Giovanni Pignoni

Development of a quantitative evaluation tool of cognitive workload in field studies through eye tracking

Master's thesis

NTNU
Norwegian University of
Science and Technology

Giovanni Pignoni

Development of a quantitative evaluation tool of cognitive workload in field studies through eye tracking

NTNU

Norwegian University of
Science and Technology

Contents

Preface

This dissertation covers the author's master's thesis project in MIXD developed at NTNU Gjøvik in the fall 2018 and spring 2019 semesters and is based on a conference paper, accepted and due to be published, titled "Development of a quantitative evaluation tool of cognitive workload in field studies through eye tracking". It will be presented in July 2019 at the HCI International 2019 Conference in Orlando (Pignoni & Komandur 2019) (see appendix A.2). The laboratory test was carried out with the generous help of the Norwegian Colour and Visual Computing Laboratory in Gjøvik, in particular with the collaboration of Peter Nussbaum. The field test was done in collaboration with the Norwegian Naval Academy (RNoNA) and Odd Sveinung Hareide (Technical Manager Electronic Navigation and Integrated Navigation Systems). The data analysis was performed with the supervision of the Prof. Frode Volden from NTNU Gjøvik.

01-05-2019
Giovanni Pignoni

Acknowledgment

Abstract

Eye tracking is primarily intended as a mean for tracking the visual attention of an observer/-operator. Still, eye tracking is also capable of recording a wider variety of psychophysical data, one of these is the trace effect of mental workload in the pupil diameter (dilatation of the pupil proportional to the level of cognitive workload). The effect of a varying visual stimulus, primarily luminance variations, on the pupil diameter has historically constrained studies on workload and pupillometry to controlled laboratory conditions, as, in field conditions, it cannot be accurately determined if the change in the pupil diameter of the subject is due to mental workload alone. Although it has been attempted to simultaneously account for the contribution of the change in pupil diameter due to luminance and mental workload, no instance has been found of an attempt to use video and luminance data dynamically to estimate the effect of the visual stimuli as it changes during a study session. Eye-tracking data will be used alongside the video feed of the field of view of the operator/observer and the data from an external luminance sensor to quantify the luminous flux from the point of view of the subject and estimate the expected pupil size. The estimated pupil size should then be used as a variating baseline to remove the effect of luminance on the pupil size and isolate the changes relative to mental workload. The newly developed methodology will be evaluated in both laboratory and field conditions such as tracking the experience of the navigator and his first officer in a high-speed marine craft of the Royal Norwegian Naval Academy alongside subjective data such as the widely used NASA-TLX for validation purposes.

1   Introduction

Interaction design in safety-critical systems differs from the process of designing consumer goods with a series of challenges peculiar to this discipline. One of particular interest is the simulation and monitoring of the operator status for the evaluation of the interface performances. Simulations are often necessary for user testing in safety-critical, when the scenarios of interest are too unpractical to replicate or would pose a considerable safety risk on the operator or the people and environment involved. It is therefore often preferable, even though it can lead to generalisability problem as the quality of the simulation itself (fidelity to the original system and environment) directly affects the quality of the findings. Monitoring the operator, either during a simulation or real life, is, therefore, the process of collecting data about his physical and cognitive state, as it changes over time. The scope of such activity in interaction design can vary, but in general, it has the scope to identify physical or mental overload, prevent bad habits as well as evaluating the qualities of the user interface. The available tool-set for monitoring an operator status during his activity has grown and is more accessible than ever, including techniques such as heart rate (ECG), heart rate variability, brain activity (fMRI, EEG), eye tracking (pupillometry, gaze position, saccades velocity). The actual application of these technologies in a field study is limited by the counteracting need of reducing confounding variables, and while the practical application is overall easier (bulky tools getting portable and less limiting for the user freedom), this remains an open problem. Correct experiment design can account and minimise this problem as well as the choice of unobtrusive data gathering tools such as eye tracking. Eye tracking has the inherent advantage of being used to evaluate a variety of parameters, not only limited to vision, with a limited or null impact on the user. Ambient illumination has historically limited the application eye tracking to controlled studies, either because of the effects of luminance on the pupil or the destructive interference of infrared (sun) light on the tracking technology used by many of the commercially available eye trackers. This research aims to understand how to account for this environmental variable and validate the use of pupillometry in a field study.

1.1   Keywords

- Eye Tracking.
- Pupillometry.
- Cognitive Workload.
- Field Study.
- Maritime Usability.

1.2   Motivation and Benefits

Cognitive workload and pupillary responses have been investigated as back as Hess & Polt (1964) where authors monitored the pupil size of a subject intent solving "simple multiplication problems" and were able to observe a link between pupillary response and difficulty level. The task-

evoked pupillary diameter was linked to memory intensive tasks by Kahneman & Beatty (1966) who report observing different pupillary responses occurring during the learning and recollect phases. Similar results have been obtained for tasks such as writing, listening, speech and mathematical problems (Kahneman et al. 1967).

Psycho-physiological studies involving pupillary response can be limited by three main factors: luminance, low-temporal-resolution and high cost.

- Luminance, perceived from the point of view (POV) of the subject provokes an involuntary response of the pupil. The majority of the reviewed literature describes laboratory conditions in which the luminance is fixed or highly controlled, limiting the generalisability of such methods to field conditions. The use of pupil diameter in field conditions requires a different approach to discern the effect of light from that of mental workload.
- Although High-temporal-resolution tracking of cognitive workload has been attempted before, Wierda et al. (2012) and Marshall (1999) in particular, it is rarely possible as it requires a constantly changing baseline value based on the visual stimuli, limiting the study to the evaluation of well-defined tasks.
- The only method that offers a solution to both the previous problems is closed source and proprietary and only supports a limited number of eye trackers (Marshall 1999).

1.2.1 Problem Description

As per the literature review, there is currently no open source, validated, and accessible method to measure cognitive workload in a field condition trough pupillometry. For a researcher that would want to add the evaluation of cognitive workload to a usability study, the available method (a method that wouldn't require the development of a custom system) is proprietary (Marshall 1999) ad thus closed source. This condition not only poses a significant economic barrier, but also disconnects the researcher from the tools and the ability to adapt it to specific research questions and have a deep understanding of the variables at play (Holmqvist & Andersson 2017). Moreover, no method has been validated for use with low-cost eye trackers, which would enable affordable data gathering, including collaborative studies with multiple eye trackers.

1.3 Research Questions

- Is it possible to reliably measure luminance from the POV of a subject using a small calibrated video camera?
- Is it possible to reliably calculate the baseline pupil size for a visual stimulus and use it to infer the cognitive state of the subject?
- Is this method reliable in field condition (where luminance variates in an unpredictable manner)?

