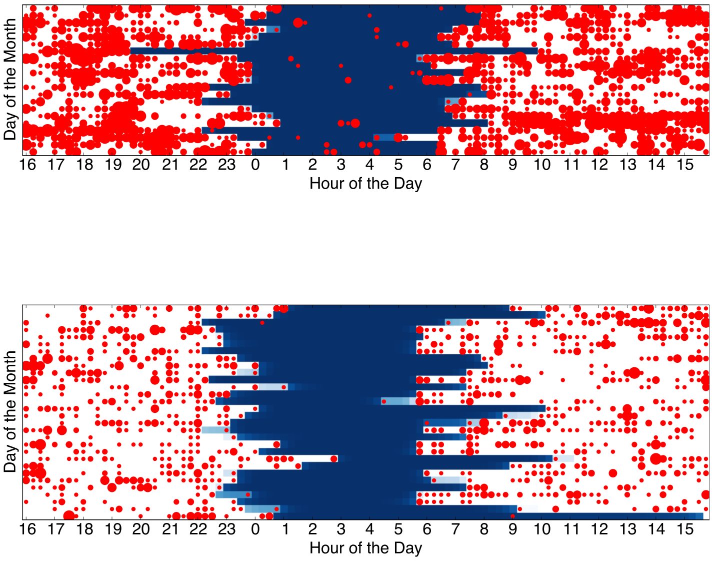


Fig 6: Sleep matrix of two sample users (21 days from dataset A top, 30 days from dataset B bottom)

	Sleep Time		Wake Time	
	Mean	(Std)	Mean	(Std)
A	23:38	(2h 16m)	7:40	(2h 2m)
B	0:35	(2h 6m)	7:55	(2h 15m)

Table 2: Aggregated sleep and wake times

Aggregated Sleep Schedules

In this section we also quantify the aggregated sleep patterns. From the posterior probability distribution functions (PDFs), $P_{t_{sleep}}(t)$ and $P_{t_{awake}}(t)$, the probability that the user is sleeping can be estimated as follows:

$$P_{sleep}(t) = P_{t_{sleep}}(t) - P_{t_{awake}}(t)$$

This is equivalent to stating that a user is currently sleeping if he has passed the time of falling asleep but has not yet passed the time waking up.

The derived values of sleep-length $t_{sleeplength}$ and mid-sleep time $t_{midsleep}$ can be calculated directly from the values of t_{sleep} and t_{awake} for each sample of the trace, and the posterior density can be estimated for these derived values in a similar way as for the model parameters. Fig. 7 shows the aggregate posterior probability density functions for t_{sleep} and t_{awake} for the 126 users of dataset A over $15-30$ days, and for the 324 users of dataset B over a selected period of 30 days (just after semester start). It may not be entirely meaningful to average the sleep patterns from all users, but it serves to illustrate the distribution of t_{sleep} and t_{awake} for a larger population. Table 2 summarizes the sleep and wake times.

Across the 30 (14-28) analyzed days for the 324 (126) users of the study, the distribution of sleep durations are as shown in Fig 8. The model allows us to easily compute such metrics. The mean value is around 8:02 (±2h 36m) for dataset A and 7:20 (±2h 28m) for dataset B. Notice how the distributions are not completely similar; this is likely due to the fact that the larger dataset B captures the sleeping behavior of students as opposed to dataset A that may have a more diverse demographic distribution.

Fig. 9 shows the probability density functions for the t_{sleep} and t_{awake} times for all users of dataset B, grouped according to weekday. Mondays to Thursdays appear quite similar, but Friday shows a much wider distribution; users typically go to bed much later on Friday and sleep in on Saturday. The distributions start to narrow down Saturday and Sunday but are more “week-like” only from Tuesday morning again.

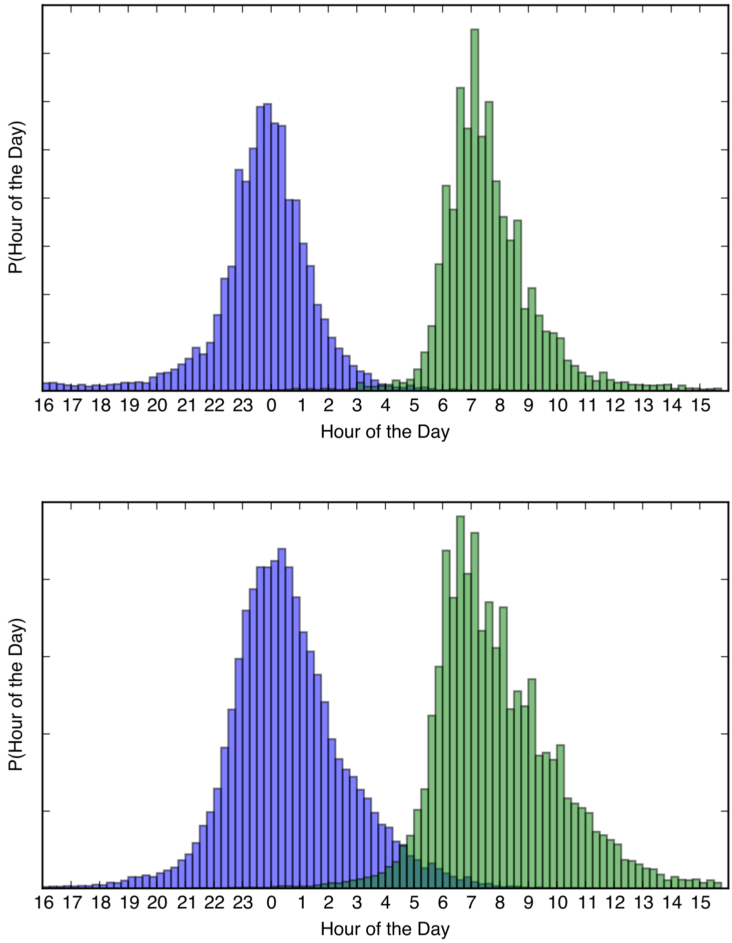
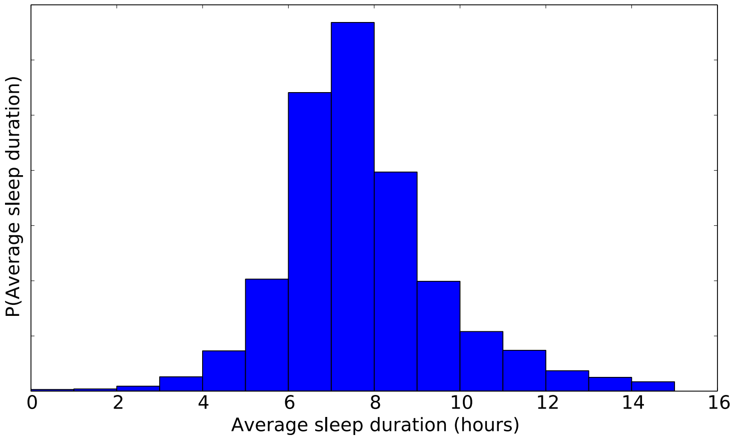


Fig 7: Aggregate Posterior Probability Distributions of t_{sleep} (blue) and t_{awake} (green) (A top, B bottom), showing what the probability is for the specific population to go to sleep or wake up at the specified time.



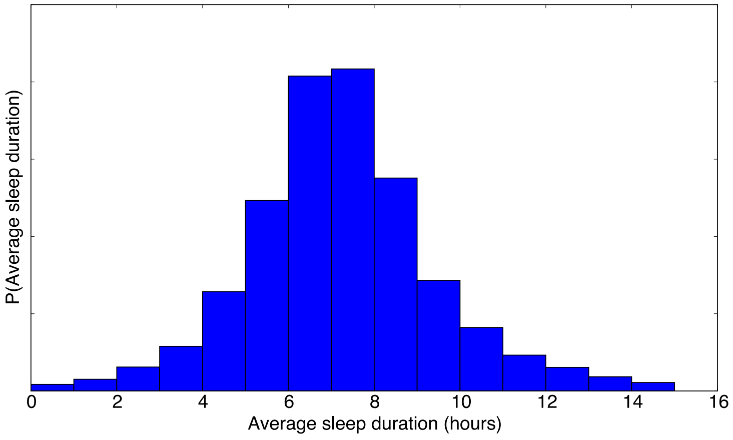


Fig 8: Aggregated Sleep Durations (A top, B bottom), based on the Posterior Probability Functions. This illustrates the probability of the length of a nights sleep within the population within the datasets.

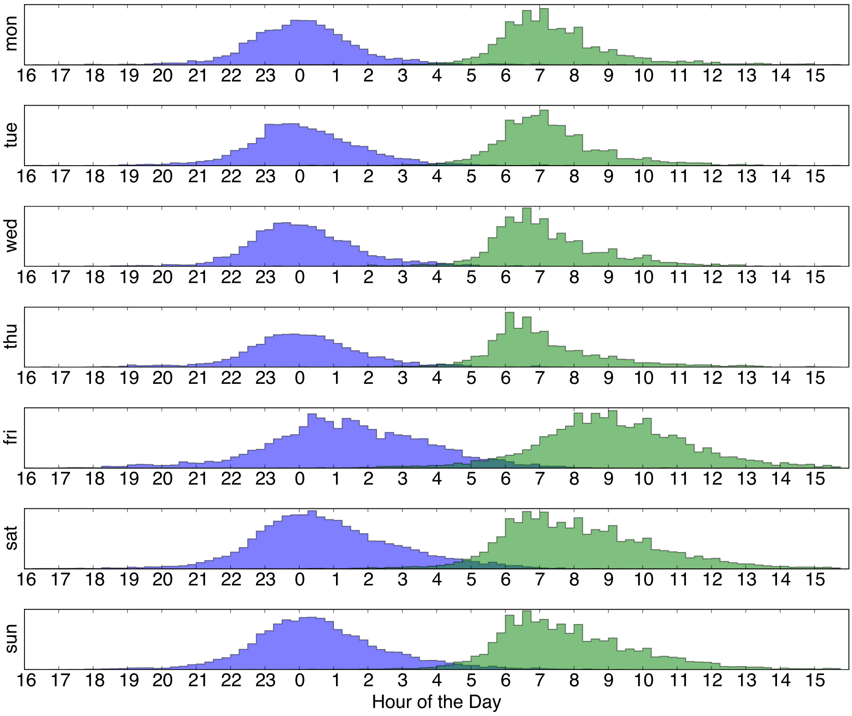


Fig 9: t_{sleep} (blue) and t_{awake} (green) over weekdays for dataset B

Discussion

The main contribution of this work is to show how simple counts of smartphone interactions can be used to infer sleep patterns with reasonably high accuracy. We have demonstrated how the seemingly weak signal of screen events carry significant information of the user status. Our method has several advantages:

- The method requires only a smartphone and can therefore be deployed without the need for special equipment or methods, such as fitness or sleep tracking bands, or sleep diaries.

- The data collection is completely automated, as no action is required from the user in setting up the tracking or remembering to log his/her activity.

- Since the model requires only screen interactions, it is absolutely non-intrusive and privacy-preserving. Although in this work we stored the data on a central server for analysis purposes, the data could remain on the phones and the sleep analysis could in principle be run directly on the phones as well.

- Compared to accelerometer or microphone-based methods, using only screen events is much more battery-efficient.

Although solutions using screen events have been proposed before [8,21], our model provides a number of key improvements:

- It is more robust to noise such as screen events generated by checking the phone at night.

- Using a Bayesian formulation allows us to provide confidence intervals for the sleep and awake times, instead of point estimates only.

- It does not depend on ad-hoc rules, but it is based on a well-defined statistical formulation.

- It is fitted and verified on a much larger userbase of over 400 users, and a longer time duration (between 2 and 4 weeks).

Demonstrating the feasibility of inferring reasonable sleep patterns from simple event counts opens the way for new exciting research directions. In particular we believe that similar methods can be applied to large datasets of user activity. For example on social network (such as Twitter, Facebook, Meetup, Gowalla) users leave a trace of their activity in the form of messages, posts, likes, etc. Another great example is Call Detail Records, the logging information kept by telecom providers about user calls and SMS. These events could be treated again as a proxy for sleep and wake cycles.

The main drawback of the proposed method is that it requires that users periodically interact with their phones during their wake time. In line with other recent polls (see for example [29–31]), we show that in most cases this does happen, as the population of users analyzed here tend to check their phone from the early morning to the late night when awake. Different populations, however, such as elderly people less accustomed to smartphone usage, may not show similar usage patterns. There is therefore a need for additional work in order to understand how increased sparsity would affect sleep pattern reconstruction.

Conclusions

We have presented a Bayesian model to infer sleep patterns from smartphone interactions, which we have applied to two datasets of more than 400 users in total. We have compared the model output with ground truth from sleep trackers, and we have shown how the model is able to recover the sleep state with a mean accuracy of 0.89 and a mean F1 score of 0.83. Furthermore, we have shown how the model is capable of producing very reasonable individual and aggregated sleep patterns. Our method represents a cost-effective, non-intrusive and automatic alternative for inferring sleep patterns, and can pave the way for large-scale studies of sleep rhythms.

Acknowledgments

AC is funded in part by the High Resolution Networks project (The Villum Foundation), as well as Social Fabric (University of Copenhagen). PB is supported in part by the Innovation Fund Denmark through the project Eye Tracking for Mobile Devices. VS and HJ received funding from Sony Mobile. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Choe EK, Consolvo S, Watson NF, Kientz JA. Opportunities for Computing Technologies to Support Healthy Sleep Behaviors. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '11. New York, NY, USA: ACM; 2011. p. 3053–3062. Available from: http://doi.acm.org/10.1145/1978942.1979395.

2. Candia J, González MC, Wang P, Schoenharl T, Madey G, Barabási AL. Uncovering individual and collective human dynamics from mobile phone records. Journal of Physics A: Mathematical and Theoretical. 2008;41(22):1–11. doi:10.1088/1751-8113/41/22/224015.

3. González MC, Hidalgo CA, Barabási AL. Understanding individual human mobility patterns. Nature. 2008;453(7196):779–782. doi:10.1038/nature06958.

4. Sekara V, Stopczynski A, Lehmann S. The fundamental structures of dynamic social networks. arXiv preprint arXiv:150604704. 2015;.

5. Onnela JP, Saramäki J, Hyvönen J, Szabó G, Lazer D, Kaski K, et al. Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(18):7332–7336. doi:10.1073/pnas.0610245104.

6. Zheng Y, Zhang L, Xie X, Ma WY. Mining interesting locations and travel sequences from GPS trajectories. Proceedings of the 18th international conference on World wide web - WWW '09. 2009;(49):791. doi:10.1145/1526709.1526816.

7. Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, et al. Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study. Journal of Medical Internet Research. 2015;17(7):e175. doi:10.2196/jmir.4273.

8. Abdullah S, Matthews M, Murnane EL, Gay G. Towards Circadian Computing: "Early to Bed and Early to Rise" Makes Some of Us Unhealthy and Sleep Deprived. Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2014; p. 673–684. doi:10.1145/2632048.2632100.

9. Smart Alarm Clock;. `https://itunes.apple.com/us/app/smart-alarm-clock-sleep-cycles/id586910133?mt=8`.

10. Sleep Cycle;. `https://itunes.apple.com/us/app/sleep-cycle-alarm-clock/id320606217?mt=8`.

11. SleepBot;. `https://play.google.com/store/apps/details?id=com.lslk.sleepbot&hl=en`.

12. Sleep as Android;. `https://play.google.com/store/apps/details?id=com.urbandroid.sleep&hl=en`.

13. Zhang M, Tillman DA, An SA. Global prevalence of sleep deprivation in students and heavy media use. Education and Information Technologies. 2015; p. 1–16.

14. Orzech KM, Grandner MA, Roane BM, Carskadon MA. Digital media use in the 2 h before bedtime is associated with sleep variables in university students. Computers in Human Behavior. 2016;55:43–50.

15. Richter CP. Biological Clocks in Medicine and Psychiatry: Shock-Phase Hypothesis. Proceedings of the National Academy of Sciences of the United States of America. 1960;46:1506–1530. doi:10.1073/pnas.46.11.1506.

16. Aschoff J. Circadian Rhythms in Man. Science (New York, NY). 1965;148(3676):1427–1432. doi:10.1126/science.148.3676.1427.

17. Ren Y, Wang C, Yang J, Chen Y. Fine-grained sleep monitoring: Hearing your breathing with smartphones. In: Computer Communications (INFOCOM), 2015 IEEE Conference on. IEEE; 2015. p. 1194–1202.

18. Hao T, Xing G, Zhou G. iSleep: unobtrusive sleep quality monitoring using smartphones. In: Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems. ACM; 2013. p. 4.

19. Chen Z, Lin M, Chen F, Lane ND, Cardone G, Wang R, et al. Unobtrusive sleep monitoring using smartphones. In: 7th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth). IEEE; 2013. p. 145–152. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6563918.

20. Gu W, Yang Z, Shangguan L, Sun W, Jin K, Liu Y. Intelligent Sleep Stage Mining Service with Smartphones. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. UbiComp '14. New York, NY, USA: ACM; 2014. p. 649–660. Available from: http://doi.acm.org/10.1145/2632048.2632084.

21. Jayarajah K, Radhakrishnan M, Hoi S, Misra A. Candy Crushing Your Sleep. In: Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers. UbiComp/ISWC'15 Adjunct. New York, NY, USA: ACM; 2015. p. 753–762. Available from: http://doi.acm.org/10.1145/2800835.2804393.

22. Hossain HMS, Roy N, Khan MAAH. Sleep Well: A Sound Sleep Monitoring Framework for Community Scaling. In: Proceedings of the 2015 16th IEEE International Conference on Mobile Data Management - Volume 01. MDM '15. Washington, DC, USA: IEEE Computer Society; 2015. p. 44–53. Available from: http://dx.doi.org/10.1109/MDM.2015.42.

23. Sony SmartWear;. http://www.sonymobile.com/global-en/products/smartwear/.

24. Stopczynski A, Sekara V, Sapiezynski P, Cuttone A, Madsen MM, Larsen JE, et al. Measuring large-scale social networks with high resolution. PloS one. 2014;9(4):e95978. doi:10.1371/journal.pone.0095978.

25. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. vol. 2. Taylor & Francis; 2014.

26. Patil A, Huard D, Fonnesbeck CJ. PyMC: Bayesian stochastic modelling in Python. Journal of statistical software. 2010;35(4):1.

27. Fonnesbeck CJ. PyMC version 3; 2015. Available from: https://github.com/pymc-devs/pymc3.

28. Berg A, Meyer R, Yu J. Deviance information criterion for comparing stochastic volatility models. Journal of Business & Economic Statistics. 2004;22(1):107–120.