1.4 Relevance

In the design of a complex interactive system, it can be of critical importance to asses and design the load that such a system has on the user. Unregulated cognitive load can have detrimental effects on the user's performances in cases of both under-load and overload. Onboard interfaces on cars are an example of human-computer interaction that has the potential to increase the

operator (driver) workload to unsafe levels (Reimer et al. 2009). A workload-aware system can help prioritise and filter information based on the driving situation and driver state. Similarly, autonomous driving systems will require monitoring of the driver state to account for the operator readiness in taking back control of the vehicle in lower tier automation. Level two and three automation requires the driver to be attentive and take control of the vehicle at any time. The safety of the system is then tied not only on the automation but also on the constant supervision of the operator, potentially resulting in under-load for the driver. The performance to load curve is U shaped, under-load is to be considered as dangerous as over-load, in this case resulting in the inability for the operator to take action in time.

Safety critical systems that require online attention of the user are dependent on the user's ability to respond in time and coherently to the system state. Examples of such system can be highly specialised tasks as air traffic control or everyday tasks such as driving a car. In general, any system in which the cognitive load is variable and the user does not have or has only limited control over the load. Such as the ability to take a break at any moment or to concentrate on/monitor one input at a time. The task difficulty can provoke a change in the user cognitive and emotional state, with a consequent impact on performance. Overload can be the effect of a task exceeding the user skill level, as well as the consequence of fatigue or other alteration of the psycho-physical state of the user. Overload can also cause unwanted consequences such as anxiety, increased error rate and increased fatigue. Under-load is also generally undesirable as it can potentially provoke a lack of attention, boredom, slow response rate, lack of focus and eventually lead to low productivity as well as dangerous conditions. For this reason, the control and assessment of cognitive workload should always be a design goal in the development of a safety-critical system to ensure that the best possible combination of user performance, safety and comfort. A measure of the cognitive workload can have two main applications in User Centred Design:

- Analysis of a system concerning how the workload changes over time for the operator, to identify overload and under-load situations, how they can occur and prevent/mitigate them in the design stage.
- Design and development of dynamic difficulty adjustment (DDA) (Afergan et al. 2014) systems that can respond to the user state and system state to adjust the task in real-time as well as identify and respond to potentially dangerous situations (e.g. prompt the operator to take a break, re-route some or all the load to another operator, slow down or stop a process in a safe manner until normal operation can be resumed).

As noted in the literature review, the workload has been measured with a variety of tools: performance-based measures, self-report questionnaires, behavioural observation, and physiological measures (such as eye tracking). When the subject of the research includes a dynamically changing condition, performance and physiological based methods have an inherent advantage to observation or self-report tests (Mehler et al. 2009), most of all, being objective and continuous measures. Moreover, physiological indices can be more sensitive than performance-based in detecting low-level changes in workload occurring before the task performance is reduced in a measurable manner (Mehler et al. 2009).

2   Background

2.1   Notes on Previous Publication

The following section is based on a conference paper, accepted and due to be published, titled "Development of a quantitative evaluation tool of cognitive workload in field studies through eye tracking". It will be presented in July 2019 at the HCI International 2019 conference in Orlando (Pignoni & Komandur 2019) see appendix A.2. The paper includes a wider literature review and a "work in progress" description of the methods as they were planned out at the end of 2018.

2.2   Existing Research

The monitoring and quantification of cognitive workload, as either a subjective or objective measure, has evolved and branched over the years.

Council (1997) provides a simplified definition of what can be intended as Cognitive workload as the ratio between the time required to complete a task and the time available to do so. This can be a good way to introduce the concept, but the shortcomings are apparent when analysing situations in which the workload is provoked by other factors, such as insufficient training, anxiety, fear, fatigue, visual or auditory overload. A more generalised approach would be to consider workload as the ratio between the resources available and the resources necessary to complete a task with time considered as a secondary product of workload (e.g. for a given task, lower training results in a higher workload and possibly longer time).

Mental demands could be the primary source of workload even if low in magnitude. Conversely, the time pressure under which the task is performed could act as the primary source of workload for an otherwise simple task. Hart & Staveland (1988) defines workload as a subjective, "perceived" metric, related to the amount of "information processing" and "decision-making" required by a task. A physiological and mental demand related to the performance of a task or a combination of tasks. "Workload is not an inherent property, but rather it emerges from the interaction between the requirements of a task, the circumstances under which it is performed, and the skills, behaviours, and perceptions of the operator."

The several mental workload measurement methods that are currently available can be organised into three groups (Di Stasi et al. 2013), (Rusnock & Borghetti 2018), (Hart & Staveland 1988) :

- "Subjective Ratings" Subjective-empirical measures of perceived effort as rated by the subject. They offer high face validity and are generally "user acceptable" which ultimately makes them easy to administer, but are generally post-test, thus relying on memory and self-perception of the participant.
- "Performance Data" Performance of the subject in a controlled task. Either a primary or secondary task is used as a metric to evaluate the workload, this can be suitable to find overload and under-load situations but mixes the definition of workload to the definition

of performance. Moreover the addition of a secondary task can create itself overload and can result in low user acceptance.

- "Physiological Measures" Physiological indices of cognitive state are nonintrusive and provide data over time. They can be difficult to administer as they might require extensive setup.

2.2.1 Subjective Ratings

Subjective reports are often structured as "paper-and-pencil" questionnaires and multidimensional ratings (e.g. the NASA-TLX Hart & Staveland (1988)), they record the user perceived workload and thus are an indirect measure. Although usually easy and cheap to administer as well as providing high face validity, they have several limitations (Tsang & Velazquez 1996). Post-facto evaluations rely on the subject's perception and memory of the task; as such they could show low inter-rater reliability, thus requiring initial baseline procedures, and often give very little temporal information as they do not track changes throughout the task. Subjective ratings of interest are: Multiple Resources Questionnaire (MRQ) (Boles & Adair 2001), Subjective Workload Assessment Technique (SWAT) (Rubio et al. 2004) , Overall Workload Level (OWL) (Jung & Jung 2001) and Integrated Workload Scale (IWS) (Pickup et al. 2005); the NASA Task Load Index was selected (Hart & Staveland 1988) for its widespread use and relevance.

NASA Task Load Index

The NASA Task Load Index is a multi-dimensional rating procedure that provides an overall workload score based on a weighted average of ratings on six subscales: Mental Demands, Physical Demands, Temporal Demands, Own Performance, Effort and Frustration (Human Performance Research Group n.d.).

The NASA-TLX is a standardised tool developed at the NASA-Ames Research Center for the evaluation of workload across a variety of activities. From an initial focus on in-flight activities, it evolved to be a generalised and widely used assessment tool. Earlier revisions of the scale included nine sub-scales in order to reduce inter-rater variability with the inclusion of a priori workload definitions of subjects to weight and average sub-scale rating (Human Performance Research Group n.d.), but were restructured in the later revisions (Hart & Staveland 1988) to reduce the number of sub-scales and improve the practical application in simulations and operational environment. Like other forms of self-report, it does not record changes in cognitive load over time (multiple questionnaires can be used to assist a complex task if divisible in sub-tasks).