29. Tecmark survey finds average user picks up their smartphone 221 times a day; 2014. http://www.tecmark.co.uk/smartphone-usage-data-uk-2014/.

30. Newport F. Most U.S. Smartphone Owners Check Phone at Least Hourly; 2015. http://www.gallup.com/poll/184046/smartphone-owners-check-phone-least-hourly.aspx.

31. BankOfAmerica. Trends in Consumer Mobility Report; 2015. http://newsroom.bankofamerica.com/files/doc_library/additional/2015_BAC_Trends_in_Consumer_Mobility_Report.pdf.

Paper G

The Blank Stare

This paper was originally prepared as part of a course in Biometrics, and was later reworked and submitted to a biometrics conference on request. It was, however, not accepted due to the limited amount of participants but not due to problems with the method or analysis. The scope of the paper is not to validate a new method across many different subjects but focuses on the stability of the found signatures.

<Firstname Lastname [et. al.]>(Ed.): <Book Title>,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn <Year> 1

The Blank Stare: Retrieving Unique Eye Tracking Signatures Independent of Visual Stimuli

Per Bækgaard[1] Michael Kai Petersen[2] Jakob Eg Larsen[3]

Abstract: USING LOW COST PORTABLE EYE TRACKING FOR BIOMETRIC IDENTIFICATION OR VERIFICATION: Eye tracking technologies have in recent years become available outside of specialised labs, and are starting to become integrated in tablets and virtual reality headsets[4]. This offers new opportunities for use in common office- and home environments, such as for biometric recognition (identification or verification), alone or in combination with other technologies. This paper exposes two fundamentally different approaches that have been suggested, based on spatial and temporal signatures respectively. While deploying different stimulation paradigms for recording, it also proposes an alternative way to analyze spatial domain signatures using Fourier transformation. Empirical data recorded from two subjects over two weeks, three months apart, are found to support previous results. Further, variations and stability of some of the proposed signatures are analyzed over the extended timeframe and under slightly varying conditions.

Keywords: Biometric Systems, Eye Tracking

1 Introduction

Biometric identification and verification are applied and subject to active standardization work [IS13]; in particular within the areas of fingerprints (minutiae) and face- and iris images, and is extending to dynamic signature time series, vascular images, hand geometry silhouette, voice data, DNA data and palm crease images. Less work has seemingly been done on the inherently traceless, non-invasive and contactless *eye tracking signatures* like *fixations* and *saccades* [Ho11] for behavioural biometric purposes.

Temporal Features – Oculomotor Plant Model Analysis of Saccades: [Ko10], [Ko12] and [Ko14] however suggest methods for biometric identification and verification using eye trace recordings. These are largely based on observing eye movement traces from which individual *saccades* are used to estimate parameters of a proposed *Oculomotor Plant Model* (OPM), thus serving as an extracted feature vector from the biometric probe, comparable to previous feature vectors of one or more biometric references from a biometric enrolment database. In the more recent work [Ko14], *Equal Error Rates* (EER) of the best model was found at 20.3% and *Receiver Operating Characteristics* (ROC) *Area Under Curve* (AUC) levels of up to around 80% were observed in a study of 32 users performing a total of 122 unique recordings within a period of 2 weeks.

[1] Cognitive Systems, DTU Compute; Technical University of Denmark, DK-2800 Kgs. Lyngby; pgba@dtu.dk
[2] Cognitive Systems, DTU Compute; Technical University of Denmark, DK-2800 Kgs. Lyngby; mkai@dtu.dk
[3] Cognitive Systems, DTU Compute; Technical University of Denmark, DK-2800 Kgs. Lyngby; jaeg@dtu.dk

[4] E.g. at http://www.theeyetribe.com

Some of the key advantages of using such a system[5] are that it is difficult for a subversive impersonator in a presentation attack to spoof dynamic patterns that would result in a feature vector matching the intended reference; such a system will often include a liveness detection as salient on-screen stimuli can be used to trigger saccades that would be difficult for a non-human entity to mimic. Further, parameters of such a system could be estimated on a continuous basis, to ensure that a user carrying out trusted tasks is not replaced after verification. It could also be speculated that the collectability and acceptability might be high, due to the unobtrusive and readily available nature of the technology.

Spatial Features – Fixation Density Map Analysis: The method(s) above work best when a large number of saccades can be evoked and analysed. Where the OPM based proposal(s) mainly rely on exciting, and estimating characteristics of, the dynamic eye-brain system by presenting salient stimuli, the work of [RK14] proposes the use of *fixation density maps* (FDM) for biometric identification or verification in a way that does not rely on invoking specific eye movements. It rather explores characteristics of the spatial domain, which [BPL15] also found surprisingly stable within two subjects during a week of repeated experiments.

[RK14] analyses the FDM that result from observing stimuli in a free-viewing condition, noting that the FDMs also embeds some temporal information as the density grows with the amount of time the eye remains within an area. The FDMs of 15 subjects observing faces, 100 subjects reading text and the same 100 subjects observing video-scenes were collected and analysed, and a min-based metric as well as Euclidean Kullback-Leibler divergence (KLD) *dissimilarity scores* were used to compare the FDMs and to compute *dissimilarity matrices*. The effect of applying different size gaussian kernels on the FDM were also evaluated. In the best possible configuration, EER rates of 18.3% were reported, with variations up to 34.5%, but still above chance level.

2 Biometric Brain-Eye Feature Extraction

Extracted temporal- and spatial features complement each other well. Whereas the temporal features to a large degree reflect the dynamic responses of the brain-eye system with its neuronal-nerve-muscle interactions, the spatial features are, to a much larger degree, independent of many of the OPM parameters, and thus may reflect more of the processes related to components of the *attention networks* in the brain.

The present work looks at both temporal- and spatial feature extraction and comparisons, using low-cost eye tracking equipment in office conditions where subjects are free to move around somewhat and/or be distracted temporarily by other events in the environment. In addition to replicating selected aspects of previous works, it also hints at the longer-term stability of the analysed features, and proposes applying a *Discrete Fourier Transformation* (DFT) to better highlight desired attention-driven qualities and find similarities in the spatial domain.

[5] Other, although apparently few, proposals that also explore biometric features extraction from eye movements exists; see the introduction section of [RK14] for a summary of some of the key ideas.

Temporal (Bottom-Up) Features – Saccades and Time-to-Target: For the temporal domain, a simple feature extraction and analysis is done: From the OPM model, it can be expected that the effective *time-to-target* (TTT) or *saccadic latency*, measured from presentation of a salient stimuli until the first eye fixation has reached an area nearby, will depend on the direction of the eye movement and will vary between individuals. The hypothesis is that these TTT measures should remain relatively stable within subjects even over varying conditions and over longer periods.

Spatial (Top-Down) Features – Fixation Density Map Metrics: For the spatial domain, this present work proposes a novel approach to analysis by applying a DFT to the FDM, which here appears to improve the comparison performance significantly.

Whereas [RK14] calculate spatial domain dissimilarity metrics directly from the resulting FDM after a gaussian kernel convolution, it is here proposed to first apply a DFT transformation, discarding the phase information and then use a box-filter to extract only the lower spatial-frequency components of the resulting spectrum before calculating feature vector distances. The filtering is not unlike applying a gaussian filter to the FDM, but the process as proposed discards translations of the FDM and focuses more on the spatial-frequency distributions instead of the actual FDM areas themselves, and may thus emphasize eye patterns, including some microsaccades, used when exploring areas of interest.

It is hypothesised that as the observed scene(s) present areas of interest in a given spatial configuration, the brain may explore different temporal and spatial combinations of these. The resulting FDM will likely not only differ between two users observing the same scene, but also to some extent when one user observe the same scene multiple times. However, if the top-down driven eye movements are more a result of individual preferences or characteristics related to attention governing the way a particular scene or areas of interest are explored, it might be possible to better extract and characterize some of these patterns in the DFT'ed domain.

The DFT function $F_{a,b}$ of a NxN square 2D image, $f_{x,y}$, is:

$$F_{a,b} = \sum_{x,y}^{N} f_{x,y} e^{-i2\pi(\frac{ax}{N}+\frac{by}{N})}$$

and applying a box-filter simply means assigning $F_{a,b} = 0$ when $|a| > l \vee |b| > l$ for some limit l, chosen suitably to the size of the FDM (values around 5% of the original image size has been used in the present case).

For comparison with [RK14], a similar gaussian kernel is applied to the FDMs, but this is not strictly required for the DFT distance metrics to be calculated; only for the FDM domain distance metrics they are essential.

To compare the extracted 2D NxM features, consider a P (probe) and a R (reference) – both based on either a FDM directly or a DFT'ed FDM – for which the distance metrics are calculated. This will be done similarly to [RK14], with the addition of a simple *Mean Square Error* (MSE) metric.

Before calculating any of the distance metrics, it is ensured that the feature vectors are normed independently such that $\sum_{x,y} P_{x,y} = 1$ and $\sum_{x,y} R_{x,y} = 1$. The distance metrics are then as follows:

The *Mean Square Error* (MSE) distance calculates the sum of the square of the difference at each 2D point in the feature space:

$$D_{\mathrm{MSE}}(P,R) = \sum_{x,y} (P_{x,y} - R_{x,y})^2$$

The *1-MIN based Similarity Metric* (1-Min) distance calculates the sum of the min value of either P or R at each 2D point in the feature space:

$$D_{\text{1-MIN}}(P,R) = \sum_{x,y} \min(P_{x,y}, R_{x,y})$$

It is in this case, where the FDMs could be thought of as representing a probability function, somewhat similar to a continuous version of the Hamming distance that is conventionally deployed when only binary point-to-point comparisons are possible.

The *Kullback-Leibler divergence* [KL51] (KLD), or relative entropy, is a non-symmetric measure defined as follows:

$$D_{\mathrm{KL1}}(P,R) = \sum_{x,y} P_{x,y} ln \frac{P_{x,y}}{R_{x,y}}$$

but in this present case, disregarding any performance considerations, a symmetric distance is desirable, and hence a value proportional to the harmonic mean of $D_{\mathrm{KL1}}(P,R)$ and $D_{\mathrm{KL1}}(R,P)$ is used, i.e.

$$D_{\mathrm{KLD}}(P,R) = \frac{2}{\frac{1}{D_{\mathrm{KL1}}(P,R)} + \frac{1}{D_{\mathrm{KL1}}(R,P)}}$$

Finally, the *Euclidean* (Eucl) distance is calculated, based on not only comparing two feature sets from FDMs or DFT'ed FDMs but using the entire previously calculated KLD dissimilarity matrix[6] as follows: Since the D_{KLD} in our case is symmetric, the dissimilarity matrix will also be symmetrical. Each row (or equivalently column) can be considered as a feature vector that holds the D_{KLD} distance from the corresponding probe to all other templates[7]. This can be compared to all other similar feature vectors by computing the

[6] As the D_{Eucl} distance requires computing the entire dissimilarity matrix based on all participating feature vectors, it may be less suited to large biometric databases. It may work best with neither too small nor too large template databases against which a probe is to be compared. Also, the Euclidean distance between two feature vectors will change according to how many and which particular feature vectors are included in the total comparison, so a cut-off threshold learned from a training set cannot directly be applied to another test set; the classifications derived from the training set will have to be used to recalculate a new threshold value in the combined training and test sets.

[7] Fom this calculation, there is no difference as to whether compared samples are other probes or references.

Euclidean distance between each vector pair. Hence, a new dissimilarity matrix can be built by computing:

$$D_{\text{Eucl}}(D_{\text{KLD}}) = \text{Euclidean}(D_{\text{KLD}}, D_{\text{KLD}}^{T})$$

as proposed by [RK14][8] where it also is subsequently normed so that $max(D_{\text{Eucl}}) = 1$

3 Experimental Setup

Two male right-handed subjects, average 55 years old, participated in the trials, which took place over several weeks[9] [10] 3 months apart[11]. In each week, the experiment was executed one or more times during most of the weekdays, at alternating hours and between two different everyday offices. In total 34 trials were run; 16 with subject A and 18 with subject B. The subjects were not instructed to follow any specific viewing patterns.

During the experiment 24 sequences were presented, where 8 colored squares (3 degrees wide) were sequentially presented as salient stimuli on the screen, each sequence alternating between the colors blue, yellow, green, yellow, white and black. Each presented square appeared for 2 seconds against their complementary color as background. This was followed by 4 seconds of solid complementary color[12], in total 480 secs of visual stimuli for each of 34 experiments performed, in aggregate 2h16m of stimuli for each participant.

The trial stimuli were presented on a conventional MacBook Pro 13â at 60 Hz screen refresh rate, running the PsychoPy software [Pe07]. The Eye Tribe mobile eye tracking device, connected via USB, was used to retrieve the eye traces through the associated API [Th16], using PeyeTribe [Bæ15], and was running at 60 Hz. An initial calibration, using the vendor provided interface, was performed at the beginning of each experiment.

The raw traces from the trials were first analysed to identify fixations by applying a DBSCAN [Es96] derived density-based clustering approach. They were then epoched according to the presence or lack of the stimuli squares, and FDMs were built from the fixations that occurred during the combined 4-second periods where only a solidly coloured background was present. This was used for the top-down FDM analysis.

For the bottom-up time-to-target analysis, the appearance or relocation of a stimuli square was used as the basis for calculating the time to the first fixation hereafter nearby the presented square. To remove outliers and be able to identify those occasions where the reaction could be assumed to be the result of a bottom-up response to the stimuli, time-to-target values of less than 0.1 sec or more than 0.4 secs were discarded as outliers[13].

[8] In this case, D_{KLD} is symmetric and square, so the transposed matrix will be identical to the original matrix.

[9] This footnote removed in order to keep the present work anonymous.

[10] The first week of the experiment has been described and qualitatively analysed in [BPL15].

[11] Although this does not prove much in terms of uniqueness or the overall permanence of the extracted features, it is nevertheless based on a longer period compared to previously noted work.

[12] Hence the title of this paper THE BLANK STARE.

[13] [Ho11] cites typical saccadic latencies of 200ms and never shorter than 100ms; the latter only found when attention is released before onset of a new stimuli (the *gap effect*), which is not the case in this present study.

Apparent inaccurate calibrations, resulting in a shift of the fixation locations compared to the presented stimuli, are further compensated: A simple affine transformation can be calculated by minimizing the MSE from each measured vs target position, and can subsequently be applied to the fixations. These are the *Recalibrated* scenarios, tested to see if they would result in better alignment and more comparable FDMs across trials.

4 Results

Temporal Features – Time-to-Target: Fig. 1 shows the resulting median time-to-target values over all trials for both subjects. The distributions are different, but wide and with a long tail towards higher values. An individual time-to-target value cannot be contributed to either subject with confidence.

Table 1 show mean, median and standard deviation values for the time-to-target of all trials. As can be seen, even between trials, the distribution is quite wide and the differences between the mean or median values between subjects is of the same magnitude as the standard deviation. Hence, when comparing two trials, the difference is not statistically significant. The same data is also shown graphically in Fig. 2.

Fig. 3 shows time-to-target when only looking at saccades in the upwards, right, downwards and left direction, respectively. The distributions are different within-subject when comparing the downward direction with the other three; for both subjects, downwards time-to-target values are somewhat longer (around 50ms) although the distributions still overlap. The other three time-to-target values are very similar within-subject.

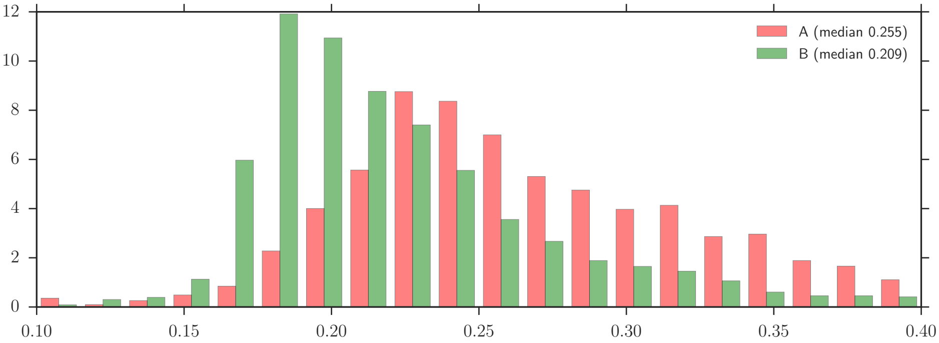


Fig. 1: Time-to-target histogram (all trials) for subject A (red) and subject B (green). The median value for subject A is 0.255s and for subject B 0.209s.

Spatial Features – Fixation Density Map Metrics: Examples of normal FDMs from both subjects[14] are shown in Fig. 4 with their corresponding DFT versions. They are generated from eye traces recorded during the solidly coloured screen epochs and appear very

[14] A parallel study with multiple participants, not included in this experiment, indicate that the FDM patterns across subjects are similarly characterized by the amount of central symmetric diffusion, spatial scattering, and vertical/horizontal asymmetrical skewness in their distribution of fixations.

	Trial	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
A	N	151	152	112	119	131	147	139	143	154	148	153	75	60	152	93	119		
	Mean	0.255	0.255	0.261	0.256	0.260	0.270	0.264	0.263	0.258	0.256	0.265	0.256	0.269	0.270	0.298	0.276		
	Median	0.241	0.245	0.250	0.249	0.249	0.263	0.257	0.258	0.252	0.246	0.258	0.248	0.265	0.262	0.300	0.261		
	σ	0.054	0.048	0.055	0.056	0.053	0.055	0.054	0.059	0.048	0.059	0.056	0.050	0.059	0.056	0.056	0.055		
B	N	167	165	174	180	177	170	175	170	166	171	169	170	169	177	166	173	162	171
	Mean	0.224	0.213	0.215	0.206	0.209	0.209	0.230	0.207	0.209	0.219	0.209	0.225	0.257	0.220	0.222	0.227	0.226	0.247
	Median	0.214	0.203	0.205	0.199	0.207	0.202	0.227	0.198	0.202	0.208	0.193	0.210	0.246	0.205	0.209	0.218	0.208	0.233
	σ	0.045	0.046	0.042	0.036	0.032	0.037	0.045	0.042	0.040	0.043	0.049	0.049	0.053	0.059	0.052	0.046	0.058	0.051

Tab. 1: Time-to-target values (all trials). N is the number of correctly identified target fixations within the 0.1s to 0.4s window. Mean, median and Standard Deviation are also listed.