The definition of workload variates among experimenters and subjects (reducing inter-rater variability), but it was found to be not as significant as the definition of the specific sources of loading making up the tasks, creating what is defined as "workload experiences" (Human Performance Research Group n.d.) which lead to the inclusion of subscale weighted according to the subjective importance of each the task. Ratings of the workload components, deemed most significant by the raters, are given more weight in the workload. The application manual Human Performance Research Group (n.d.) suggests that data can either be obtained in an operational setting (on-line) or retrospectively (videotaped/regenerated activities are suggested as mnemonic aid) (Human Performance Research Group n.d.). The NASA Task Load Index is divided into two parts: weights and ratings. Weights are the contribution of each factor to the

workload of a specific task (increase inter-rater reliability and provide a definition of the task workload). Ratings are the second part of the index, reflecting the magnitude of each factor in a given task. Ratings can be obtained during the task, in a pause or at the end of an entire task.

2.2.2 Performance Data

Primary Task Performance

Mental workload has to be measured indirectly, as a manifestation of the mental state, when not relying on the user self-report, an efficient but simplistic approach consists in the evaluation of the output efficiency of the subject. That is the comparison of the expected performance against the actual performance of a user for a well-defined task. The analysis of performance needs eventually to be interpreted as in itself can only directly represent the instantaneous load level. The context in which the task is performed, as well as the presence of overload or under-load, are not covered by simple performance data, hence the limitation of this model. Still, when the task can be evaluated and planned in terms of expected loads, such as a sequence of tasks that have a predictable and differentiable load, performance data can be a critical source of information.

Secondary Task Performance

When the primary task is not well defined or controllable, such as a long driving session, in which the workload is stable with distributed peaks, a similar analysis can be done on a secondary task. Attention and cognitive workload are intertwined psychological constructs. As per the Kahneman's model of attentional resources Kahneman (2012),attention can be seen as a limited resource. A dual-task performance measure implies that the secondary task is executed in a realm of limited and variable resources (resources used by the primary task). The variable load relative to the primary task is consequently affecting the performance of the secondary task. Changes (speed, accuracy, response time) in the execution of the secondary task can be interpreted as a difference in cognitive/visual workload related to the primary task. The ISO defined Detection Response Task (DRT) (ISO 2016) is an example of a performance-based method, Čegovnik et al. (2018) used a tactile DRT to validate the use of a Tribe eye tracker for the measurement of the cognitive load of the subjects using oculography and pupillometry. The DTR allows the assessment of cognitive load through the analysis of response rate and miss-rate of the response task: a stimulus is delivered through a vibrator attached to the subject in a random sequence; the test measures the response time (time needed to press a button attached to the steering wheel in a driving simulator). The use of a secondary task should be carefully planned considering the effect of the controls tasks on the main task as well as the low temporal resolution of the events that can be measured. The introduction of a secondary, artificial task potentially distracts the user or affects the primary performance. Furthermore, the relationship linking cognitive workload and task performance is not linear but follows an inverted U-shape as defined in the Hebb-Yerkes-Dodson Law (Čegovnik et al. 2018). Both over-load and under-load can lead to decreased performances creating the need for subjective data.

2.2.3 Physiological Indices

In light of the various limitations of subjective data and performance data, it becomes clear why the most active field of study for the cognitive workload is lead by the interest in non-intrusive

techniques to directly measure cognitive function. Psychophysiological techniques, in general, are based on a relationship between a behavioural phenomenon and a measurable activity of the central nervous system. Physiological indices indirectly connect a measurable psychophysical parameter to an expected mental workload. In the case of mental workload, this can include heart rate and heart rate variability, respiratory rate, galvanic skin response, brain activity (EEG, fRMITCD), as well as eye activity (Cacioppo et al. 2007). Eye tracking and pupillometry have the inherent advantage of being unobtrusive and less of a constraint for the user than the majority of the techniques mentioned above. The application of eye tracking of mental workload is promising (Di Stasi et al. 2013) and for usability studies, it is a natural extension of the more general use of gaze tracking. Psychophysiological measures also have several limitations to consider when applied in field conditions (Čegovnik et al. 2018). In most of the investigations mentioned, the experimental design involved two or more tasks with different levels of difficulty indirectly estimating the workload required for each of the tasks. Often resulting in the analysis of weak signals, subject to artefacts and noise with psychophysiological equipment that developed in In laboratory settings and that poorly suits the field, in terms of size and configuration. The physiological and neurological models employed in the psychophysiological methods have to be specially designed and trained to fit a particular task-evoked neural activity. It is, therefore, difficult to compare the results to a generalised measure of workload.

Pupil

In the Handbook of Psychophysiology, (Cacioppo et al. 2007, p. 443) defines the pupillary system as a "dually innervated organ". The pupil size is determined by the concurring action of both parasympathetically innervated constricting muscles and sympathetically innervated radial dilator muscles. The parasympathetic activity is dominant, responding to light reflexes and determine the varying pupil size baseline, the sympathetic activity is instead connected to behavioural and stress contexts and can be used as a psychophysiological parameter of cognitive activity. (Pignoni & Komandur 2019)

Task-evoked Pupillary Response

Palinko et al. (2010) estimated the driver's cognitive load from pupil size measurements finding that the pupillary response is correlated with the measured driving performances and this, as similar studies, seems to confirm the reliability of pupillometry as a measure of cognitive workload. However, the analysis is limited to a simulated task with low variability of target luminance. Palinko et al. (2010) introduced a pupillometric cognitive load measure for real-time cognitive load changes (every several seconds). (Pignoni & Komandur 2019)

Light and cognitive load effects on pupil diameter.

Palinko & Kun (2011) follows up the previous study with a proof of the possible separation between cognitive and ambient light components of pupil dilatation. The study was conducted using a driving simulator as a controlled environment for both an Aural Vigilance Task, Illumination Task (with different brightness targets) and a combined validation task. The study results show that it should be theoretically pos-

sible to model the psychophysical functions of the pupillary response over time to light stimuli and shows the measured trend over time. It also notes how the transitions bright/dark/bright are not equal as different muscle groups are involved in the contraction and dilation movements. The bright light reaction is quick "to protect the retina from overexposure", while the reaction to darkness is slower and gradual. The psychophysical function to predict an expected baseline pupil diameter should, therefore, take into account multiple parameters: current light level, previous light level, the rate of change, as well as age, and target. The study concludes that it is possible to discern the effects of luminance and cognitive load on pupil diameter and that the "proof of concept" predictor works in the limited experimental setting. (Pignoni & Komandur 2019)

Discern between mentally and visually workload.

Recarte & Nunes (2000) validated the use of pupillary response as workload index in a field scenario ignoring the effect of illumination changes as the variable was impossible to control. The data they collected shows consistent results across the different driving task (no task, verbal task and visual task) collected during multiple driving sessions, to such a degree that cannot be explained by the sole different lighting conditions. (Pignoni & Komandur 2019)

Unified formula for light-adapted pupil size.