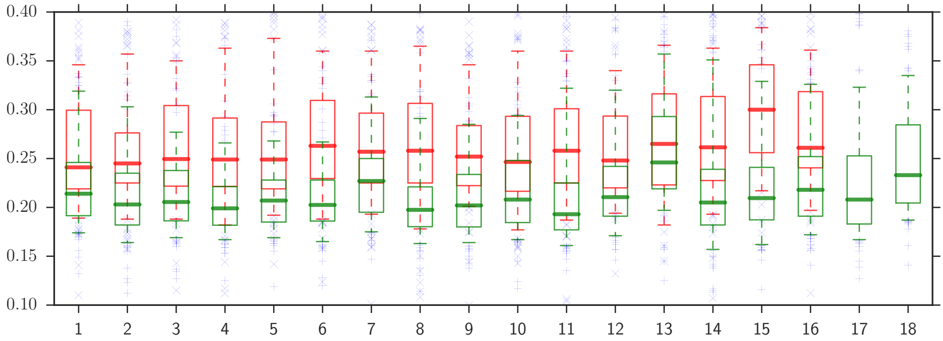


Fig. 2: Time-to-target for subjects A (red) and B (green). The bold line is the median, boxes outline quartiles, whiskers 5% and 95% confidence intervals. Numbers on the x-axis are for convenience.

different. The maps shown are somewhat typical for each subject, and although there is variation within-subject, they can generally be classified correctly by a human observer.

Dissimilarity matrices are calculated for the 4 metrics (MSE, 1-Min, KLD and Eucl), for all trials in the first week, for all in the second week and for all in both weeks combined. An example of the best comparison matrices can be seen in Fig. 7. From these, ROC and Detection Error Trade-Off (DET) curves are plotted by varying the detection threshold levels, and the accuracy (ACC, proportion of correctly classified outcomes) at the maximum value of F1 (the harmonic mean of precision and recall) is calculated, as is the Area Under Curve (AUC) for the ROC curve and the Equal Error Rate (EER)[15]. The ROC and DET curves corresponding to the matrices in Fig. 7 can be seen in Fig. 6. All of the computed performance metrics are listed Table 2.

As can be seen, the DFT based spatial-frequency approaches generally show a significantly better performance than basing the comparison on the spatial domain of the original FDMs only, and the 1-Min and Eucl metrics consistently[16] show the best performance in all four domains. The recalibration has a significantly positive effect on the comparison performance in the spatial domain in week 2, but deteriorates the performance of week 1

[15] F1, accuracy, area under curve and equal error rates are computed as conventionally.

[16] This holds when looking at optimal threshold values based on, and used within, the same analysed set of of feature vectors.

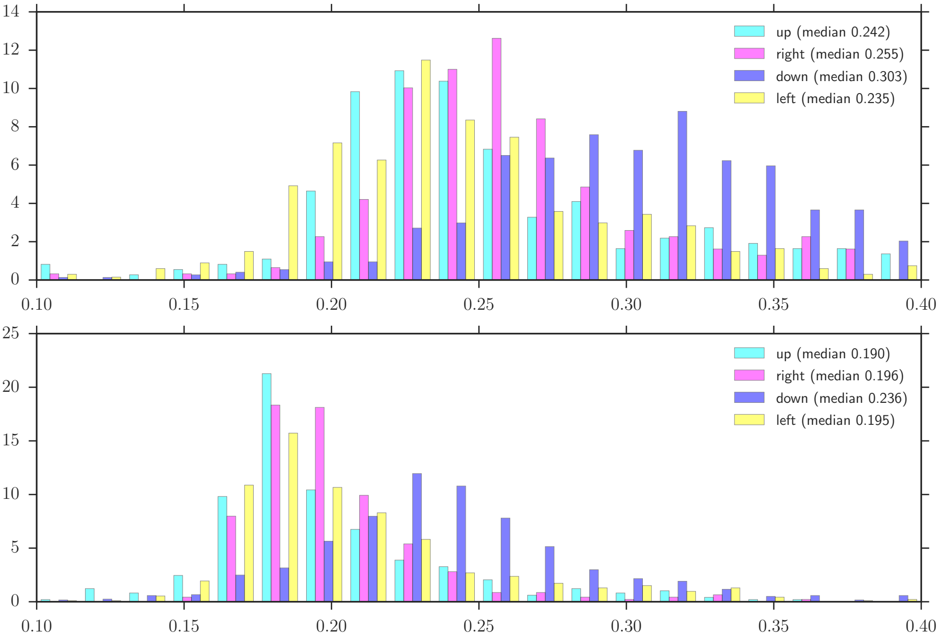


Fig. 3: Time-to-target histogram over all trials for subjects A (Top) and B (Bottom). Cyan, magenta, blue and yellow bars correspond to upwards, right, downwards and left saccades, with medians 0.242s, 0.255, 0.303s and 0.235s for subject A and 0.190s, 0.196s, 0.236s and 0.195s for subject B.

spatial domain comparisons[17]. The effect of the re-calibration is also inconclusive in the spatial-frequency domain, but has a much smaller effect here.

The best performing set of dissimilarity scores show an EER of 0.7% within week 1 and 0% (no errors in classification) within week 2; this variation is within the expected statistical variation for a system operating at this level. Between the two weeks, the best result is at 2.4%. This does *not* however imply that a system trained on one week would automatically give excellent results when tested on the other weeks data. As can be seen in Fig. 5, the optimal threshold value resulting in lowest error levels for one week neither always coincides with that of the other week nor that of the weeks combined.[18]

The relatively stable and similar performance between and over weeks might give hints on the stability of the FDMs, although the very limited number of subjects makes it impossible to draw any firm conclusions.

Other features extracted from the FDMs have also been tested, such as the spatial variance and skewness of the distributions (for instance, subject A has a tendency to have more

[17] This could point to the need for an improved re-calibration routine in case there are outliers in the data.

[18] There are differences in the behaviour of the metrics: MSE shows inconsistent alignment of optimal threshold values between weeks in the spatial domain but appear more consistent in the DFT domain, where all metrics generally appear more consistent. This mandates further study before one metric can be recommended in favour of the others.

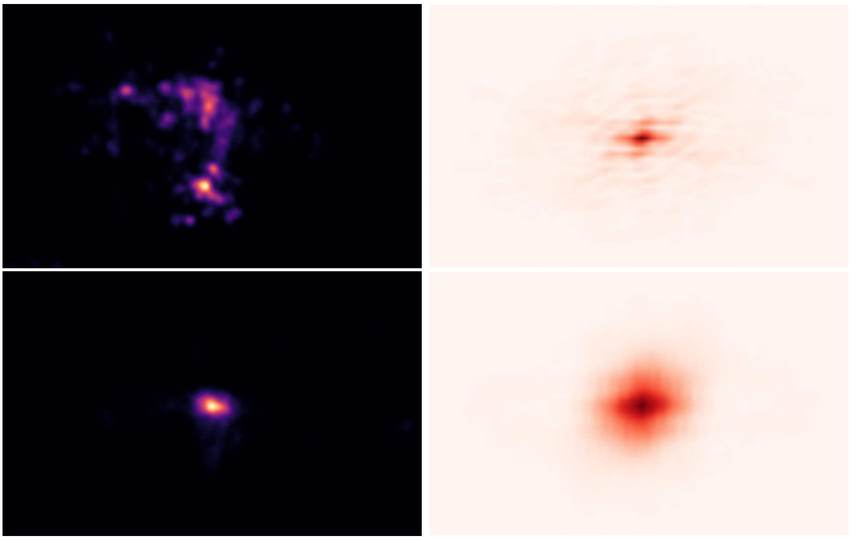


Fig. 4: (LEFT) Typical FDM recorded during presentation of solidly coloured screens for subjects A (TOP) and subject B (BOTTOM). (RIGHT) DFT'ed FDM corresponding to the ones to the left.

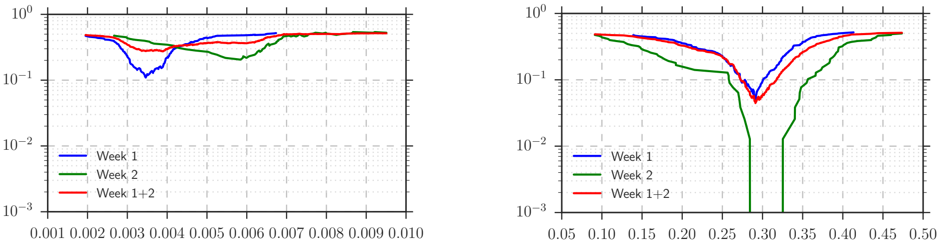


Fig. 5: Examples of total error levels $(1 - \text{ACC})$ vs threshold value for (LEFT) 'misaligned' MSE based comparisons for original DFMs and (RIGHT) 'better aligned' 1-Min based comparison in the DFT domain. Blue, green and red denote week 1, week 2 and weeks 1+2 combined respectively.

fixations to the left side of the screen), but even though there is a clear clustering of data values for each subject they apparently do not provide better comparison scores, and also hold lower entropy, and hence will not be discussed further here.

5 Conclusion

It has been demonstrated that using a DFT FDM spatial-frequency domain based approach in some instances can give improved performance to biometric recognition systems based on dissimilarity scores, compared to earlier proposed gaussian-filtered FDM spatial domain dissimilarity scores. Performance over the three month period achieved an EER of 2.4% overall and 0.7%/0.0% within each individual week in the best possible case, with corresponding AUC scores of 99.6% and 100.0% within weeks.

		Week 1			Week 2			Week 1+2		
		ACC	AUC	EER	ACC	AUC	EER	ACC	AUC	EER
FDM	MSE	89.0	89.7	15.5	79.5	71.5	35.3	72.2	73.6	34.5
	1-MIN	85.7	92.2	15.0	87.2	94.1	16.0	79.3	85.9	21.7
	KLD	76.2	84.8	26.0	83.3	91.0	18.6	73.1	80.8	26.6
	Eucl	79.0	88.0	21.2	87.2	94.0	18.6	77.7	81.7	25.6
FDM’	MSE	79.5	89.7	21.2	97.4	99.5	3.2	74.3	82.3	27.2
	1-MIN	72.9	81.2	29.8	96.2	99.1	5.8	79.1	88.8	21.3
	KLD	72.9	74.9	36.0	91.0	92.9	10.9	76.1	83.0	25.9
	Eucl	76.2	84.0	29.3	94.9	97.2	9.6	85.6	92.0	15.8
DFT	MSE	94.3	98.6	6.4	100.0	100.0	0.0	91.8	96.0	8.5
	1-MIN	90.5	96.8	10.7	100.0	100.0	0.0	93.0	98.2	7.6
	KLD	89.5	96.4	11.2	100.0	100.0	0.0	93.0	98.0	8.3
	Eucl	99.0	100.0	1.7	100.0	100.0	0.0	97.7	99.6	2.4
DFT’	MSE	97.1	99.1	5.0	100.0	100.0	0.0	91.4	94.9	11.1
	1-MIN	95.2	98.6	6.9	100.0	100.0	0.0	95.5	99.0	4.5
	KLD	92.9	97.8	9.7	100.0	100.0	0.0	94.8	98.8	5.6
	Eucl	99.5	100.0	0.7	100.0	100.0	0.0	97.7	99.6	2.4

Tab. 2: Effects of using different dissimilarity scores and domains, showing how well the biometric samples are correctly classified as a match or a non-match, compared to ground truth of mated vs non-mated origins. The Accuracy (ACC) is given at the threshold value with the highest F1 score. (AUC) is the Area Under Curve and (ERR) the Equal Error Rate. The metrics are based on comparing either original FDMs (FDM), recalibrated FDMs (FDM’), DFT’ed FDMs (DFT), or DFT’ed recalibrated FDMs (DFT’), with each dissimilarity score: MSE, 1-Min , KLD , and Eucl. Trivially identical comparisons are not included in the metrics.

No significant differences between weeks, time of day, office location or environmental condition was noted, hinting at the stability of the signatures within each individual even over extended periods.

Differences in simple time-to-target values between subjects were also demonstrated, as were directionally dependent differences, and although they also appeared stable over time and conditions, they did not allow statistically significant independent recognition of the subjects. Hence, in a setup similar to the present, with low resolution eye tracking equipment in varying environmental conditions, a temporal-spatial approach deploying the full OPM might be required; simple time-to-target methods might initially be more useful as part of liveness detection.

However, as the top-down and bottom-up modes complement each other, inherently reflecting different human behavioural systems, combining dissimilarity scores based on both would suggest itself as an enhancement compared to treating the two domains independently and should receive continued research efforts. As a means to solicit suitable

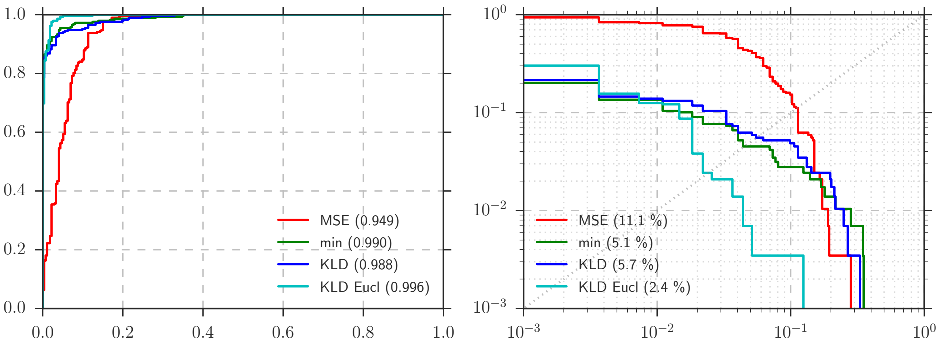


Fig. 6: Best performing ROC and DET curves, corresponding to Fig. 7, with dissimilarity scores: MSE (RED), 1-Min (GREEN), KLD (BLUE), and Eucl (CYAN) over DFT'ed FDMs across all trials. The ROC curve (LEFT) plots the *True Positives* (y-axis) vs the *False Positives* with resulting AUC of 94.9%, 99.9%, 98.8% and 99.6% respectively. The DET curve (RIGHT) plots the *False Negatives* (y axis) vs the *False Positives* on logarithmic scales. The intersection with the line from origo to $(1,1)$ is the point of Equal Error Rate (EER), of 11.1%, 4.5%, 5.6% and 2.4% respectively. See Tab. 2.

eye trace responses, a combination of salient stimuli with a known timing and non-salient stimuli, even solidly coloured screens, appears within reason and should also be explored further.

References

[Bæ15] Bækgaard, P.: , Simple python interface to the Eye Tribe eye tracker. http://github.com/baekgaard/peyetribe/, 2015. Accessed: 2015-06-21.

[BPL15] Bækgaard, Per; Petersen, Michael Kai; Larsen, Jakob Eg: Thinking Outside of the Box or Enjoying Your 2 Seconds of Frame? Universal Access in Human-Computer Interaction. Access to the Human Environment and Culture, 9178:186–195, 2015.

[Es96] Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). volume 96. AAAI Press, pp. 226–231, 1996.

[Ho11] Holmqvist, Kenneth; Nyström, Marcus; Andersson, Richard; Dewhurst, Richard; Jarodzka, Halszka; Van de Weijer, Joost: Eye tracking: A comprehensive guide to methods and measures. Oxford University Press, 2011.

[IS13] ISO 19794-2 to 19794-15: Information technology – Biometric data interchange formats – Parts 2 through 15. ISO/IEC 19794–2 through 19794–15, International Organization for Standardization, Geneva, Switzerland, 2007,2011,2013.

[KL51] Kullback, S.; Leibler, R. A.: On Information and Sufficiency. Ann. Math. Statist., 22(1):79–86, 03 1951.

[Ko10] Komogortsev, Oleg V; Jayarathna, Sampath; Aragon, Cecilia R; Mahmoud, Mechehoul: Biometric identification via an oculomotor plant mathematical model. In: Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications. ACM, pp. 57–60, 2010.

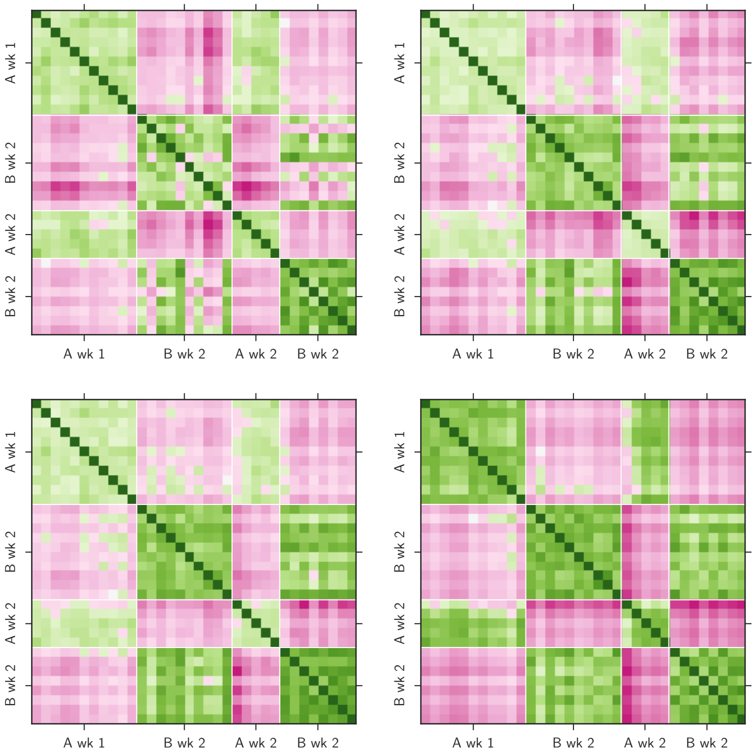


Fig. 7: Dissimilarity Matrices based on scores: MSE (TOP LEFT), 1-Min (TOP RIGHT), KLD (BOTTOM LEFT), and Eucl (BOTTOM RIGHT) over all DFT'ed FDMs. Each matrix is divided into 4x4 groups. The first two groups (TOP and LEFT) represents week 1, and the last two groups (BOTTOM and RIGHT) week 2. Within weeks, the first subgroup is subject A and the last subject B. Each row/column represent a single trial. Stronger red indicates higher scores (decreasing similarity); darker greens indicate a lower score. The threshold between red and green corresponds to the highest f1 score achievable. See Fig. 6 for corresponding ROC and DET curves.

[Ko12] Komogortsev, O.V.; Karpov, A.; Price, L.R.; Aragon, C.: Biometric authentication via oculomotor plant characteristics. In: Biometrics (ICB), 2012 5th IAPR International Conference on. pp. 413–420, March 2012.

[Ko14] Komogortsev, Oleg; Holland, Corey; Karpov, Alex; Price, Larry R: Biometrics via oculomotor plant characteristics: Impact of parameters in oculomotor plant model. ACM Transactions on Applied Perception (TAP), 11(4):20, 2014.

[Pe07] Peirce, Jonathan W: PsychoPyâ–Psychophysics software in Python. Journal of neuroscience methods, 162(1):8–13, 2007.

[RK14] Rigas, Ioannis; Komogortsev, Oleg V: Biometric recognition via fixation density maps. In: SPIE Defense+ Security. International Society for Optics and Photonics, pp. 90750M–90750M, 2014.

[Th16] TheEyeTribe: , API Reference eyetribe-docs. http://dev.theeyetribe.com/api/, 2016. Accessed: 2016-06-11.

Paper H

Your Heart Might Give Away Your Emotions

This paper was presented by Michael Kai Petersen at the 1st International Workshop on Multimedia Affective Computing in Chengdu China, 2014.

The author of this thesis is not the primary author on the paper, but co-supervised the work and took part of developing methods, analysing data and writing the paper.

YOUR HEART MIGHT GIVE AWAY YOUR EMOTIONS

David Kristian Laundav, Camilla Birgitte Falk Jensen, Per Bækgaard,
Michael Kai Petersen and Jakob Eg Larsen

Cognitive Systems, DTU Compute, Building 321,
Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark
s083130@student.dtu.dk,{cbfj,pgba,mkai,jaeg}@dtu.dk

ABSTRACT

ESTIMATING EMOTIONAL RESPONSES TO PICTURES BASED ON HEART RATE MEASUREMENTS: Variations in Heart Rate serves as an important clinical health indicator, but potentially also as a window into cognitive reactions to presented stimuli, as a function of both stimuli, context and previous cognitive state.

This study looks at single-trial time domain mean Heart Rate (HR) and frequency domain Heart Rate Variability (HRV) measured while subjects were passively viewing emotionally charged images, comparing short random presentations with grouped sequences of either neutral, highly arousing pleasant or highly arousing unpleasant pictures.

Based on only a few users we were not able to demonstrate HRV variations that correlated with randomly presented emotional content due to the inherent noise in the signal. Nor could we reproduce results from earlier studies, which based on averaged values over many subjects, revealed small changes in the mean HR only seconds after presentation of emotional images. However for longer sequences of pleasant and unpleasant images, we found a trend in the mean HR that could correlate with the emotional content of the images. Suggesting a potential for using HR in single user Quantified Self applications to assess fluctuations over longer periods in emotional state, rather than dynamic responses to emotional stimuli.

Index Terms— Heart Rate Variability, Emotion, Cognitive State, Affective Computing

1. INTRODUCTION

With the increased availability of low cost biometric sensors (for instance the Zephyr BioPatch[1], Basis Band[2] and Fitbit[3]) comes a wish for determining a person's cognitive or emotional state based on simple, directly observable biometric markers, such as Heart Rate (HR), Galvanic Skin Response (GSR) or changes in Body Temperature. Many commercial health-care oriented solutions already utilize and combine these biometric data, collected by torso- or wrist-mounted sensors, enabling anyone to monitor or track various aspects of their everyday life. This aspect is a central part of the *Quantied Self*[4] (QS) movement, which – in contrast to scientific studies of a large number of subjects – provides only a limited although often instantaneous analysis of ones own biometric markers. However, most existing applications which grasp the concept of QS do currently not make any attempt at deducing the emotional or cognitive state/response of the user. Additionally, on the basis of utilising the full potential of HR components, only the mean HR is currently considered in such applications.

Even if we no longer view the heart as the sole origin of emotions, several studies have documented links between emotional state and HR or Heart Rate Variability (HRV). Emotions are often ranked in two different dimensions, valence and arousal. Valence orders emotions from highly negative and unpleasant through neutral to highly positive and pleasant, whereas arousal orders emotions from boring and non-engaging to highly exciting.

HR is measured in beats pr. minute (bpm), and is for healthy adults at rest typically between 60 and 100 bpm. It is conventionally measured by counting the beats over a period of, say 30 seconds, then calculating the corresponding number of beats to a full minute. However, an instantaneous HR can at any time be calculated, and is inversely proportional to the current beat-to-beat period, which is often determined as the time interval between two subsequent specifically named peak points, "R", in the electrocardiogram (the "RR" time).

In addition to looking at the HR as bpm in the time domain, the frequency domain power spectrum, here referred to as HR(V), contains a lot of information about the subject: The HR is modulated both by the sympathetic and the bidirectional cardiac vagal (parasympathetic) nervous systems [1].

[1] *Zephyr BioPatch*: http://www.zephyranywhere.com/products/biopatch/ (Online; accessed 31-March-2014)

[2] *Basis Band*: http://www.mybasis.com/ (Online; accessed 01-April-2014)

[3] *Fitbit*: http://www.fitbit.com/uk (Online; accessed 01-April-2014)

[4] *Quantified Self*: http://quantifiedself.com/ (Online; accessed 01-April-2014)

Connected brain networks that deal with emotions can therefore impact HR through increased activity of the sympathetic nervous system, which leads to an increased HR, whereas increased activity of the parasympathetic nervous system leads to a decreased HR [2]. Further, the impact of the sympathetic nervous system is primarily seen in the HR(V) power spectrum below 0.15 Hz whereas the parasympathetic regulations are present in the entire spectrum and are thus the major contributor to the HF component above 0.15 Hz [1]. A person that is able to adapt well to presented stimuli will generally show modulations in this HF area, whereas absence or a low level of HF variation is indicative of an abnormal or stressed state that eventually might have health implications.

Even though LF variations in the power spectrum may not be directly linked to modulations caused by emotional responses, changes in the distribution and in the LF/HF ratio can vary over a longer timescale, reflecting changes to the subjects emotional and cognitive state.

The studies that have served as inspiration here are generally based on average values observed over multiple trials with up to 100 participants. Some focus mainly on trends of the mean HR whereas others look at variations in the frequency domain power spectrum of the HR(V); both do so as a function of the presented stimuli and over time.

Presenting unpleasant arousing pictures from a select set of 96 IAPS pictures to 27 students resulted [3] in an average HR deceleration of -1.76 bpm 3 s after onset, more that twice the impact observed when presenting neutral or pleasant pictures. In another study [4] [5] involving a more complex scenario (anticipatory imagined predator-prey survival), a somewhat longer 15 s HR deceleration for threatening (”gun”) and neutral scenarios followed an anticipatory HR acceleration for rewarding (”money”) and threatening scenarios. The deceleration was peaking at -3.5 bpm on average over the 29 participants analysed.

Both of these studies support findings of earlier studies [6], reporting even significant progressively decreasing HR for non-phobic subjects in response to a sequence of unpleasant pictures vs neutral or pleasant pictures, not unlike the ”fear bradycardia” found in animals [7].

Showing film clips or recalling memories to induce emotions were demonstrated [8] to cause a subduction of HF HRV power spectrum components compared to emotionally neutral activities; the biggest response was observed in films conveying a happy emotion. The observed response was further found to co-vary with activity in 4 specific brain networks, including the Medial Prefrontal Cortex (mPFC) BA10, which is linked to representing own emotions/mental state and relating that to those of others.

Even participants in a study listening to the H.C. Andersen ”The Ugly Duckling” fairy tale [9] demonstrated a significant positive correlation between the HR(V) LF/HF ratio and the perceived story intensity across the population of 21 participants of the study, but only hinted at trends related to the valence.

Other studies [2] are proposing HRV as an index into how the ”top-down” cortical-subcortical appraisal modulates the autonomic responses, and thus could serve as an important window into understanding stress and health.

It is therefore a justified expectation that presenting stimuli with varying emotional engagement will lead to emotionally co-varying responses. This could be observed as a decreasing HR or as a change in the LF/HF ratio of the HR(V) frequency domain power spectrum. Viewing arousing unpleasant pictures can be expected to cause a decelerating HR, while engaging in higher intensity/arousing emotional activities might show a lowered HF content, and consequently a higher LF/HF ratio (under comparable conditions), in the HR(V) frequency domain.

We have tried to reproduce these results of emotionally dependent HR and HR(V) responses to sequences of images in single run trials with only a few participants. In order to assess how reliable and stable such responses are, and how applicable they might be for not only predicting emotional responses, but more generally serve as a window into assessing fluctuating mental states, while taking into account the noise inherent to the HR signal.

2. METHOD

2.1. Chosen pictures

To do so, several sequences of previously user rated pictures are displayed to the participants. These pictures are part of a larger set from the International Affective Picture System (IAPS) [10], where the valence and arousal have been rated systematically and are provided as part of the set.

The pictures chosen for the trial were mainly portraying people in different situations. The selected pictures were categorized as either neutral or highly arousing unpleasant or pleasant, based on their valence and arousal rating. Unpleasant picture were all high arousal, with a content that ranged from severe burns, mutilated bodies and babies with tumors to model pictures with a perceived threatening content. No phobia inducing pictures (spiders, snakes, etc.) were chosen, to exclude potentially person-dependent effects hereof. Neutral pictures typically contained people with neutral facial expression, in common, everyday situations or non-engaging everyday objects. Pleasant pictures were likewise ranging from highly arousing erotic content to portraits of people getting married.

2.2. Experimental design

The experiment is divided into three continuous trials. The first displays 48 pictures in a random order, in regards to the valence category hereof. The two subsequent trials displays 18 respectively 30 pictures, in sequences of 3 respectively 5 of equal valence category. In doing so allows to both 1) inspect the contrast between showing pictures in a random order and in a sequence, and 2) whether or not the signal stabilises when increasing the amount of pictures shown in a sequence.

Prior to each presented picture, a grey-scale picture is first shown, then followed by a scrambled picture of identical luminance and finally the real IAPS picture; each phase lasts 6 seconds and thus in total 18 seconds. The 6 seconds were originally chosen based on the length of the response observed by [3]. No identical pictures were shown to the same subject during the three continuous trials.

2.3. Pictures displayed during the experiment

2.3.1. Trial one

In the first part of the experiment, 48 random pictures were presented for each participant where the stimuli consisted of 83 unique pictures. Their IAPS picture id is listed in the footnotes[5]. The collection of pictures consisted of 28 pleasant (mean valence/arousal = 7.1, 5.6), 25 neutral (mean valence/arousal = 5.0, 3.9) and 30 unpleasant (mean valence/arousal = 2.2, 5.9) pictures.

2.3.2. Trial two

In the second part of the experiment, 18 pictures were shown for each participant as a series of 3 pictures of 3 equal stimuli. The stimuli itself consisted of 48 unique pictures[6]. The collection consisted of 18 pleasant (mean valence/arousal = 7.1, 5.0), 12 neutral (mean valence/arousal = 4.9, 3.6) and 18 unpleasant pictures (mean valence/arousal = 2.2, 6.2).

2.3.3. Trial three

In the last part of the experiment, 30 pictures in total was shown for every participant as a series of 5 pictures with 5 equal stimuli. In total, the stimuli consisted of 63 unique pictures[7]. The collection of pictures in this trial consisted of 21 pleasant (mean valence/arousal = 7.0, 5.3), 17 neutral (mean valence/arousal = 4.9, 3.7) and 25 unpleasant pictures (mean valence/arousal = 2.2, 6.0).

2.4. Hardware

The pictures were presented to subjects sitting at rest upright in a chair, using a computer that ran a custom script from within Matlab. Data collection was done with *BioSemi Software*[8], using a BioSemi finger pletysmography sensor to determine RR-intervals and R-Onsets. Plethysmography should provide sufficient similarities to conventional Electrocardiograms in this (resting) context [11].

The pictures were displayed on a Viewsonic G90fB[9] which was placed approximately 50cm (19.69 in.) from where the subject sat. An oscilloscope was utilised to measure the delay of the picture onset, which proved to be insignificant. This procedure was performed both prior and after the trial was completed, no drift (i.e. increased delay) was found.

2.5. Data collection and reduction

During the experiment, biometric data as well as the onset timing for each picture were recorded. The data collection rate during the experiment, for collecting biometric data, was set to 1024 Hz to ensure a high resolution and thereby minimize the risk of losing data. After collecting the data, it was first imported to EDFbrowser[10] from where all R-Onsets (i.e. a detection of a peak in the electriocardiogram) were exported. Afterwards, the R-Onsets were imported to gHRV[11] for further analysis.

Data reduction and normalisation was performed as required, enabling analysis of the Mean HR and the HRV LF/HF ratio [1]: Outliers were first removed whereafter the data was interpolated in gHRV, to a sampling frequency of 4 Hz. The window size was set to 36 seconds, equalling the time it takes to present two sequences of pictures (scrambled and grey-scale pictures included) to reduce the inherent noise. The window shift was set to 3 seconds, to provide a high resolution. The power spectrum was calculated using the standard, built-in FFT method of gHRV.

Next, the *episodes* were imported, making it possible to view and analyse the data as a response to specific events; in this case, the valence category (pleasant, neutral or unpleasant).

[5] ’9341’, ’2730’, ’8420’, ’8041’, ’3030’, ’6838’, ’8380’, ’3530’, ’2870’, ’2220’, ’4653’, ’8465’, ’4611’, ’2590’, ’2800’, ’3015’, ’8116’, ’8320’, ’4617’, ’2810’, ’2200’, ’4666’, ’2205’, ’8090’, ’2208’, ’4659’, ’6834’, ’3100’, ’3500’, ’5621’, ’3053’, ’2383’, ’3180’, ’3181’, ’9405’, ’2890’, ’8300’, ’2752’, ’2260’, ’7550’, ’2240’, ’2520’, ’2501’, ’6821’, ’8161’, ’2650’, ’6211’, ’8370’, ’4687’, ’2850’, ’9254’, ’8200’, ’2441’, ’3400’, ’2221’, ’8280’, ’2278’, ’5455’, ’9041’, ’3210’, ’4694’, ’2120’, ’2210’, ’8080’, ’2270’, ’2215’, ’2214’, ’2491’, ’4640’, ’2579’, ’2235’, ’8180’, ’9800’, ’2410’, ’2020’, ’2393’, ’2750’, ’6210’, ’2514’, ’9921’, ’2749’, ’2493’, ’2691’

[6] ’2590’, ’6834’, ’2750’, ’2235’, ’2745.1’, ’4611’, ’8465’, ’8400’, ’4659’, ’8320’, ’2208’, ’2560’, ’2221’, ’2749’, ’2260’, ’2240’, ’2501’, ’2393’, ’8041’, ’2441’, ’2810’, ’8370’, ’9254’, ’9250’, ’3400’, ’8330’, ’8280’, ’3170’, ’9210’, ’3053’, ’8120’, ’8080’, ’2270’, ’2215’, ’2579’, ’3500’, ’2493’, ’2214’, ’3030’, ’2278’, ’3015’, ’2752’, ’6212’, ’6211’, ’6210’, ’9921’, ’2691’, ’2250’

[7] ’2590’, ’2205’, ’8420’, ’3030’, ’2650’, ’2745.1’, ’2890’, ’2190’, ’3530’, ’2870’, ’8300’, ’8400’, ’2730’, ’2800’, ’2221’, ’4659’, ’8320’, ’4617’, ’2200’, ’3210’, ’2520’, ’8116’, ’3550’, ’2560’, ’3100’, ’2220’, ’2579’, ’2383’, ’3181’, ’9405’, ’2749’, ’8090’, ’8120’, ’2501’, ’8041’, ’8161’, ’2441’, ’2120’, ’6838’, ’2850’, ’9250’, ’8330’, ’8200’, ’3170’, ’9041’, ’3053’, ’4694’, ’4687’, ’8080’, ’2270’, ’2214’, ’2491’, ’2272’, ’2235’, ’2278’, ’9800’, ’3500’, ’2752’, ’6212’, ’6211’, ’6210’, ’9921’, ’2250’

[8] Biosemi software: http://www.biosemi.com/ (Online; accessed 10-February-2014)

[9] Viewsonic G90fB: http://www1.viewsonic.com/products/archive/g90fb.h (Online; accessed 10-February-2014)

[10] EDFBrowser: http://www.teuniz.net/edfbrowser/ (Online; accessed 10-February-2014)

[11] gHRV: http://milegroup.github.io/ghrv/ (Online; accessed 29-March-2014)

2.6. Participants

Upon arrival, the subjects were told about the procedure of the experimental design, as explained in Section 2.2. Next, they were placed in a chair in a soundproof, dimly lit room, and the biometric sensor was attached to the subject.

The subjects were three physically and mentally healthy men who were (at the time of experiment) not medicated. The average age of the three men was 44 years with a standard deviation of +/- 13.9 years.

3. RESULTS

For results, please refer to figures 1 and 2 and their captions.

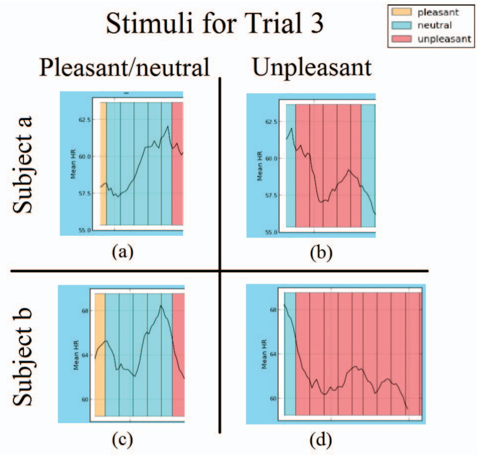


Fig. 1. The plot illustrates the mean heart rate divided into categories of valence and subject. Pleasant and neutral stimuli possibly displays a heart rate acceleration trend, whereas unpleasant stimuli results in decelerations.

4. DISCUSSION

Although no clear trend was observed in the LF/HF ratio of the HRV power spectrum, possible trends in the Mean HR were seen in longer sequences of unpleasant and pleasant/neutral stimuli, respectively, for 2 of 3 participants. When the subjects were shown a series of unpleasant pictures, as seen in Figure 1, a heart rate deceleration occurred, whereas an acceleration was found during pleasant/neutral stimuli. The somewhat unexpected heart rate acceleration may simply be a result of returning to a certain ”base level” after previous stimuli of e.g. unpleasant stimuli. In cell *d* it is furthermore seen, that the heart rate continues to de-accelerate over a longer period of time until it stabilises. This indicates that the signal seemingly can retain its properties throughout an extended period of time. The 3rd participant did not respond in any pattern that could be related to the stimuli.

The HR deceleration as a response to unpleasant stimuli is in line with previous studies [3] [4], as described in Section 1.

The first part of the trials, in which pictures were shown in random order in regards to the valence categories hereof, did not suggest any significant trends for any of the subjects. A high amount of variation of the mean heart rate was found throughout the entire trial, even over the 36 s period used to even out the inherent noise when calculating the mean HR: Standard Deviations in the raw data were seldomly below 3 – 4 bpm, which is larger than the earlier documented Mean HR variations observed as a response to presented stimuli.

This might indicate that the reaction seen in the mean HR is better suited to observe slower reactions and changes of over emotional/cognitive state, on the order of 10 s of seconds, consistent with the fact that longer sequences of equally rated stimuli provide better observable trends.