Since the pupil diameter can be deconstructed as the result of multiple concurring factors, in order to correctly differentiate the cognitive workload from the pupillary light response, it can be useful to compute the expected pupil diameter for a given brightness condition and use the resulting value as a baseline value upon which to calculate the cognitive-driven component of the measured pupil size. (Pignoni & Komandur 2019)

In a recent paper, Watson & Yellott (2012) (NASA Ames Research Center and University of California) have reviewed seven different published psychophysical functions defining the relation between target luminance (cd/m2) and expected pupil diameter. In the same paper, they also published a newly developed unified formula. The calculated baseline would work in the range of 2 to 8 mm, the reliability of the unified formula has to be tested to ensure that the small variability of the pupil size provoked by cognitive workload is preserved (<+-1mm)(Palinko & Kun 2011). The unified formula (Watson & Yellott 2012) is valid only for a light-adapted condition with stable illuminant and point of view (PoV) as it does not account for the adaptation state or the "pupillary unrest" (low-frequency random fluctuation in the range of 0.02 to 2.0 Hz and amplitude within +-0.25 mm). (Pignoni & Komandur 2019)

Independent variables in the unified equation:

- Luminance.
- Age (The maximum pupil diameter, as well as total range, declines as the age grows).

- Field diameter (deg).
- Number of eyes stimulated, the final diameter is dependent on the number of eyes that are adapted to the light condition they defined the "Effective corneal flux density" (the variable controlling the effective pupil diameter) as dependent on the number of eyes (attenuated by a factor of 10 for one eye).

$$F = LaM(e)$$

(F)Flux density as the product of (L)luminance, (a)area, and (M(e))monocular effect.

Illuminance measurement using a digital camera.

"Luminance as a measure needed to dynamically estimate the pupil size of a subject, candela per square meter cd/m2, is the quantity of light radiating from a source. An illuminance meter is an expensive and bulky device, in more than one instance this has resulted in attempts to use a camera as cheaper and more flexible alternatives ((Hiscocks & Eng 2014) and (Wüller & Gabele 2007)). (Pignoni & Komandur 2019)

A digital sensor is at its core an array of Illuminance sensors. Each pixel measures the number of photons hitting the photoelectric surface. The presence of a Bayer filter for colour photography makes it so that to reconstruct the information from the entire visible spectrum, so multiple pixels have to be analysed at the same time. Each pixel in the final image has reconstructed values from the neighbouring pixels for all the RGB channels and in itself would be sufficient to reconstruct the illuminance of the scene, in order to reduce noise and gain reliability multiple pixels should be used, the number of pixels used is effectively the field of view of the instrument (Hiscocks & Eng 2014). The formula proposed by Hiscocks & Eng (2014) has been optimised for a DSLR camera, not all the parameters are accessible when using an embedded digital video camera. (Pignoni & Komandur 2019)

Parameters:

- Pixel value, 0-255 for an 8bit monochrome image.
- Shutter Speed; In a video camera, this is limited by the frame rate, e.g. 1/30s for a 30fps camera) and aperture or focal ratio.
- Iso or film speed.
- Camera Constant, the calibration constant for a specific camera model that has to be determined with a known instrument.

D. Wuller Wüller & Gabele (2007) suggest a different model, reversing the colour processing of the camera, first from gamma-compressed RGB to linear RGB and then from linear RGB to CIE XYZ, extrapolating then $y(\lambda)$ as the relative luminance (luminance as defined by the luminosity function, reproducing the spectral luminous efficiency of the human eye). The author notes that to access the linear response of

9

an image sensor the correct inverse gamma has to be applied and that this could deviate from the standard 2.2 (sRGB), the relative luminance can then be converted to luminance through a linear relation specific for a particular sensor/camera settings combination. (Pignoni & Komandur 2019)

3   Methodology and Implementation

3.1   Pupil Software

The Pupil Headset eye tracker has been developed alongside an open source software framework. The software is predominantly written in Python although the "heavy lifting" is done in C++. The modular software structure encourages the development of plugins to add extra functionalities in a flexible manner.

Pupil Capture

Pupil Capture is the application used to create recordings with the Pupil Headset. It collects the video streams from both the world camera and the eye camera as well as doing on-line pupil detection, gaze tracking and markers tracking in the environment (different sets of markers are used to define surfaces, areas of interest or for gaze calibration). The mobile version of Pupil Capture can run on most Android phones equipped with a USB-C connector; this version is limited to recording or streaming of the video feeds. Therefore calibration and pupil/gaze tracking have to be done off-line on the pupil player software or remotely if streaming.

In order to interpret the output data correctly from the supplied software, it is essential to consider that the different components are always running independently, at a different frequency that may or may not be in sync. The video frames from the World and eye camera are timestamped as Unix Epoch Time: floating point number that corresponds to the number of seconds since Jan 1, 1970, 00:00:00. This timestamp is then inherited by the resulting artefacts (detected pupil, gaze position, marker tracking, etc.).

Pupil Player

Pupil Player is the software needed to open and visualise the data recorded with Pupil Capture or the Pupil Mobile (Android) software. It includes tools for off-line calibration, artefacts detection and export of formatted data-sets. The data-sets are exported as a spreadsheet (.csv files) that can be easily imported in a variety of statistical analysis software.The two main files of interest are "pupil_positions.csv" and "gaze_positions.csv".

The most relevant (for this research) data saved in the "pupil_positions.csv" file:

- timestamp - Timestamp (Unix Epoch Time) of the source eye image frame.
- index - Number of the closest world video frame.
- confidence - An output value of the pupil detector, index of the confidence on the measurement. "0" indicates no confidence "1" indicates perfect confidence, "0.6" is the suggested threshold for meaningful data.
- diameter - Diameter of the pupil in pixels as observed in the eye video frame; this is the raw value, not corrected for perspective.
- diameter 3d - Diameter of the pupil available while using the "3D mode", it is scaled to mm based on an average eyeball diameter, and it is corrected for perspective. The reliability of

11

this value is questionable; the scaling can be off by several mm if the actual eye diameter is significantly different from the model and the value has proven unstable over long recording.

This file includes data from the eye detector algorithm, the role of the eye detector is to recognise the eye features from the video stream and fit a geometrical model to the image to calculate the gaze position. Other artefacts are recorded alongside the gaze direction, such as the pupil size or blinks. Two pupil detection algorithms are available, the "2D detection mode" offers high accuracy of less than one visual degrees but its very sensible to moment and misalignment of the headset. The "3D mode" uses a 3d model of the eye that is continuously updated with information from the video feed. The 3D mode can compensate for movements of the headset but has a lower accuracy of 1.5-2.5 deg. The choice between the two models has to be weighed for each experimental conditions, considering the duration of the experiment, or just how well each model is performing with the particular physiognomy of the participant's eye. The same eye-video can be re-analysed multiple times adjusting the parameters or switching between the algorithms in the Pupil Player interface.

The most relevant (for this research) data saved in the "gaze_positions.csv" file:

- timestamp - Timestamp (Unix Epoch Time) of the source world image frame
- Index - //
- Confidence - //
- norm_pos_x - x position in the world image frame in normalised coordinates
- norm_pos_y - y position in the world image frame in normalised coordinates
- base_data - "timestamp-id timestamp-id ..." of pupil data that this gaze position is computed from.