A trend was seemingly observable for 2 out of 3 test persons presented to longer sequences of similarly rated unpleasant vs. neutral/unpleasant pictures. This might indicate that mean HR changes are not just a simple and predictable function of the presented stimuli, but rather should be viewed as a representation of top-down emotional and cognitive states in response to a history of previous and current stimuli, when analysed in single-trial single-subject cases.

This is consistent with one person seemingly not showing any stimuli related changes, possibly here reflecting a personal preference in regards to the content being displayed.

Some of these responses could be derived with higher temporal resolution using brain imaging, based on Event Related Potential (ERP) responses in Electroencephalography (EEG) recordings. However, whereas an ERP represents a fast response, typically on the timescale of 100 ms for autonomous responses, 300 ms for attention priority related responses and up to 500 ms for contextual dependent responses, a HR derived signal varies on a much longer timescale, of up to 10s of seconds. It may therefore be better suited to assess cognitive state and responses that depends on the overall health, alertness, stress and ability to cope with complex dynamically changing situations [2] [1].

We therefore suggest that Mean HR variations and LF/HF HRV ratios are to be further analysed as a potential window into top-down complex and context dependent cognitive and emotional states.

One could argue that the findings in this study might be gender specific since only male subjects into account. Michela Sarlo et al. [12] previously investigated the gender differences and found that the patterns of HR and skin conductance reactivity across all emotional categories were sim-

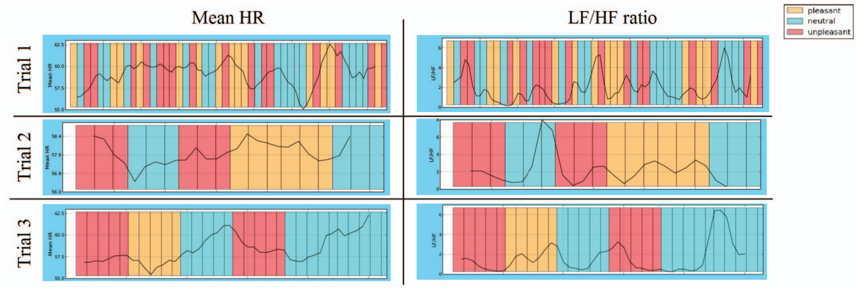


Fig. 2. This plot illustrates the three continuous trials for subject a) also seen in Figure 1. The first trial displays pictures in random order in regards to the valence category. The second and third trials shows sequences of pictures of 3 respectively 5 pictures with equal valence category. Note that the time scale is not aligned, as different amounts of pictures are shown in each trial: 48 for trial 1, 18 for trial 2 and 30 for trial 3. The distance between each vertical line is approximately 18 seconds. As seen, a high amount of variation during trial 1 occurs, which indicates that it can be difficult to deduct a pattern from the signal. In comparison, a more clear indication of a pattern can be derived from trial 2 and 3 on the basis of the mean HR, where the noise of the signal is also seemingly reduced. No clear signal is however found when looking at the LF/HF ratio.

ilar for both female and male subjects. Blood pressure on the other hand for stimuli related to sexual content showed an increased reactivity for men in comparison to women. Taken together, these findings indicate that the patterns of emotional reaction as seen in HR might be independent of gender.

We also suggest that other types of stimuli should be analyzed. The pictures, in the way presented here, do not contain any context nor a storyline, as would e.g. a movie trailer or a film clip. We suspect that stimuli designed to engage and tell a story might have a stronger affect on the subjects and might induce a stronger measurable response. Setting up experiments that are not simply based on passing viewing/listening but contains elements of risk, punishment and rewards, like [4], might even further increase the subjects' engagement and lead to increased measurable responses.

Other studies, such as [3], have found small HR changes on a timescale of few seconds. Given the observed noise and seemingly random variations in the HR signal, this proved difficult to reproduce in a single-trial single-subject study like the present one.

Seen from a commercial point of view, these findings might prove to be highly valuable within industries which provide a context over a longer period of time where emotional state is of the essence. Examples hereof are the gaming-, movie- and music industries, which might thus be able to adapt content to the emotional responses of users.

Consumers have only recently gained access to low cost wearable computing devices that are normally equipped with multiple sensors to measure different biometric makers. The LF/HF ratio of HRV is however not normally accounted for, which may be a potential limitation. Even though this study found no direct link between the cognitive state and the LF/HF ratio of HRV, [8] previously found a subduction in the HF HRV component in which the biggest response was emotionally pleasant movies. This indicates, that HRV, like the mean heart rate, could be of potential interest for future studies as well as for commercial purposes.

When utilising biometric sensors, the analysis of the data is often instantaneous or given as simple mean values. As this study indicates, the emotional function is not simply a function of the presented stimuli, but rather a top-down representation in response to a history of the presented stimuli and context, also when analysing data from single subjects. Thus, more suitable solutions, such as online storage and better analysis tools, might be required to better utilize biometric sensors to their full potential.

In conclusion, we support the point of view that simple biometric markers has the potential to serve as a window into a more complex emotional and cognitive state, but more research will be required and low cost biometric sensors and related tools needs improvements in order to better utilize these commercially.

5. REFERENCES

[1] Marek Malik, J Thomas Bigger, A John Camm, Robert E Kleiger, Alberto Malliani, Arthur J Moss, and Peter J Schwartz, "Heart rate variability standards of measurement, physiological interpretation, and clinical use," *European heart journal*, vol. 17, no. 3, pp. 354–381, 1996.

[2] Julian F Thayer, Fredrik Åhs, Mats Fredrikson, John J Sollers III, and Tor D Wager, "A meta-analysis of

heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health," *Neuroscience & Biobehavioral Reviews*, vol. 36, no. 2, pp. 747–756, 2012.

[3] Margaret M Bradley, Laura Miccoli, Miguel A Escrig, and Peter J Lang, "The pupil as a measure of emotional arousal and autonomic activation," *Psychophysiology*, vol. 45, no. 4, pp. 602–607, 2008.

[4] Peter J Lang and Margaret M Bradley, "Emotion and the motivational brain," *Biological psychology*, vol. 84, no. 3, pp. 437–450, 2010.

[5] Andreas Löw, Peter J Lang, J Carson Smith, and Margaret M Bradley, "Both predator and prey emotional arousal in threat and reward," *Psychological Science*, vol. 19, no. 9, pp. 865–873, 2008.

[6] Margaret M Bradley, Steven Hamby, Andreas Löw, and Peter J Lang, "Brain potentials in perception: picture complexity and emotional arousal," *Psychophysiology*, vol. 44, no. 3, pp. 364–373, 2007.

[7] Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, et al., "Motivated attention: Affect, activation, and action," *Attention and orienting: Sensory and motivational processes*, pp. 97–135, 1997.

[8] Richard D Lane, Kateri McRae, Eric M Reiman, Kewei Chen, Geoffrey L Ahern, and Julian F Thayer, "Neural correlates of heart rate variability during emotion," *Neuroimage*, vol. 44, no. 1, pp. 213–222, 2009.

[9] Mikkel Wallentin, Andreas Højlund Nielsen, Peter Vuust, Anders Dohn, Andreas Roepstorff, and Torben Ellegaard Lund, "Amygdala and heart rate variability responses from listening to emotionally intense parts of a story," *Neuroimage*, vol. 58, no. 3, pp. 963–973, 2011.

[10] Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert, "International affective picture system (iaps): Technical manual and affective ratings," Gainesville, FL: The Center for Research in Psychophysiology, University of Florida, 1999.

[11] Nicholas D Giardino, Paul M Lehrer, and Robert Edelberg, "Comparison of finger plethysmograph to ecg in the measurement of heart rate variability," *Psychophysiology*, vol. 39, no. 2, pp. 246–253, 2002.

[12] Michela Sarlo, Daniela Palomba, Giulia Buodo, Rita Minghetti, and Luciano Stegagno, "Blood pressure changes highlight gender differences in emotional reactivity to arousing pictures," vol. 70, pp. 188–196. Elsevier, 2005.

Paper I

Key Terms in Eye Movement Research

The following is a working paper (by the author) that summarize some of the key terminologies in eye movement research, prepared as input to Project Principia and also intended as a brief introduction for students that work with eye tracking.

In the present thesis, similar topics are explained in chapters 1 and 2, and the present paper has been used as the (initial) basis for some of the explanations.

Key Terms in Eye Movement Research

Per Baekgaard pgba@dtu.dk

Nov 2015

Eye Movements

Eye movements to a large degree reflects attention.

The Visual System

The visual process starts at the eye where light is converted to electrical signals passed on to the visual cortex at the rear of the brain.

The Eye

At the center of the field of view is an area of high resolution called the fovea; it occupies less than 2 degrees [Holmqvist 2011] of visual angle – roughly the size of your thumb when held at arms distance [O'Shea 1991].

Where you look, i.e. the center of what is projected inside the fovea, is called the *gaze point*. When looking at a screen, the pixel coordinates of the gaze point are often referred to as the *gaze coordinates*.

Classification of Eye Movements

Fixations and Saccades, Microsaccades and Glissades

The eyes tend to remain within the same small area only for a short period of time, typically 200-300 milliseconds but occasionally as short as some tens of milliseconds or as long as a few seconds [Holmqvist 2011]. These periods are referred to as *fixations*, although they are not completely absent of eye movements; very small movements take place all the time, often called *microsaccades*. Microsaccades are also an important part of the visual processing [Rucci 2015].

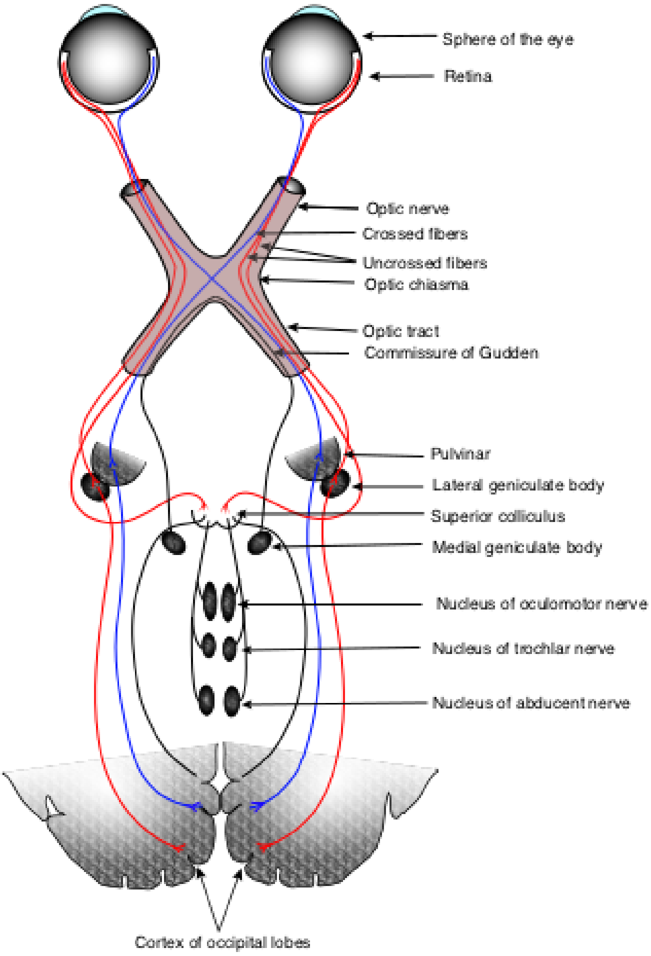


Figure 1: The Visual Processing System

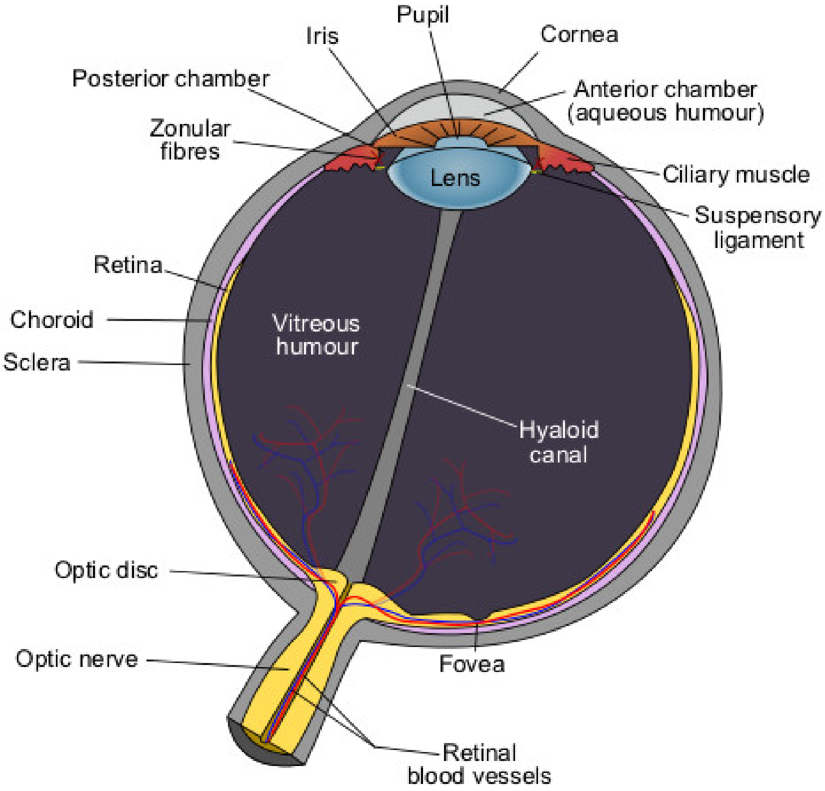


Figure 2: The Eye

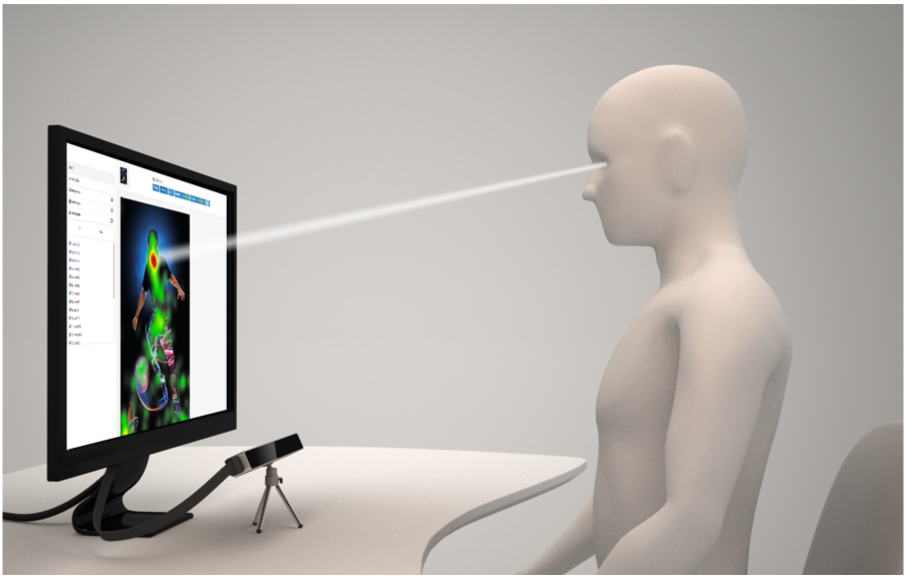


Figure 3: Gaze Coordinates *NOTE: Image not under CC and needs to be replaced*

Fixations are the element of eye movements that are conventionally most directly linked to *attention* to a specific visual area. [1].

Fixations can be characterised by for instance their duration, frequency, coordinates and spread.

Between fixations, the eyes move rapidly in a near-ballistic mode; these are the fastest movement the body is capable of [Holmqvist 2011] and last only for 30-80 milliseconds. Such eye movements are called *saccades*, during which the visual system appear blinded; no conscious visual processes takes place during saccades [Holmqvist 2011], [Burr 1994]. Sometimes a saccade does not hit its target as intended; in this case corrective *glissades* [Bahill 1975a] will take place at the end of the saccade to place the following fixation at the desired gaze point.

Saccades can be quantified by for instance their speed (visual angle/time), duration, length, frequency, latency vs. visual stimuli or accuracy [Holmqvist 2011] as well as their direction.

Saccades are affected by task difficulties, workload, whether one is trying to get an overview or is carefully inspecting, high-frequency content of the scene, reading abilities, and many other factors [Holmqvist 2011].

[1] Sometimes, two different "modes" are suggested: 1) Cognitive controlled *top-down* or *goal-oriented* processing where e.g. specific objects are being searched for, and 2) *saliency* or *stimulus driven* processing, where the attention of the eye is (often unconsciously) directed toward changes in the present visual scene [Corbetta 2002]

Smooth Pursuit

The eyes can also move in another mode during which conscious visual processing takes place. This can *only* happen when the eyes follow an object in relative motion to the head, such as when tracking a car passing by [2]. These smooth movements are called *smooth pursuit*.

Blinks

Blinks also need to be considered when inspecting eye movements or the visual processing system. Blinks happen spontaneously without a conscious effort: The majority of *Endogenous Eyeblinks* [Stern 1984] last between 100 and 300 milliseconds [Holmqvist 2011], with visual intake being reduced during as well as immediately before and after each blink. Also *reflex blinks* and *voluntary blinks* may take place [Stern 1984].

Blinks are often characterised by their frequency [3] or by either their total duration or the duration of specific phases (closing, closed, opening).

Blinks are modulated by drowsiness, loss of vigilance or mental workload [Holmqvist 2011], [Caffier 2003] and can often be linked to the completion of specific tasks. During (visually or attentionally) demanding tasks, blinks are often inhibited, and then blinks and saccadic eye movements sometimes takes place “in the same gap”, apparently minimising the time the visual intake is inhibited [Coltheart 2010, p. 129: “Thinking Affects Blinking”].

Pupil Reactions

The pupil primarily controls the amount of light that hits the photosensitive areas of the eyes; the higher the luminosity, the smaller the pupil becomes [Bouma 1962].

It is, however, also impacted by processing [Hyona 1995] or cognitive workload [Pomplun 2003] as well as emotional arousal [Bradley 2008]. For task specific workloads, the *pupil dilation* – small and short dilations of the pupil size that co-occur with specific conditions – can be used as a marker of this processing. Drowsiness or fatigue may conversely decrease the pupil diameter [Holmqvist 2011]. Further, recent studies [Laeng 2012] hint at *pupilometry* being a “window” into the preconscious.

[2] When keeping an eye on an object while turning your head another effect also come into play: the vestibulo-ocular reflex

[3] Some studies have reported average blink-rates of 7 to 11 pr. minute when reading, and 13 increasing to 40 pr. minute during extended driving tasks [Stern 1984]

Use of Eye Tracking

Eye Tracking (the process of capturing Eye Movements) can be used for 1) *passive*, sometimes “after-the-fact” analysis, for 2) *active* control or as part of 3) an *adaptive* system that observes the user and adjusts its behaviour to the inferred state of user.

Passive

In passive mode, *fixation density maps* or *heatmaps* are often used as a tool, although they simplify the information by removing parts of the temporal domain.

Specific parts of an observed scene that receive fixations are called *areas of interest*, and heatmaps may show their received intention as a proportion of time one or multiple observes has fixated at or near the area. [Engelke 2013]

It is, however, no simple task to identify fixations. The most efficient methods needs high resolution (>200 Hz) gaze points in order to correctly identify saccades [Nystrom 2010]. For lower resolution data, simpler clustering algorithms, such as a modified version of DBSCAN that breaks fixations after 70-90 milliseconds of gaze points outside the current cluster, can be useful.

Passive eye tracking can be seen in use in e.g. marketing research or when comparing different designs of layouts of packaging, web pages, grocery stores, etc.

Active

In active mode, eye movements are used to directly control a device. This could be used in games for controlling direction of movement or intent, or for hands free typing, or in other applications.

As eye movements are normally just an implicit part of the visual and cognitive processing, active eye tracking paradigms needs to be designed carefully in order achieve ease of use and efficiency while avoiding eye fatigue and increased cognitive workload/mental stress. See [Hansen 2012] or [MacKenzie 2012] (also available online) for examples of how to evaluate such systems.

Adaptive

Adaptive eye tracking can be viewed as a hybrid between the two: a system may analyse passive eye movement to infer the users state, and adapts its present behaviour to this inferred state.

Examples could be to adjust the current level of difficulty to the inferred cognitive or emotional load of the user, or simply to suggest breaks to users that appear with increased fatigue in front of their screens.

How Eye Trackers Work

Modern Eye Trackers usually work by means of a set of infrared (IR) LEDs and an IR camera that captures the glints of the LEDs in the eye as well as the the pupil[4] itself.

By keeping the IR LED's steady vs the object being looked at (i.e. the screen), the position of the pupil relative to the glints can be projected into gaze coordinates by means of a calibration to known targets.

Demo/experiment

The Eye Tribe Tracker

Eye Tracking Demo

Links/references

[Holmqvist 2011]: Holmqvist, Kenneth, et al (2011). Eye tracking: A comprehensive guide to methods and measures. Oxford University Press.

[O'Shea 1991]: O'Shea, R. P. (1991). Thumb's rule tested: visual angle of thumb's width is about 2 deg. Perception, 20(3), 415–418. doi:10.1068/p200415

[Burr 1994]: Burr, D. C., Morrone, M. C., & Ross, J. (1994). Selective suppression of the magnocellular visual pathway during saccadic eye movements. Nature, 371(6497), 511–513. doi:10.1038/371511a0

[Bahill 1975a]: Bahill, A. T., Clark, M. R., & Stark, L. (1975). Glissades—eye movements generated by mismatched components of the saccadic motoneuronal control signal. Mathematical Biosciences, 26(3-4), 303–318. doi:10.1016/0025-5564(75)90018-8

[Stern 1984]: Stern, J. A., Walrath, L. C., & Goldstein, R. (1984). The endogenous eyeblink. Psychophysiology, 21(1), 22–33. doi:10.1111/j.1469-8986.1984.tb02312.x

[Bouma 1962]: Bouma, H. (1962). Size of the static pupil as a function of wavelength and luminosity of the light incident on the human eye. Nature, 193, 690–691. doi:10.1038/193690a0

[Hyona 1995]: Hyönä, J., Tommola, J., & Alaja, A.-M. (1995). Pupil Dilation as a Measure of Processing Load in Simultaneous Interpretation and Other Language Tasks. The Quarterly Journal of Experimental Psychology Section A, 48(3), 598–612. doi:10.1080/14640749508401407

[4] which will appear very bright in IR light

[Pomplun 2003]: Pomplun, M., & Sunkara, S. (2003). Pupil Dilation as an Indicator of Cognitive Workload in Human-Computer Interaction. Human-Centered Computing: Cognitive, Social and Ergonomic Aspects, 542–546.

[Bradley 2008]: Bradley, M. M., Miccoli, L., Escrig, M. a., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. Psychophysiology, 45(4), 602–607. doi:10.1111/j.1469-8986.2008.00654.x

[Caffier 2003]: Caffier, P. P., Erdmann, U., & Ullsperger, P. (2003). Experimental evaluation of eye-blink parameters as a drowsiness measure. European Journal of Applied Physiology, 89(3-4), 319–325. doi:10.1007/s00421-003-0807-5

[Laeng 2012]: Laeng, B., Sirois, S., & Gredeback, G. (2012). Pupillometry: A Window to the Preconscious? Perspectives on Psychological Science, 7(1), 18–27. doi:10.1177/1745691611427305

[Corbetta 2002]: Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. Nature Reviews. Neuroscience, 3(3), 201–15. doi:10.1038/nrn755

[Nystrom 2010]: Nyström, M., & Holmqvist, K. (2010). An adaptive algorithm for fixation, saccade, and glissade detection in eye tracking data. Behavior Research Methods, 42(1), 188–204. doi:10.3758/BRM.42.1.188

[Rucci 2015]: Rucci, M., & Victor, J. D. (2015). The unsteady eye: an information-processing stage, not a bug. Trends in Neurosciences, 38(4), 195–206. doi:10.1016/j.tins.2015.01.005

[Coltheart 2010]: Coltheart, V. (Ed.). (2011). Tutorials in visual cognition. Routledge.

[Engelke 2013]: Engelke, U., Liu, H., & Wang, J. (2013). Comparative study of fixation density maps. Image Processing, . . . , 22(3), 1121–33. doi:10.1109/TIP.2012.2227767

[MacKenzie 2012]: MacKenzie, I. S. (2012). Evaluating eye tracking systems for computer input. In Majaranta, P., Aoki, H., Donegan, M., Hansen, D. W., Hansen, J. P., Hyrskykari, A., & Räihä, K.-J. (Eds.). Gaze interaction and applications of eye tracking: Advances in assistive technologies, pp. 205-225. Hershey, PA: IGI Global.

[Hansen 2011]: Hansen, J. P., & Aoki, H. (2011). Methods and Measures: An Introduction. Gaze Interaction and Applications of Eye Tracking: Advances in Assistive Technologies: Advances in Assistive Technologies, 197.

Bibliography

[1] Bækgaard, P., Petersen, M. K., and Larsen, J. E. “In the twinkling of an eye: Synchronization of EEG and eye tracking based on blink signatures”. In: *4th International Workshop on Cognitive Information Processing - Proceedings of CIP 2014* (2014), pp. 14–19.

[2] Bækgaard, P., Petersen, M. K., and Larsen, J. E. “Thinking Outside of the Box or Enjoying Your 2 Seconds of Frame?” In: *Universal Access in Human-Computer Interaction. Access to the Human Environment and Culture* 9178 (2015), pp. 186–195.

[3] Bækgaard, P., Petersen, M. K., and Larsen, J. E. “Assessing Levels of Attention Using Low Cost Eye Tracking”. In: *Universal Access in Human-Computer Interaction. Methods, Techniques, and Best Practices: 10th International Conference, UAHCI 2016, Held as Part of HCI International 2016, Toronto, ON, Canada, July 17-22, 2016, Proceedings, Part I*. Ed. by Antona, M. and Stephanidis, C. Cham: Springer International Publishing, 2016, pp. 409–420.

[4] Bækgaard, P., Petersen, M. K., and Larsen, J. E. “Separating Components of Attention and Surprise”. In: *Manuscript submitted for publication* (2016).

[5] Bækgaard, P., Jalaliniya, S., and Hansen, J. P. “Pupillary Measures During an Assembly Task”. In: *Manuscript submitted for publication* (2016).

[6] Cuttone, A., Bækgaard, P., Sekara, V., Jonsson, H., Larsen, J. E., and Sune Lehmann. “SensibleSleep: A Bayesian Model for Learning Sleep Patterns from Smartphone Events”. In: *Manuscript submitted for publication* (2016).

[7] Bækgaard, P., Petersen, M. K., and Larsen, J. E. “The Blank Stare”. In: *Manuscript in preparation* (2016).

[8] Laundav, D. K., Jensen, C. B. F., Bækgaard, P., Petersen, M. K., and Larsen, J. E. “Your heart might give away your emotions”. In: *Proceedings of the 2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. 2014, pp. 1–6.

[9] Bækgaard, P. “Key Terms in Eye Movement Research (Project Principia; Working Paper)”. 2015.

[10] Bækgaard, P. “Simple python interface to the Eye Tribe eye tracker [baekgaard/peyetribe]”. In: *GitHub repository* (2014).

[11] Bækgaard, P. “Attention Network Test implemented in PsychoPy [baekgaard/ant]”. In: *GitHub repository* (2016).

[12] Keith, R. J. “The Marketing Revolution”. In: *Journal of Marketing* 24.3 (1960), pp. 35–38.

[13] Melton, T. “The Benefits of Lean Manufacturing”. In: *Chemical Engineering Research and Design* 83.6 (2005), pp. 662–673.

[14] Hassenzahl, M. and Tractinsky, N. “User experience - a research agenda”. In: *Behaviour & Information Technology* 25.2 (2006), pp. 91–97.

[15] Forlizzi, J. “Understanding Experience in Interactive Systems”. In: *DIS ’04 Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques*. 2004, pp. 261–268.

[16] Vuori, T. O. and Huy, Q. N. “Distributed Attention and Shared Emotions in the Innovation Process: How Nokia Lost the Smartphone Battle”. In: *Administrative Science Quarterly* (2015), pp. 1–43.

[17] Sullivan, M. *What I Learned Working With Jony Ive’s Team On The Apple Watch (Online)*. 2016.

[18] Eklund, A., Nichols, T. E., and Knutsson, H. “Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates”. In: *Proceedings of the National Academy of Sciences* 113.33 (2016), p. 201602413.

[19] Swan, M. “The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery”. In: *Big Data* 1.2 (2013), pp. 85–99.

[20] Petersen, M. K., Stahlhut, C., Stopczynski, A., Larsen, J. E. E., and Hansen, L. K. “Smartphones get emotional: Mind reading images and reconstructing the neural sources”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 6975 LNCS (2011), pp. 578–587.

[21] Stopczynski, A., Stahlhut, C., Larsen, J. E., Petersen, M. K., and Hansen, L. K. “The Smartphone Brain Scanner: A Mobile Real-time Neuroimaging System”. In: (2013), pp. 1–17.

[22] Stopczynski, A., Stahlhut, C., Petersen, M. K., Larsen, J. E., Jensen, C. F., Ivanova, M. G., Andersen, T. S., and Hansen, L. K. “Smartphones as pocketable labs: Visions for mobile brain imaging and neurofeedback”. In: *International Journal of Psychophysiology* 91.1 (2014), pp. 54–66.

[23] De Vos, M., Gandras, K., and Debener, S. “Towards a truly mobile auditory brain–computer interface: Exploring the P300 to take away”. In: *International Journal of Psychophysiology* 91.1 (2014), pp. 46–53.

[24] De Vos, M. and Debener, S. “Mobile EEG: towards brain activity monitoring during natural action and cognition”. In: *International Journal of Psychophysiology* 91.1 (2014), pp. 1–2.

[25] Ferhat, O., Vilariño, F., and Sánchez, F. J. “A Cheap Portable Eye–tracker Solution for Common Setups”. In: *Journal of Eye Movement Research* 7.3 (2014), pp. 1–10.

[26] The Eye Tribe. *Web Page*.

[27] Dwoskin, E. and Rusli, E. M. *The Technology that Unmasks Your Hidden Emotions*. Jan. 2015.

[28] Affectiva. *Affectiva Web Page*. 2016.

[29] Poh, M.-Z., Swenson, N., and Picard, R. “A Wearable Sensor for Unobtrusive, Long-Term Assessment of Electrodermal Activity”. In: *IEEE Transactions on Biomedical Engineering* 57.5 (2010), pp. 1243–1252.

[30] Strauss, M., Reynolds, C., Hughes, S., Park, K., McDarby, G., and Picard, R. W. “The HandWave Bluetooth skin conductance sensor”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 3784 LNCS (2005), pp. 699–706.

[31] Empatica. *Empatica Web Page*. 2016.

[32] Swan, M. “Sensor Mania! The Internet of Things, Wearable Computing, Objective Metrics, and the Quantified Self 2.0”. In: *Journal of Sensor and Actuator Networks* 1.3 (2012), pp. 217–253.

[33] Kivikangas, J. M., Chanel, G., Cowley, B., Ekman, I., Salminen, M., Järvelä, S., and Ravaja, N. “A review of the use of psychophysiological methods in game research”. In: *Journal of Gaming & Virtual Worlds* 3.3 (2011), pp. 181–199.

[34] Mulert, C., Jäger, L., Schmitt, R., Bussfeld, P., Pogarell, O., Möller, H. J., Juckel, G., and Hegerl, U. “Integration of fMRI and simultaneous EEG: Towards a comprehensive understanding of localization and time-course of brain activity in target detection”. In: *NeuroImage* 22.1 (2004), pp. 83–94.

[35] Zhukov, L. and Weinstein, D. “Independent Component Analysis for EEG Source Localization”. In: June (2000), pp. 87–96.

[36] Koles, A. L. “Trends in EEG source localization”. In: *Electroencephalography and Clinical Neurophysiology* 106 (1998), pp. 127–137.

[37] Michel, C. M., Murray, M. M., Lantz, G., Gonzalez, S., Spinelli, L., and Grave De Peralta, R. “EEG source imaging”. In: *Clinical Neurophysiology* 115.10 (2004), pp. 2195–2222.

[38] Coull, J. T. *Neural correlates of attention and arousal: Insights from electrophysiology, functional neuroimaging and psychopharmacology*. 1998.

[39] Hillyard, S. A., Hink, R. F., Schwent, V. L., and Picton, T. W. “Electrical Signs of Selective Attention in the Human Brain”. In: *Science 182* October (1973), pp. 177–180.

[40] Squires, N. K., Squires, K. C., and Hillyard, S. A. “Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man”. In: *Electroencephalography and Clinical Neurophysiology* 38.4 (1975), pp. 387–401.

[41] Picton, T. W. “The P300 wave of the human event-related potential”. In: *Journal of Clinical Neurophysiology* 9.4 (1992), pp. 456–479.

[42] Polich, J. *Updating P300: An integrative theory of P3a and P3b*. 2007.

[43] Kutas, M. and Hillyard, S. A. *Reading senseless sentences: brain potentials reflect semantic incongruity.* 1980.

[44] Debener, S., Emkes, R., De Vos, M., and Bleichner, M. “Unobtrusive ambulatory EEG using a smartphone and flexible printed electrodes around the ear.” In: *Scientific reports* 5 (2015), p. 16743.

[45] Boucsein, W. *Electrodermal activity*. Vol. 3. 2012, p. 618.

[46] Johnson, L. C. and Lubin, A. “Spontaneous electrodermal activity during waking and sleeping.” In: *Psychophysiology* 3.1 (1966), pp. 8–17.

[47] Lykken, D. T. and Venables, P. H. “Direct measurement of skin conductance: A proposal for standardization”. In: *Psychophysiology* 8.5 (1971), pp. 656–672.

[48] Critchley, H. D., Elliott, R., Mathias, C. J., and Dolan, R. J. “Neural activity relating to generation and representation of galvanic skin conductance responses: a functional magnetic resonance imaging study.” In: *The Journal of neuroscience : the official journal of the Society for Neuroscience* 20.8 (2000), pp. 3033–3040.

[49] Healey, J. A. and Picard, R. W. “Detecting Stress During Real-World Dring Tasks Using Physiological Sensors”. In: *IEEE Transactions on Intelligent Transportation Systems* 6.2 (2005), pp. 156–166.

[50] Sano, A. and Picard, R. W. “Stress Recognition Using Wearable Sensors and Mobile Phones”. In: *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on* (2013), pp. 671–676.

[51] Holmqvist, K. *Eye Tracking: a comprehensive guide to methods and measures*. Oxford University Press, 2011.

[52] Picard, R., Fedor, S., and Ayzenberg, Y. "Multiple Arousal Theory and Daily-Life Electrodermal Activity Asymmetry". In: *Emotion Review* 8.1 (2016), pp. 62–75.

[53] Picard, R., Fedor, S., and Ayzenberg, Y. "Response to Commentaries on "Multiple Arousal Theory and Daily-Life Electrodermal Activity Asymmetry"". In: *Emotion Review* 8.1 (2016), pp. 84–86.

[54] Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. "Heart rate variability. Standards of measurement, physiological interpretation, and clinical use. T". In: *European Heart Journal* 17.3 (1996), pp. 354–81.

[55] Wallentin, M., Nielsen, A. H., Vuust, P., Dohn, A., Roepstorff, A., and Lund, T. E. "Amygdala and heart rate variability responses from listening to emotionally intense parts of a story." In: *NeuroImage* 58.3 (2011), pp. 963–73.

[56] Taelman, J., Vandeput, S., Spaepen, A., and Huffel, S. V. "Influence of Mental Stress on Heart Rate and Heart Rate Variability". In: *Ecifmbe 2008* 29.1 (2009), pp. 1366–1369.

[57] Löw, A., Lang, P. J., Smith, J. C., and Bradley, M. M. "Both predator and prey: Emotional arousal in threat and reward". In: *Psychological Science* 19.9 (2008), pp. 865–873.

[58] Lane, R. D., McRae, K., Reiman, E. M., Chen, K., Ahern, G. L., and Thayer, J. F. "Neural correlates of heart rate variability during emotion". In: *NeuroImage* 44.1 (2009), pp. 213–222.

[59] Thayer, J. F., Ahs, F., Fredrikson, M., Sollers III, J. J., and Wager, T. D. "A meta-analysis of heart rate variability and neuroimaging studies: Implications for heart rate variability as a marker of stress and health." In: *Neuroscience and Biobehavioral Reviews* 36.2 (2012), pp. 747–756.

[60] Kemper, K. J., Hamilton, C., and Atkinson, M. "Heart rate variability: Impact of differences in outlier identification and management strategies on common measures in three clinical populations". In: *Pediatric Research* 62.3 (2007), pp. 337–342.

[61] Vila, J., Palacios, F., Fernandez-Delgado, M., Felix, P., and Barro, S. "Time-Frequency Analysis of Heart-Rate Variability". In: *Engineering in Medicine and Biology Magazine* 16.5 (1997), pp. 119–126.

[62] Rodríguez-Liñares, L., Lado, M. J., Vila, X. a., Méndez, a. J., and Cuesta, P. "GHRV: Heart rate variability analysis made easy". In: *Computer Methods and Programs in Biomedicine* 116.1 (2014), pp. 26–38.