3.2   Unpack the Camera Data

The Camera

The Pupil Headset is equipped with an egocentric video camera, over time different specification were available, including stereo (3d) or the support for the Intel RealSense module. The camera on the unit used for this research is the quite simpler "high-speed camera" which still provides a variety of combinations of speed and resolutions: 1920x1080 @30fps, 1280x720 @60fps, 640x480 @120fps. The camera comes with two lenses: 60deg and 100 deg FOV. Multiple combinations were tested, the 100 deg FOV at 1920x1080 produces a considerable distortion and vignetting it was ultimately used at a lower resolution.

Libuvc and the inability to calibrate

The acquisition process is based on the open source library libuvc (Tossel n.d.) "a cross-platform library for USB video devices, built on top of libusb". It enables fine-grained control over USB video devices exporting the standard USB Video Class (UVC) interface, enabling developers to write drivers for previously unsupported devices, or access UVC devices in a generic fashion." The Pupil Capture software access the camera trough pyuvc "python bindings for libuvc" and allows access to a variety of UVC parameters from the front-end (zoom, focus, brightness and exposure).

The software in its current state (v1.11) comes with two significant limitations. The manual exposure setting is erratic, exposure time and aperture are not operated independently, being tied to a single numerical value that does not linearly relate to an exposure coefficient. A calibration of the camera was attempted but was ultimately unsuccessful (to model the response of the sensor at the different settings to be able to reconstruct a cd/m2 value from a relative RGB luminance). Sliding the exposure control setting from left to right increases the lightness in the picture as it would be expected, this apparently is changing the exposure time, but, as the exposure time reaches the maximum (at about 1/4 of the slider) another parameter, aperture, is triggered and changes by about one stop, while the exposure time turns low again. This results in a control slider that cyclically, over its length, turns the image brighter and darker. This alone meant that a calibration of the camera would have been a lengthy and imprecise process, not justifiable in light of the second shortcoming in the software. After a brief analysis of the pyuvc code, it seemed clear that the software does not provide read access to the camera settings over time, which severely limits the use of automatic exposure. Without access to a log of the exposure settings it would be impossible to differentiate between a change in luminance in the scene and automatic change in exposure settings, this would ultimately mean having to choose between calibration or automatic exposure. Fixed exposure could be sufficient in specific controlled environments but would not perform well in a field condition in which a recording would unpredictably contain scenes at significantly difference luminance resulting in under or overexposed video. Given the shortcomings of the camera/software combination, it became necessary to find an alternative source of absolute luminance data, eventually culminating in the form of an external light sensor.

3.3   Light Sensor

The TSL2591 (AMS 2013) light-to-digital converter transforms light intensity to a digital signal. It includes two photodiodes, a broadband unit (visible plus infrared) and an infrared-responding unit. The channels are measured independently. Illuminance (lux) is derived from applying an empirical formula that approximates the human eye response. The formula is not correctly documented in the data-sheet, but in brief, its purpose is to match the sensor curve, a radiometric unit to a photometric measure, as a simulation of human vision. The conversion from irradiance to illuminance is further complicated by the fact that the sensory response is monochromatic (does not provide tri-stimuli information as an RGB sensor would). The general approach is to subtract the infrared reading from the broadband reading with a coefficient. Such a coefficient has to be generalised as it should variate for each different type of light sources with different spectral compositions, the sensor characteristics are described in figure 1 and 2.

Empirical testing, comparing the calculated pupil size (based on the sensor reading) against the measured pupil size seems to indicate that the "less than perfect" matching between the human eye response and the sensor response is still sufficiently accurate for this application. The resulting Lux value can be used as an average luminance value in front of the user, dividing the Lux by the solid angle in front of the sensor in steradians (2.2 was found to be a good approximation). The unified formula for light-adapted pupil size (Watson & Yellott 2012) takes as primary input variable the characteristics of standardised visual stimuli: the diameter (an angle expressed in degrees of view of the subject) and luminance (cd/m2) of a lighted circle on

a dark background, positioned in front of the subject. This model is a simplified stimulus, and it can be challenging to map a complex scene to it. A way to do so is to utilise the sensor reading to characterise and idealised stimuli as large as the maximum binocular field of view (200 deg (w) x 135 deg (h)).

Experiential testing over a variety of outdoor and indoor conditions seems to support this method, as long as the user is presented with a diffuse luminance field. Conditions such as walking indoor or outdoor, in which most of the scene presented to the user has similar luminance or conditions in which the user has to interact with a single device, such as a laptop, and the vision can be considered as adapted to that specific source of light. The shortcomings of having a single "average" value, representing the condition in front of the user, relative to the head movement and not to gaze position, are visible in more variable conditions. Such as when the user has to switch between multiple devices/areas of interest, or in which the luminance of the device is considerably different from the average luminance in the environment (e.g. a smartphone in a dark room, or the difference between looking at the speedometer and then looking out the windshield while driving).

An attempt was made to integrate the data of the sensor with the video data and the gaze data to attenuate this issue; will be discussed in the next few pages.

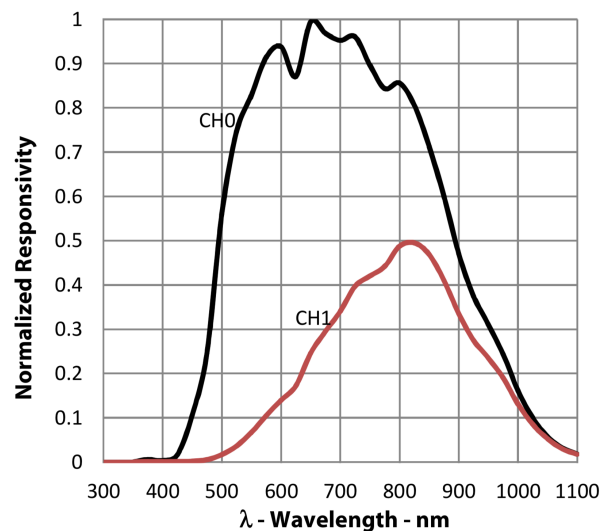Figure 1: Normalised spectral Responsivity (sensitivity to light at different wavelength) of the two channels, CH0(broadband) and CH1(infrared) (AMS 2013) of the TSL2591 light-to-digital converter.