[63] Pradhan, L. and Islam, M. "Replacing outliers with existing data in Inter Beat Interval Signal for Heart Rate Variability analysis". In: *IBIS* 1 (2010), p. 10.

[64] Porges, S. W. "The polyvagal perspective". In: *Biological Psychology* 74.2 (2007), pp. 116–143.

[65] Giardino, N. D., Lehrer, P. M., and Edelberg, R. "Comparison of finger plethysmograph to ECG in the measurement of heart rate variability." In: *Psychophysiology* 39.2 (2002), pp. 246–253.

[66] Mitkidis, P., McGraw, J. J., Roepstorff, A., and Wallot, S. "Building trust: Heart rate synchrony and arousal during joint action increased by public goods game". In: *Physiology & Behavior* 149 (2015), pp. 101–106.

[67] McDuff, D., Kaliouby, R. el, Senechal, T., Amr, M., Cohn, J. F., and Picard, R. "Affectiva-MIT Facial Expression Dataset (AM-FED): Naturalistic and Spontaneous Facial Expressions Collected &#x0022;In-the-Wild&#x0022;" in: *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2013), pp. 881–888.

[68] Ekman, P. and Friesen, W. V. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. 1978.

[69] Bunk, E. "The Influence of Mental State on Eye Movements". M.Sc. DTU, 2016.

[70] Komogortsev, O. V. and Khan, J. I. "Eye Movement Prediction by Kalman Filter with Integrated Linear Horizontal Oculomotor Plant Mechanical Model". In: *Eye Tracking Research & Applications (ETRA) Symposium* (2008), pp. 229–236.

[71] Robinson, D. A. "The Mechanics of Human Saccadic Eye Movement." In: *The Journal of physiology* 174 (1964), pp. 245–264.

[72] Robinson, D. A. "The mechanics of human smooth pursuit eye movement." In: *The Journal of Physiology* 180.3 (1965), pp. 569–591.

[73] Robinson, D. A. "The Oculomotor Control System: A Review". In: *Proceedings of the IEEE* 56.6 (1968), pp. 1032–1049.

[74] Robinson, D. A., O'meara, D. M., Scott, A. B., and Collins, C. C. "Mechanical components of human eye movements". In: *Journal of Applied Physiology* 26.5 (1969), pp. 548–553.

[75] Robinson, D. A. “Models of the saccadic eye movement control system.” In: *Kybernetik* 14.2 (1973), pp. 71–83.

[76] Young, L. R. and Stark, L. “Variable Feedback Experiments Testing a Sampled Data Model for Eye Tracking Movements”. In: *IEEE Transactions on Human Factors in Electronics* HFE-4.1 (1963), pp. 38–51.

[77] Sparks, D. L. “The brainstem control of saccadic eye movements”. In: *Nature reviews. Neuroscience* 3.12 (2002), pp. 952–964.

[78] Leigh, R. J. and Zee, D. S. *The Neurology of Eye Movements*. Oxford University Press, USA, 2015.

[79] Fan, J., McCandliss, B. D., Sommer, T., Raz, A., and Posner, M. I. “Testing the Efficiency and Independence of Attentional Networks”. In: *Journal of Cognitive Neuroscience* 14.3 (2002), pp. 340–347.

[80] Peirce, J. W. “PsychoPy-Psychophysics software in Python”. In: *Journal of Neuroscience Methods* 162.1-2 (2007), pp. 8–13.

[81] Dodge, R. and Cline, T. S. “The angle velocity of eye movements.” In: *Psychological Review* 8 (1901), pp. 145–157.

[82] Delabarre, E. B. “A Method of Recording Eye-Movements”. In: *The American Journal of Psychology* 9.4 (1898), pp. 572–574.

[83] Duchowski, A. *Eye Tracking Methodology Theory and Practice*. Springer, 2007.

[84] Horsley, M., Eliot, M., Knight, B. A., and Reilly, R. *Current trends in eye tracking research*. 2014, pp. 1–345.

[85] Dalmaijer, E. S. “Is the low-cost EyeTribe eye tracker any good for research”. In: *PeerJ PrePints* (2014), pp. 1–35.

[86] Ooms, K., Lapon, L., Dupont, L., and Popelka, S. “Accuracy and precision of fixation locations recorded with the low-cost Eye Tribe tracker in different experimental set-ups”. In: *Journal of Eye Movement Research* 8.1 (2015), pp. 1–24.

[87] Cornsweet, T. N. and Crane, H. D. “Accurate two-dimensional eye tracker using first and fourth Purkinje images.” In: *Journal of the Optical Society of America* 63.8 (1973), pp. 921–928.

[88] Young, L. R. and Sheena, D. “Survey of eye movement recording methods”. In: *Behavior Research Methods & Instrumentation* 7.5 (1975), pp. 397–429.

[89] Guestrin, E. D. and Eizenman, M. “General theory of remote gaze estimation using the pupil center and corneal reflections”. In: *IEEE Transactions on Biomedical Engineering* 53.6 (2006), pp. 1124–1133.

[90] Morimoto, C. H. and Mimica, M. R. M. *Eye gaze tracking techniques for interactive applications*. 2005.

[91] Tobii. *Tobii Web Page*. 2016.

[92] Guenter, B., Finch, M., Drucker, S., Tan, D., and Snyder, J. “Foveated 3D graphics”. In: *ACM Transactions on Graphics* 31.6 (2012), p. 1.

[93] SMI. *Web Page*. 2016.

[94] *SMI Foveated Rendering*. 2016.

[95] Duchowski, A. T. and Çöltekin, A. “Foveated gaze-contingent displays for peripheral LOD management, 3D visualization, and stereo imaging”. In: *ACM Transactions on Multimedia Computing, Communications, and Applications* 3.4 (2007), pp. 1–18.

[96] Komogortsev, O. V. and Khan, J. I. “Eye movement prediction by oculomotor plant Kalman filter with brainstem control”. In: *Journal of Control Theory and Applications* 7.1 (2009), pp. 14–22.

[97] Komogortsev, O. V. and Jayarathna, U. K. S. “2D oculomotor plant mathematical model for eye movement simulation”. In: *8th IEEE International Conference on BioInformatics and BioEngineering, BIBE 2008*. 2008.

[98] Han, P., Saunders, D. R., Woods, R. L., and Luo, G. “Trajectory prediction of saccadic eye movements using a compressed exponential model”. In: *Journal of Vision* 13.8 (2013), p. 27.

[99] Duchowski, A. T. *A Breadth-First Survey of Eye Tracking Applications*. 2002.

[100] Karray, F., Alemzadeh, M., Saleh, J. a., and Arab, M. N. “Human-Computer Interaction: Overview on State of the Art”. In: *International Journal on Smart Sensing and Intelligent Systems* 1.1 (2008), pp. 137–159.

[101] Jacob, R. J. K. “The use of eye movements in human-computer interaction techniques: what you look at is what you get”. In: *ACM Transactions on Information Systems* 9.3 (1991), pp. 152–169.

[102] Møllenbach, E., Hansen, J. P., and Lillholm, M. “Eye Movements in Gaze Interaction”. In: *Journal of Eye Movement Research* 6 (2013), pp. 1–15.

[103] Posner, M. I., Snyder, C. R. R., and Davidson, B. J. “Attention and the detection of signals.” In: *Journal of Experimental Psychology: General* 109.2 (1980), pp. 160–174.

[104] Posner, M. I. “Orienting of attention”. In: *Quarterly Journal of Experimental Psychology* 32 (1980), pp. 3–25.

[105] Hyrskykari, A., Ovaska, S., Majaranta, P., Räihä, K., and Lehtinen, M. "Gaze Path Stimulation in Retrospective Think-Aloud". In: *Journal of Eye Movement Research* 2.4 (2008), pp. 1–18.

[106] Bol, N., Boerman, S. C., Romano Bergstrom, J. C., and Kruikemeier, S. "An Overview of How Eye Tracking Is Used in Communication Research". In: *Universal Access in Human-Computer Interaction. Methods, Techniques, and Best Practices, Volume 9737 of the series Lecture Notes in Computer Science*. Springer International Publishing, 2016, pp. 421–429.

[107] Borji, A. and Itti, L. "State-of-the-art in visual attention modeling". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.1 (2013), pp. 185–207.

[108] Velichkovsky, B. M., Dornhoefer, S. M., Pannasch, S., and Unema, P. J. "Visual fixations and level of attentional processing". In: *Proceedings of the symposium on Eye tracking research & applications - ETRA '00* (2000), pp. 79–85.

[109] Martinez-Conde, S., Macknik, S. L., and Hubel, D. H. "The role of fixational eye movements in visual perception". In: *Nature Reviews Neuroscience* 5.3 (2004), pp. 229–240.

[110] Martinez-Conde, S. and Macknik, S. L. "Fixational eye movements across vertebrates: comparative dynamics, physiology, and perception." In: *Journal of vision* 8.14 (2008), pp. 28.1–16.

[111] McCamy, M. B., Otero-Millan, J., Macknik, S. L., Yang, Y., Troncoso, X. G., Baer, S. M., Crook, S. M., and Martinez-Conde, S. "Microsaccadic Efficacy and Contribution to Foveal and Peripheral Vision". In: *Journal of Neuroscience* 32.27 (2012), pp. 9194–9204.

[112] Rucci, M. and Victor, J. D. "The unsteady eye: an information-processing stage, not a bug". In: *Trends in Neurosciences* 38.4 (2015), pp. 195–206.

[113] Mannan, S. K., Ruddock, K. H., and Wooding, D. S. "Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-D images." In: *Spatial vision* 9.3 (1995), pp. 363–86.

[114] Mannan, S. K., Ruddock, K. H., and Wooding, D. S. "Fixation patterns made during brief examination of two-dimensional images". In: *Perception* 26.8 (1997), pp. 1059–1072.

[115] Levenshtein, V. I. *Binary codes capable of correcting deletions, insertions, and reversals*. 1966.

[116] Wilson, A. T., Rintoul, M. D., and Valicka, C. G. “Exploratory Trajectory Clustering with Distance Geometry”. In: *Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience, Volume 9744 of the series Lecture Notes in Computer Science*. 2016, pp. 263–274.

[117] Bridgeman, B., Hendry, D., and Stark, L. “Failure to detect displacement of the visual world during saccadic eye movements”. In: *Vision Research* 15.6 (1975), pp. 719–722.

[118] Burr, D. C., Morrone, M. C., and Ross, J. *Selective suppression of the magnocellular visual pathway during saccadic eye movements.* 1994.

[119] Bahill, A. T. and Troost, B. “Types of saccadic eye movements”. In: *Neurology* 29.8 (1979), pp. 1150–1152.

[120] Tole, J. and Young, L. “Digital filters for saccade and fixation detection”. In: *Eye movements: Cognition and visual Perception*. Hillsdale, NJ: Lawrence Erlbaum., 1981, pp. 185–199.

[121] Behrens, F., Mackeben, M., and Schröder-Preikschat, W. “An improved algorithm for automatic detection of saccades in eye movement data and for calculating saccade parameters.” In: *Behavior research methods* 42.3 (2010), pp. 701–708.

[122] Nyström, M. and Holmqvist, K. “An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data”. In: *Behavior Research Methods* 42.1 (2010), pp. 188–204.

[123] Ester, M., Kriegel, H. P., Sander, J., and Xu, X. “A density-based algorithm for discovering clusters in large spatial databases with noise.” In: *Kdd* (1996).

[124] Salvucci, D. D. and Goldberg, J. H. “Identifying Fixations and Saccades in Eye-Tracking Protocols”. In: *Proceedings of the Eye Tracking Research and Applications Symposium* (2000), pp. 71–78.

[125] Kasneci, E., Kasneci, G., Kübler, T. C., and Rosenstiel, W. “The applicability of probabilistic methods to the online recognition of fixations and saccades in dynamic scenes”. In: *Etra* JANUARY (2014), pp. 323–326.

[126] Urruty, T., Lew, S., Djeraba, C., and Simovici, D. A. “Detecting eye fixations by projection clustering”. In: *Proceedings - 14th International Conference on Image Analysis and Processing Workshops, ICIAP 2007* 3.4 (2007), pp. 45–50.

[127] Li, B., Wang, Q., Barney, E., Hart, L., Wall, C., Chawarska, K., De Urabain, I., Smith, T., and Shic, F. “Modified DBSCAN algorithm on oculomotor fixation identification”. In: *Eye Tracking Research and Applications Symposium (ETRA)* 14 (2016), pp. 337–338.

[128] Lorenceau, J. “Cursive writing with smooth pursuit eye movements”. In: *Current Biology* 22.16 (2012), pp. 1506–1509.

[129] Esteves, A., Velloso, E., Bulling, A., and Gellersen, H. “Orbits: Gaze Interaction for Smart Watches using Smooth Pursuit Eye Movements”. In: *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology - UIST ’15* 1 (2015), pp. 457–466.

[130] Stern, J. a., Walrath, L. C., and Goldstein, R. “The endogenous eyeblink.” In: *Psychophysiology* 21.1 (1984), pp. 22–33.

[131] Siegle, G. J., Ichikawa, N., and Steinhauer, S. “Blink before and after you think: Blinks occur prior to and following cognitive load indexed by pupillary responses”. In: *Psychophysiology* 45 (2008), pp. 679–687.

[132] Stern, J. A., Boyer, D., and Schroeder, D. “Blink Rate: A Measure of Fatigue”. In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 36.2 (1994), pp. 285–297.

[133] Stern, J. A., Boyer, D. J., Schroeder, D. J., and Touchstone, R. M. *Blinks, Saccades, and Fixation Pauses During Vigilance Task Performance: II. Gender and Time of Day*. Tech. rep. DTIC Document, 1996.

[134] Schleicher, R., Galley, N., Briest, S., and Galley, L. “Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired?” In: *Ergonomics* 51.7 (2008), pp. 982–1010.

[135] Caffier, P. P., Erdmann, U., and Ullsperger, P. “Experimental evaluation of eye-blink parameters as a drowsiness measure.” In: *European journal of applied physiology* 89.3-4 (2003), pp. 319–325.

[136] Gilland, J. “Driving, eye-tracking and visual entropy: Exploration of age and task effects”. PhD. 2008, p. 192.

[137] Shultz, S., Klin, A., and Jones, W. “Inhibition of eye blinking reveals subjective perceptions of stimulus salience”. In: *Proceedings of the National Academy of Sciences of the United States of America* 108.52 (2011), pp. 21270–21275.

[138] Lins, O. G., Picton, T. W., Berg, P., and Scherg, M. “Ocular Artifacts in EEG and event related potentials II: scalp topography”. In: *Brain topography* 6.1 (1993), pp. 51–63.

[139] Hoffmann, S. and Falkenstein, M. “The Correction of Eye Blink Artefacts in the EEG: A Comparison of Two Prominent Methods”. In: *PLoS ONE* 3.8 (2008), e3004.

[140] Humboldt-Universität. *Natural Reading, Dataset 3*. 2013.

[141] Berg, P. and Scherg, M. “Dipole models of eye movements and blinks”. In: *Electroencephalography and clinical neurophysiology* 79.1 (1991), pp. 36–44.

[142] Ellis, C. J. “The pupillary light reflex in normal subjects.” In: *The British journal of ophthalmology* 65.11 (1981), pp. 754–9.

[143] Heller, P. H., Perry, F., Jewett, D. L., and Levine, J. D. “Autonomic components of the human pupillary light reflex”. In: *Investigative Ophthalmology and Visual Science* 31.1 (1990), pp. 156–162.

[144] Beatty, J. and Lucero-Wagoner, B. *The Pupillary System*. Ed. by Cacioppo, J., Tassinary, L., and Berntson, G. Cambridge University Press, 2000. Chap. 6, pp. 142–162.

[145] Bradshaw, J. L. “Background light intensity and the pupillary reponse in a reaction time task”. In: *Psychonomic Science* 14.6 (1969), pp. 271–272.

[146] Research, S. *SR Research Web Page*. 2016.

[147] Lang, P. J. and Bradley, M. M. “Emotion and the motivational brain”. In: *Biological Psychology* 84.3 (2010), pp. 437–450.

[148] Wang, W., Li, Z., Wang, Y., and Chen, F. “Indexing cognitive workload based on pupillary response under luminance and emotional changes”. In: *Proceedings of the 2013 international conference on Intelligent user interfaces*. ACM, 2013, pp. 247–256.

[149] Klingner, J., Kumar, R., and Hanrahan, P. “Measuring the Task-Evoked Pupillary Response with a Remote Eye Tracker”. In: *Proceedings of the 2008 symposium on Eye tracking research & applications.* 1.212 (2008), pp. 69–72.

[150] Stark, L., Campbell, F. W., and Atwood, J. “Pupil Unrest: An Example of Noise in a Biological Servomechanism”. In: *Nature* 182.4639 (Sept. 1958), pp. 857–858.

[151] Komogortsev, O. V., Karpov, A., Price, L. R., and Aragon, C. “Biometric authentication via oculomotor plant characteristics”. In: *2012 5th IAPR International Conference on Biometrics (ICB)*. IEEE, Mar. 2012, pp. 413–420.

[152] Holland, C. D. and Komogortsev, O. V. “Biometric verification via complex eye movements: The effects of environment and stimulus”. In: *2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)* (2012), pp. 39–46.

[153] Rigas, I. and Komogortsev, O. V. “Biometric recognition via fixation density maps”. In: ().

[154] May, J. G., Kennedy, R. S., Williams, M. C., Dunlap, W. P., and Brannan, J. R. “Eye movement indices of mental workload”. In: *Acta Psychologica* 75.1 (1990), pp. 75–89.

[155] Recarte, M. A. and Nunes, L. M. “Mental workload while driving: effects on visual search, discrimination, and decision making.” In: *Journal of experimental psychology. Applied* 9.2 (2003), pp. 119–137.

[156] Kahneman, D. and Beatty, J. “Pupil Diameter and Load on Memory”. In: *Science* 154.3756 (1966), pp. 1583–1585.

[157] Beatty, J. *Task-evoked pupillary responses, processing load, and the structure of processing resources*. 1982.

[158] Hyönä, J., Tommola, J., and Alaja, A.-M. “Pupil Dilation as a Measure of Processing Load in Simultaneous Interpretation and Other Language Tasks”. In: *The Quarterly Journal of Experimental Psychology Section A* 48.3 (1995), pp. 598–612.

[159] Einhäuser, W., Stout, J., Koch, C., and Carter, O. “Pupil dilation reflects perceptual selection and predicts subsequent stability in perceptual rivalry.” In: *Proceedings of the National Academy of Sciences of the United States of America* 105.5 (2008), pp. 1704–9.

[160] Privitera, C. “Pupil dilation during visual target detection”. In: *Journal of . . .* 10.2010 (2010), pp. 1–14.

[161] Preuschoff, K., ’t Hart, B. M., and Einh??user, W. “Pupil dilation signals surprise: Evidence for noradrenaline’s role in decision making”. In: *Frontiers in Neuroscience* 5.SEP (2011), pp. 1–12.

[162] Siegle, G. J., Steinhauer, S. R., Carter, C. S., Ramel, W., and Thase, M. E. “Do the seconds turn into hours? Relationships between sustained pupil dilation in response to emotional information and self-reported rumination”. In: *Cognitive Therapy and Research* 27.3 (2003), pp. 365–382.

[163] Gee, J. W. de, Knapen, T., and Donner, T. H. “Decision-related pupil dilation reflects upcoming choice and individual bias.” In: *Proceedings of the National Academy of Sciences of the United States of America* 111.5 (2014), E618–25.

[164] Kuchinsky, S. E., Pandža, N. B., and Haarmann, H. J. “Linking Indices of Tonic Alertness: Resting-State Pupil Dilation and Cingulo-Opercular Neural Activity”. In: *Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience: 10th International Conference, AC 2016, Held as Part of HCI International 2016, Toronto, ON, Canada, July 17-22, 2016,*

Proceedings, Part I. Ed. by Schmorrow, D. D. and Fidopiastis, M. C. Cham: Springer International Publishing, 2016, pp. 218–230.

[165] Kuipers, J. R. and Thierry, G. “N400 amplitude reduction correlates with an increase in pupil size.” In: *Frontiers in human neuroscience* 5.June (2011), p. 61.

[166] Mill, R. D., O’Connor, A. R., and Dobbins, I. G. “Pupil dilation during recognition memory: Isolating unexpected recognition from judgment uncertainty”. In: *Cognition* 154 (2016), pp. 81–94.

[167] Bradley, M. M. and Lang, P. J. “Memory, emotion, and pupil diameter: Repetition of natural scenes”. In: *Psychophysiology* (2015).

[168] Zekveld, A. A., Heslenfeld, D. J., Johnsrude, I. S., Versfeld, N. J., and Kramer, S. E. “The eye as a window to the listening brain: Neural correlates of pupil size as a measure of cognitive listening load”. In: *NeuroImage* 101 (2014), pp. 76–86.

[169] Partala, T. and Surakka, V. “Pupil size variation as an indication of affective processing”. In: *International Journal of Human Computer Studies* 59.1-2 (2003), pp. 185–198.

[170] Stanners, R. F., Coulter, M., Sweet, A. W., and Murphy, P. “The pupillary response as an indicator of arousal and cognition”. In: *Motivation and Emotion* 3.4 (1979), pp. 319–340.

[171] Bradley, M. M., Miccoli, L., Escrig, M. a., and Lang, P. J. “The pupil as a measure of emotional arousal and autonomic activation”. In: *Psychophysiology* 45.4 (2008), pp. 602–607.

[172] Geng, J. J., Blumenfeld, Z., Tyson, T. L., and Minzenberg, M. J. “Pupil diameter reflects uncertainty in attentional selection during visual search”. In: *Frontiers in Human Neuroscience* 9.August (Aug. 2015), pp. 1–14.

[173] Laeng, B., Ørbo, M., Holmlund, T., and Miozzo, M. “Pupillary stroop effects”. In: *Cognitive Processing* 12.1 (2011), pp. 13–21.

[174] Beatty, J. “Phasic Not Tonic Pupillary Responses Vary With Auditory Vigilance Performance”. In: *Psychophysiology* 19.2 (1982), pp. 167–172.

[175] Alnæs, D., Sneve, M. H., Espeseth, T., Endestad, T., Pavert, S. H. P. van de, and Laeng, B. “Pupil size signals mental effort deployed during multiple object tracking and predicts brain activity in the dorsal attention network and the locus coeruleus”. In: *Journal of Vision* 14.4 (2014), p. 1.

[176] Murphy, P. R., O'Connell, R. G., O'Sullivan, M., Robertson, I. H., and Balsters, J. H. "Pupil diameter covaries with BOLD activity in human locus coeruleus". In: *Human Brain Mapping* 35.8 (2014), pp. 4140–4154.

[177] Wilson, G. F., Caldwell, J. a., and Russell, C. a. "Performance and Psychophysiological Measures of Fatigue Effects on Aviation Related Tasks of Varying Difficulty". In: *The International Journal of Aviation Psychology* 17.906692079 (2007), pp. 219–247.

[178] Liao, H.-I., Yoneya, M., Kidani, S., Kashino, M., and Furukawa, S. "Human Pupillary Dilation Response to Deviant Auditory Stimuli: Effects of Stimulus Properties and Voluntary Attention". In: *Frontiers in Neuroscience* 10.February (2016).

[179] Gabay, S., Pertzov, Y., and Henik, A. "Orienting of attention, pupil size, and the norepinephrine system." In: *Attention, perception & psychophysics* 73.1 (2011), pp. 123–9.

[180] Eldar, E., Cohen, J. D., and Niv, Y. "The effects of neural gain on attention and learning." In: *Nature neuroscience* 16.8 (2013), pp. 1146–53.

[181] Galley, N. "Saccadic eye movement velocity as an indicator of (de) activation: A review and some speculations." In: *Journal of Psychophysiology* August (1989), pp. 229–244.

[182] McGregor, D. K. and Stern, J. a. "Time on task and blink effects on saccade duration." In: *Ergonomics* 39.4 (1996), pp. 649–60.

[183] Di Stasi, L. L., Catena, A., Ca??as, J. J., Macknik, S. L., and Martinez-Conde, S. "Saccadic velocity as an arousal index in naturalistic tasks". In: *Neuroscience and Biobehavioral Reviews* 37.5 (2013), pp. 968–975.

[184] Dambacher, M. and Kliegl, R. "Synchronizing timelines: Relations between fixation durations and N400 amplitudes during sentence reading". In: *Brain Research* 1155.1 (2007), pp. 147–162.

[185] Kliegl, R., Dambacher, M., Dimigen, O., Jacobs, A. M., and Sommer, W. "Eye movements and brain electric potentials during reading". In: *Psychological Research* 76.2 (2012), pp. 145–158.

[186] James, W. *The principles of psychology, Vol I.* 1890.

[187] Von Helmholtz, H. *Handbuch der physiologischen Optik*. 1867.

[188] Gibson, J. J. "A critical review of the concept of set in contemporary experimental psychology." In: *Psychological Bulletin* 38.9 (1941), pp. 781–817.

[189] Posner, M. I. and Petersen, S. E. “The attention system of the human brain.” In: *Annual review of neuroscience* 13 (1990), pp. 25–42.

[190] Petersen, S. and Posner, M. “The Attention System of the Human Brain: 20 Years After”. In: *Annual review of neuroscience* 21.35 (2012), pp. 73–89.

[191] Corbetta, M. and Shulman, G. L. “Control of goal-directed and stimulus-driven attention in the brain.” In: *Nature reviews. Neuroscience* 3.3 (2002), pp. 201–15.

[192] Shulman, G. L. and Corbetta, M. “Two Attentional Networks”. In: *Cognitive Neuroscience of Attention, 2nd ed.* 2012, pp. 113–128.

[193] Velichkovsky, B. M., Joos, M., Helmert, J. R., and Pannasch, S. “Two Visual Systems and their Eye Movements: Evidence from Static and Dynamic Scene Perception”. In: *Proceedings of the XXVII conference of the cognitive science society* (2005), pp. 2283–2288.

[194] Aston-Jones, G., Rajkowski, J., and Cohen, J. “Role of locus coeruleus in attention and behavioral flexibility”. In: *Biological Psychiatry* 46.9 (1999), pp. 1309–1320.

[195] Joshi, S., Li, Y., Kalwani, R. M., and Gold, J. I. “Relationships between Pupil Diameter and Neuronal Activity in the Locus Coeruleus, Colliculi, and Cingulate Cortex”. In: *Neuron* 89.1 (2016), pp. 221–234.

[196] Aston-Jones, G. and Cohen, J. D. “An Integrative Theory of Locus Coeruleus-Norepinephrine Function: Adaptive Gain and Optimal Performance”. In: *Annual Review of Neuroscience* 28.1 (2005), pp. 403–450.

[197] Bouret, S. and Sara, S. J. “Reward expectation, orientation of attention and locus coeruleus-medial frontal cortex interplay during learning”. In: *European Journal of Neuroscience* 20.3 (2004), pp. 791–802.

[198] Geva, R., Zivan, M., Warsha, A., and Olchik, D. “Alerting, orienting or executive attention networks: differential patters of pupil dilations.” In: *Frontiers in behavioral neuroscience* 7.October (2013), p. 145.

[199] Ambrosini, E., Vastano, R., Montefinese, M., and Ciavarro, M. “Functional specificity of the locus coeruleus-norepinephrine system in the attentional networks”. In: *Frontiers in behavioral neuroscience* 7 (2013), pp. 1–2.

[200] Yerkes, R. M. and Dodson, J. D. “The relation of strength of stimulus to rapidity of habit-formation”. In: *Journal of Comparative Neurology and Psychology* 18.5 (1908), pp. 459–482.

[201] Kahneman, D. *Attention and effort*. Prentice-Hall, Inc, 1973, pp. 1–339.

[202] Jepma, M. and Nieuwenhuis, S. “Pupil diameter predicts changes in the exploration-exploitation trade-off: evidence for the adaptive gain theory.” In: *Journal of cognitive neuroscience* 23.7 (2011), pp. 1587–1596.

[203] Chmielewski, W. X., Mückschel, M., Ziemssen, T., and Beste, C. “The norepinephrine system affects specific neurophysiological subprocesses in the modulation of inhibitory control by working memory demands”. In: *Human Brain Mapping* 00.August (2016), pp. 1–14.

[204] Stark, L. and Sherman, P. M. “A servoanalytic study of consensual pupil reflex to light”. In: *Journal of neurophysiology* 20.1 (1957), pp. 17–26.

[205] Privitera, C. M. and Stark, L. W. “A binocular pupil model for simulation of relative afferent pupil defects and the swinging flashlight test”. In: *Biological Cybernetics* 94.3 (2006), pp. 215–224.

[206] Laeng, B., Sirois, S., and Gredeback, G. “Pupillometry: A Window to the Preconscious?” In: *Perspectives on Psychological Science* 7.1 (2012), pp. 18–27.

[207] Hess, E. H. and Polt, J. M. “Pupil Size in Relation to Mental Activity during Simple Problem-Solving”. In: *Science* 143.3611 (1964), pp. 1190–1192.

[208] Hess, E. H. and Polt, J. M. “Pupil size as related to interest value of visual stimuli”. In: *Science* 132 (1960), pp. 349–50.

[209] Smith, J. D., Ichinose, L. Y., Masek, G. A., Watanabe, T., and Stark, L. “Midbrain Single Units Correlating with Pupil Response to Light”. In: *Science* 162.3859 (1968), pp. 1302–1303.

[210] Ahern, S. and Beatty, J. “Pupillary responses during information processing vary with Scholastic Aptitude Test scores”. In: *Science* 205.4412 (1979), pp. 1289–1292.

[211] Wang, C. A. and Munoz, D. P. *A circuit for pupil orienting responses: Implications for cognitive modulation of pupil size*. 2015.

[212] Costa, V. D. and Rudebeck, P. H. “More than Meets the Eye: The Relationship between Pupil Size and Locus Coeruleus Activity”. In: *Neuron* 89.1 (2016), pp. 8–10.

[213] Varazzani, C., San-Galli, A., Gilardeau, S., and Bouret, S. “Noradrenaline and Dopamine Neurons in the Reward/Effort Trade-Off: A Direct Electrophysiological Comparison in Behaving Monkeys”. In: *Journal of Neuroscience* 35.20 (May 2015), pp. 7866–7877.

[214] Ang, Y.-S., Manohar, S., and Apps, M. A. J. “Commentary: Noradrenaline and Dopamine Neurons in the Reward/Effort Trade-off: A Direct Electrophysiological Comparison in Behaving Monkeys”. In: *Frontiers in Behavioral Neuroscience* 9.November (Nov. 2015), p. 310.

[215] Iqbal, S. T., Zheng, X. S., and Bailey, B. P. “Task-evoked pupillary response to mental workload in human-computer interaction”. In: *Extended abstracts of the 2004 conference on Human factors and computing systems CHI 04* (2004), p. 1477.

[216] Palinko, O., Kun, A. L., Shyrokov, A., and Heeman, P. “Estimating cognitive load using remote eye tracking in a driving simulator”. In: *Proceedings of the 2010 symposium on eye-tracking research & applications*. ACM. 2010, pp. 141–144.

[217] Marshall, S. P. “The Index of Cognitive Activity: measuring cognitive workload”. In: *Proceedings of the IEEE 7th Conference on Human Factors and Power Plants* (2002), pp. 5–9.

[218] Marshall, S. P. “Identifying cognitive state from eye metrics”. In: *Aviation Space and Environmental Medicine* 78.5 II (2007), pp. 165–175.

[219] Lew, R., Dyre, B. P., Soule, T., Ragsdale, S. A., and Werner, S. “Assessing Mental Workload From Skin Conductance and”. In: *Proceedings of tehe Human Factors and Ergonomics Society* (2010), pp. 254–258.

[220] Pinzon-Morales, R.-D. and Hirata, Y. “Customization of Wavelet Function for Pupil Fluctuation Analysis to Evaluate Levels of Sleepiness”. In: *Journal of Communication and Computer* 10 (2013), pp. 585–592.

[221] Pedrotti, M., Mirzaei, M. A., Tedesco, A., Chardonnet, J.-r., Mérienne, F., Benedetto, S., and Baccino, T. “Automatic stress classification with pupil diameter analysis”. In: *Int. Journal of Human-Computer Interaction* 30.March (2014), pp. 1–17.

[222] Lew, R., Dyre, B. P., Werner, S., Wotring, B., and Tran, T. “Exploring the Potential of Short-Time Fourier Transforms for Analyzing Skin Conductance and Pupillometry in Real-Time Applications”. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (2016), pp. 1536–1540.

[223] Nakayama, M. and Shimizu, Y. “Frequency analysis of task evoked pupillary response and eye-movement”. In: *Proceedings of the Eye tracking research & applications symposium on Eye tracking research & applications - ETRA’2004* (2004), pp. 71–76.

[224] Lüdtke, H., Wilhelm, B., Adler, M., Schaeffel, F., and Wilhelm, H. "Mathematical procedures in data recording and processing of pupillary fatigue waves". In: *Vision Research* 38.19 (1998), pp. 2889–2896.

[225] Klingner, J. "Fixation-aligned pupillary response averaging". In: *Proceedings of the 2010 Symposium on Eye-Tracking . . .* 1.212 (2010), pp. 275–282.

[226] Hoeks, B. and Levelt, W. J. M. "Pupillary dilation as a measure of attention: a quantitative system analysis". In: *Behavior Research Methods, Instruments, & Computers* 25.1 (1993), pp. 16–26.

[227] Wierda, S. M., Rijn, H. van, Taatgen, N. a., and Martens, S. "Pupil dilation deconvolution reveals the dynamics of attention at high temporal resolution". In: *Proceedings of the National Academy of Sciences* 109.22 (2012), pp. 8456–8460.

[228] Gollan, B. and Ferscha, A. "Modeling Pupil Dilation as Online Input for Estimation of Cognitive Load in non-laboratory Attention-Aware Systems". In: *COGNITIVE 2016, The Eighth International Conference on Advanced Cognitive Technologies and Applications* c (2016), pp. 55–61.

[229] Daugman, J. G. *US Patent US 5,291,560 A: Biometric personal identification system based on iris analysis*. 1994.

[230] Daugman, J. "How iris recognition works". In: *Circuits and Systems for Video Technology, IEEE Transactions on* 14.1 (2004), pp. 21–30.

[231] Boles, W. "A security system based on human iris identification using wavelet transform". In: *Engineering Applications of Artificial Intelligence* 11.1 (1998), pp. 77–85.

[232] Hill, R. B. *US Patent 4,109,237: Apparatus and method for identifying individuals through their retinal vasculature patterns*. 1978.

[233] Jain, A., Hong, L., and Pankanti, S. "Biometric identification". In: *Commun. ACM* 43.2 (2000), pp. 90–98.

[234] Kasprowski, P. and Ober, J. "Eye Movements in Biometrics 2 Physiology of Eye Movements 3 Previous Researches Concerning Eye Movements". In: (2004), pp. 248–258.

[235] Komogortsev, O., Holland, C., Karpov, A., and Price, L. R. "Biometrics via Oculomotor Plant Characteristics : Impact of Parameters in Oculomotor Plant Model". In: 11.4 (2014), pp. 1–17.

[236] Bednarik, R., Kinnunen, T., Mihaila, A., and Franti, P. "Eye-movements as a biometric". In: *Image Analysis, Proceedings* 3540 (2005), pp. 780–789.

[237] Andersen, E. F. “Eye tracking of patterns for detecting personal state of attention (Student Paper)”. 2015.

[238] Kullback, S. and Leibler, R. a. “On Information and Sufficiency”. In: *The Annals of Mathematical Statistics* 22.1 (1951), pp. 79–86.

[239] Fischer, B. and Boch, R. “Saccadic eye movements after extremely short reaction times in the rhesus monkey”. In: *Brain Research* 260 (1983), pp. 21–26.

[240] Williams, R. S., Biel, A. L., Wegier, P., Lapp, L. K., Dyson, B. J., and Spaniol, J. “Age differences in the Attention Network Test: Evidence from behavior and event-related potentials”. In: *Brain and Cognition* 102 (2016), pp. 65–79.

[241] Bellaera, L. and Mühlenen, A. von. “The effect of induced sadness and moderate depression on attention networks”. In: *Cognition and Emotion* 9931.June (2016), pp. 1–13.

[242] Bouret, S. and Sara, S. J. “Network reset: A simplified overarching theory of locus coeruleus noradrenaline function”. In: *Trends in Neurosciences* 28.11 (2005), pp. 574–582.

[243] Harrysson, M., Métayer, E., and Sarrazin, H. “The strength of ’weak signals’”. In: *McKinsey Quarterly* (2014), p. 4.

[244] Lehmann, S. and Lassen, D. D. *Sensibe DTU Web Page*. 2016.

[245] Abdullah, S., Matthews, M., Murnane, E. L., and Gay, G. “Towards Circadian Computing : “ Early to Bed and Early to Rise ” Makes Some of Us Unhealthy and Sleep Deprived”. In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2014), pp. 673–684.

Technical University of Denmark
Cognitive Systems
Richard Petersens Plads
DK-2800, Kgs. Lyngby
Denmark
Phone: (+45) 45 25 30 31
Email: compute@compute.dtu.dk
www.compute.dtu.dk