Mounting Hardware

A small bracket was modelled to fit over the Pupil Headset camera housing to secure the sensor board to the eye tracker. It has been thought out as a temporary modification to the headset
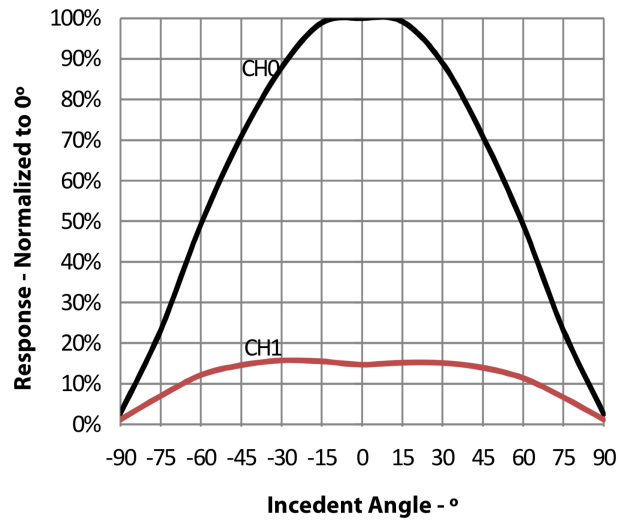
Figure 2: White light sources LED Angular Response (AMS 2013): normalised response of the two channels of the TSL2591 sensor, CH0(broadband) and CH1(infrared), to white light at different angles. Even though the response of the sensor is weighted toward the centre of the field of view, a plastic hood was added to eliminate the effect of incident light coming from the sides of the sensor. This was in clear contrast to the initial attempt to flatten the angular response through the use of a diffuser.

and can be assembled in minutes. A series of clips serve as cable management, securing the connection between the sensor and the data logger. The sensor is positioned close to the camera so that the FOV of the two is overlapping as much as possible. A plastic hood and cover protect the sensor as well as limiting the FOV to about +-50 degrees. The model can be printed or downloaded for free and then printed with a generic filament printer from Shapeways.

3.4   On-line Camera Calibration

A small introduction on what makes the video analysis possible.

Open CV

OpenCV (Open Source Computer Vision) (OpenCV 2019) is an open source computer vision and machine learning software library. The library includes a set of computer vision and machine learning algorithms as well as a simple module for colour conversion, image processing, video extraction and much more. It has C++, Python, Java and MATLAB interfaces, for this project it was used with Python 3 on Mac OS. OpenCV is used to open the video recordings from a Pupil Capture archive, extract the frames and work on the frames. This includes blurring, colour conversion, recognition of areas of interest as visible in figure 4.
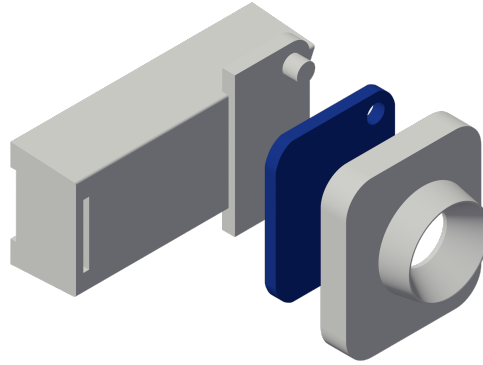
Figure 3: The mounting bracket and plastic cover 3d models, in scale with the sensor board(blue). The sensor board is fastened to the braked by melting the two plastic prongs, the plastic cover snaps on the sensor PCB.

Relative Luminance

Relative luminance is related to the photometric definition of luminance but expressed in normalised value (0-1) from a reference black to a reference white. It is an expression of the luminous flux density in a particular direction (radiometric measure), weighted by the luminosity function of the CIE Standard Observer. The relative luminance is a is calculated from the sRGB values that are first linearised (as per sRGB specification) from gamma-compressed RGB (standard $gamma$ in the experiment was $= 2.2$) and then fed into the luminosity function eq. 3.1, (W3C 2018), (Pignoni 2018).

-Relative Lumiance equation.

$$rLv = 0.2126 * R' + 0.7152 * G' + 0.0722 * B' \tag{3.1}$$

The data collected by the onboard video camera can be analysed following three main approaches:

- Apriori calibration, done once, to map the response of the sensor, so that, given fine control over the exposure, it would allow for the use of the camera as a luminance sensor. For an experimentally defined camera coefficient Kc, Luminance Cd/m2 Ls can be calculated from a relative luminance value Lr (normalised to 0-1) given that all the exposure value are available: fs Aperture (F-Stop), S ISO Setting and t Shutter Speed (seconds). The equation is adapted from (Hiscocks & Eng 2014).

$$Ls = (Lr * fs^2)/(Kc * t^S) \tag{3.2}$$

This approach has been used before with dSRL and would be the ideal solution for a self-contained, camera-based luminance sensor, unfortunately, with the current software/hardware, it is unpractical to obtain fine control over the exposure settings. Moreover using compressed and processed video could reduce the precision of the conversion significantly as details are lost (e.g.

16

aggressive compression applied to dark/bright reduces the usable dynamic range. Also, colour correction curves and gamma can alter the linear relation between the relative luminance and effective luminance over the spectrum).

- An empirical calibration that tries to match the response of the camera to the measured pupil size (e.g. for a given pupil size, reconstruct the expected luminance) and find Kc.

$$Ls = Lr/Kc \tag{3.3}$$

This approach was the first practical attempt to measure luminance, before the addition of the external sensor and after the realising that the first approach was not applicable in this case. The main limitation of an empirical calibration is that the measured pupil is, in reality, a relative measure. The unified formula for adapted pupil size (Watson & Yellott 2012) could be used to convert a measure pupil size in mm to a luminance value and, ideally, this luminance value could then be used to find the camera coefficient Kc comparing the camera relative luminance and the pupil-derived luminance value. This is unfortunately not possible as the native measure of pupil size is in pixels, and the relation pixels/mm is variable as it depends on the distance of the pupil camera from the eye (which changes between subjects or otherwise every time the eye tracker is set up). The pupil capture software attempts to do such conversion fitting an averagely sized eyeball on the eye video frames, but, being an average model, it is not reliable. The effort becomes then to obtain an absolute measure from two relative measures, which is further limited by the non-linear response of the (Watson & Yellott 2012) formula. The video processing is visualised in the figure 4 section(a).

- An on-line calibration is a final attempt made to use the video data. It relies on the external sensor and it is based on the assumption that the sensor has a similar FOV compared to the onboard camera. The sensor reading can, in fact, be used as a reference value for any point on the video frame as the average luminance measured by the sensor is the photometric measure resulting in the average relative luminance in the video frames. (e.g. for a measured luminance of x the resulting average RGB relative luminance is y (figure 4 section(a)) therefore the relative luminance in a point yP is equal to the effective luminance xP).

The sensor reading is converted to Cd/m^2 by dividing it by the solid angle (steradians, FOV of the sensor).

$$aLum = lux/2.2 \tag{3.4}$$

The maximum and minimum luminance (corresponding to RGB black and white) is calculated by dividing the sensor luminance $aLum$ by the camera average relative luminance $cameraALum$ (1/1000 is added to avoid division by zero).

$$minLum = aLum/(cameraALum * 1000 + 1) \tag{3.5}$$

$$maxLum = minLum * 1000 \tag{3.6}$$

For a given relative luminance extracted from the video (not the average luminance, a luminance specific to only a portion of the frame) $cameraSLum$ it is now possible to calculate the effective luminance with a linear interpolation between the maximum and minimum luminance value using the spot relative luminance (0-1) as weight.

$$Lum = ((maxLum * cameraSLum) + (minLum * (1 - cameraSLum)))/2 \qquad (3.7)$$

Luminance on the Gaze Point

The figure 4 section(b) and (c) are describing two analysis that can be applied on the video and gaze data to obtain a more accurate measure of luminance for the actual FOV of the subject. The average luminance measured by the sensor is bound to the head movement and not to the eye movement, as the gaze wanders away from the centre of the camera FOV the measure becomes less accurate. Similar bias can be introduced by the use of screen or otherwise areas of interest characterised by a significantly different luminance than the average luminance in the scene (e.g. a smartphone in a dark room). The gaze and video data can be combined to measure luminance in a specific spot in the scene. The figure 4 section(b) describes a method designed to measure the luminance surrounding the gaze (the Fovea, the area responsible for the sharp central vision of fewer than twenty degrees Visual field). Although this approach serves a significant improvement over the use of the sole average luminance it can be too sensitive to high contrast objects (e.g. in the case of text on a screen, the black and white components are merged into an average grey, an underestimation of the adapted luminance).

The figure 4 section(c) describes a method designed to measure the luminance of an area of interest (AOI) around the gaze point. The area of interest is defined as the portion of video surrounding the gaze point characterised by similar chroma and luminance; empirical testing has proven that this method correctly isolates high contrast AOI such as windows and screen and it is possibly the best approach among those that were explored.

A significant source of error that was not tackled in the research is vignetting, the onboard camera has significant vignetting (as well as the external lux sensor 2) if uncorrected this inevitably results in a slightly reduced average luminance for the entire scene as well as reduced measured luminance as the gaze moves toward the corners of the frame.

3.4.1 Data Logger Feather and Battery

The Sensor module was bought as part of a breakout board from Adafruit Industries (Industries 2019), the board is designed to communicate over I2C, and it is provided with an Arduino compatible library. The use of the sensor can be divided in two scenarios: with a computer or with the Android phone. With a computer it is arguably better to directly log the sensor data through serial from an Arduino, this provides data that can be timestamped with the same reference as the pupil data as they are running on the same machine. This is not possible while recording eye tracking data with the smartphone software; for this, an external data logger was a better option. Adafruit Industries produces a line of boards and accessories called Feather. Two Feather M4 Express were bought as well as two Adafruit Adalogger FeatherWing (Industries 2019) (see figure 5). The M4 Express is an Arduino compatible board, powered by a 120MHz Cortex M4 ATSAMD51J19 makes it very capable and fast. It includes a charging circuit for a 3.7 lithium battery, making it portable without the need of any extra hardware. The Adalogger board comprises of a Real Time Clock (RTC) PCF8523 with a backup battery to keep the clock update when disconnected from power as well as a micro SD card slot. This combination can be used to read the luminance sensor data, timestamp it and save it on the SD card all with one device.

a) Relative Luminance

Original image.

The luminosity function is applied to the scene to obtain a map of the relative luminance.

Average relative luminance of the entire scene, it does not include information on the actual cd/m2 as the exposure data is not accessible. It can be empirically reconstructed with a coefficient that would scale the plot to correctly match the pupil size, (only valid for fixed exposure).

b) Luminance on gaze

Original image with gaze position.

Gaussian blur is applied to the scene to obtain the average colour around the gaze position.

The average relative luminance (a) is used alongside the external sensor data to convert relative luminance in luminance. The on-line calibration of the camera data is based on the assumption that the FOV of the camera is roughly the same as the FOV of the sensor. Therefore the average luminance measured by the sensor, the photometric measure resulting in the average relative luminance of the video frames. Based on this assumption a simple interpolation is used to convert the relative luminance on the gaze point to a cd/m2 value.

c) Luminance around gaze

Original image with gaze position.

The average luminance is calculated only on the selected area and scaled as in (b).
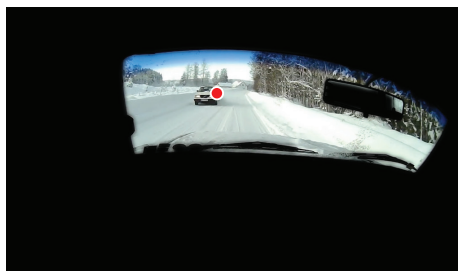
The portion of video around the gaze point is selected through a "GrabCut" algorithm. It selects the continuous area with simial chroma/luminance around the gaze pixel. This is done in order to isolate the surface/area of adapted vision around the gaze such as a window or a screen. The tolerance of the grabcut algorithm is dynamically set to one standard deviation (RGB values) of the entire frame.

Figure 4: Schematic description of the different attempts to extract useful luminance information from the video frame.
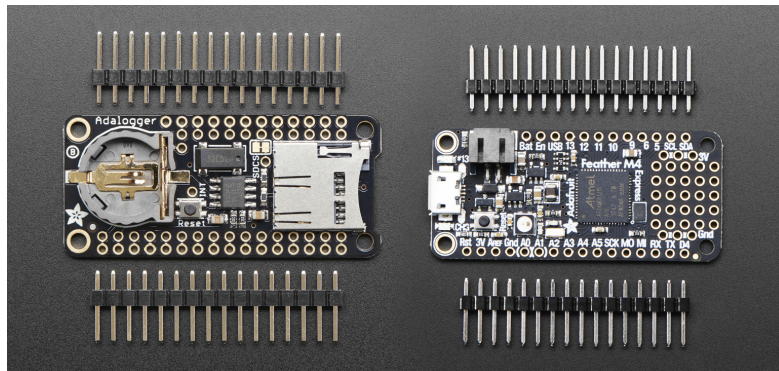
19

Figure 5: Adafruit Adalogger FeatherWing and Feather M4 Express (Industries 2019).

The RTC clock is inertly imprecise and drifts over time (as much as several seconds per day can be expected depending on temperature). This results in a delta between the timestamps of the Eye tracking data, recorded either with the smartphones or a computer and the timestamps of the light data, recorded on the external data logger. The two sources have to consequently be re-synchronised. Given the small number of recording this procedure is still acceptable even though it involves manual fine-tuning of the drift value by measuring the difference between features (peaks) in the measured pupil size and the calculated pupil size.

3.5 Code Structure

The structure of the code is represented in figure 6.

3.5.1 Implementation of the Unified Formula

The unified formula (Watson & Yellott 2012) for adapted pupil size implies a very precise stimulus, a bright circle on a dark background. The size of the circle determines the degrees of field of view (degrees) and with the luminance (cd/m2) makes up the core of the equation, defining the corneal flux density (i.e. the product of luminance and subtended area) as defined by Stanley (1995): $D = 7.75 - 5.75[(F/846)0.41/((F/846)0.41 + 2)]$ "where D is the pupil diameter (mm), and F is the corneal flux density (cdm-2deg2)".

The model implies that the pupil control mechanism reacts as a 'flux integrator', following an S-shaped curve.

As mentioned in the previous sections the standardised stimuli has a very limited application, to simplify the application the field of view angle is kept constant at 167 degrees and the man variable is the average luminance in front of the user.

A "R" compatible version of the formulas is available as part of the CVD: Color Vision Deficiencies (Gama et al. n.d.) package, this version was adapted to work in python an included in the code.

3.5.2 Filtering

The Pull diameter recording produced by the Pupil eye tracker is a measurement taken from the analysis of the frames coming from the eye camera. Sampling frequency and resolution depend
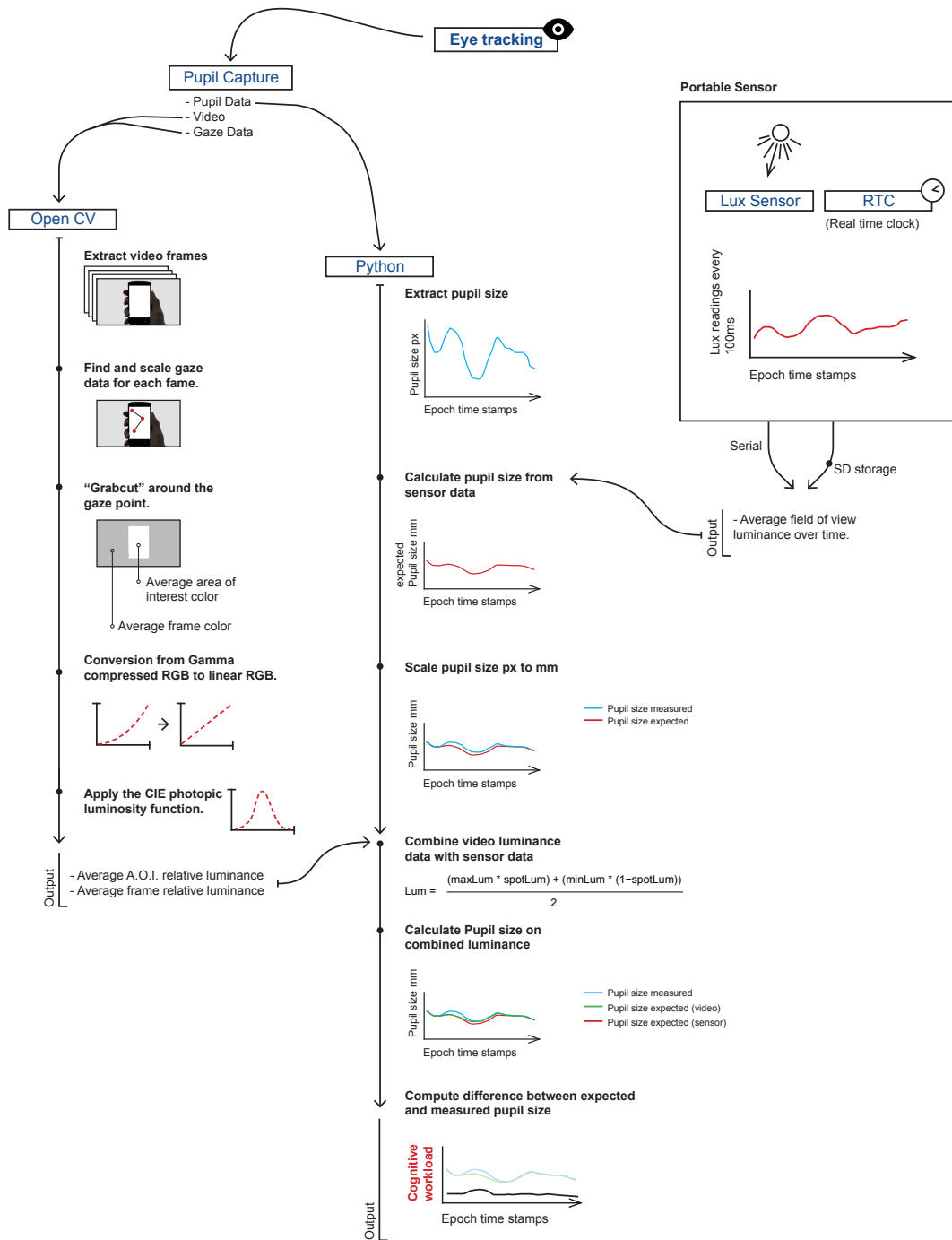
20

Figure 6: A simplified map of how the data is being processed in order to estimate the cognitive workload.

on the capabilities of the camera itself. The compromise that was chosen for most of the recording done for this thesis is 400x400px resolution at 120hz; the maximum sampling would be 200hz at a reduced resolution. The frequency of 120hz would imply a folding frequency (Nyquist frequency) of around 60hz as the highest signal frequency that can be recorded and recreated with the camera system. Unfortunately, the filtering required to clean the pupil diameter data lowers considerably the final sample rate for the "Cognitive workload" metric. The resolution of the pupil depends on framing of the eye as well as gaze direction, in indoor conditions it usually averages around 60px. The measured pupil size includes noise from multiple sources:

- Movement of the camera (drift) on the head.
- Hippus, spasmodic, rhythmic, dilatation and contraction of the pupillary muscles.
- Incorrect recognition and removal of blinks.
- Artefacts created by the recognition algorithm that produces the 3D value (corrected for perspective).
- Artefacts as an intrinsic effect of the eye movement in the 2D (raw pixel) value as the interaction between the variable gaze angle and the fixed camera angle.

The hippus can be easily removed as it is a relatively high-frequency component (~2.5 Hz) compared to the rest of the signal. An initial attempt with a simple Butterworth filter such as the one included in the scipy.signal package (community 2014) showed that a 1hz cutoff frequency was only sufficient to remove a low portion of the noise, and that lowering further the frequency to remove other artefacts (such as blinks) would reduce considerably the shape and height of the waveform, introducing distortion in the measure of pupil size and most importantly reducing the dynamic range of the signal. An alternative to the more traditional low pass filter is a family of filters often referred to as Savitzky-Golay. In the original paper (Savitzky & Golay 1964), the authors proposed an alternative method for the smoothing of constant frequency digital data points based on "local least-squares polynomial approximation" (Schafer 2011). Savitzky and Golay initially developed this methodology to smooth noisy chemical spectrum analyser data, demonstrating how least-squares smoothing reduced noise while preserving the shape/height of the waveform peaks. This makes a Savitzky–Golay filter ideal for the filtering of pupil data for two main reasons: it preserves the full dynamic range (2 to 8mm of the pupil dilatation) and has zero phases; therefore the features (peaks) of the signal are not shifted in the time domain. Not having a phase shift, avoid the de-synchronisation of the pupil signal against the video or luminance sensor data. The scipy.signal package includes the Savitzky–Golay filter that was used in the project (community 2015).

3.5.3 Scaling

As mentioned before, the measured pupil size is expressed in pixel and should be converted to mm in order to compare the measured pupil size to the calculated pupil size, output of the unified formula. Unfortunately, it is impractical to get a coefficient of the pixel density (pixel per millimetre ppmm) as it varies each time the camera is set up, being dependent upon the distance and angle between the eye camera and the eye. An empirical method to obtain the ppmm for a given recording is to place a reference object in the frame of the camera (figure 9), this procedure is unfortunately invasive as the object should be placed firmly against the participant's eyelid.
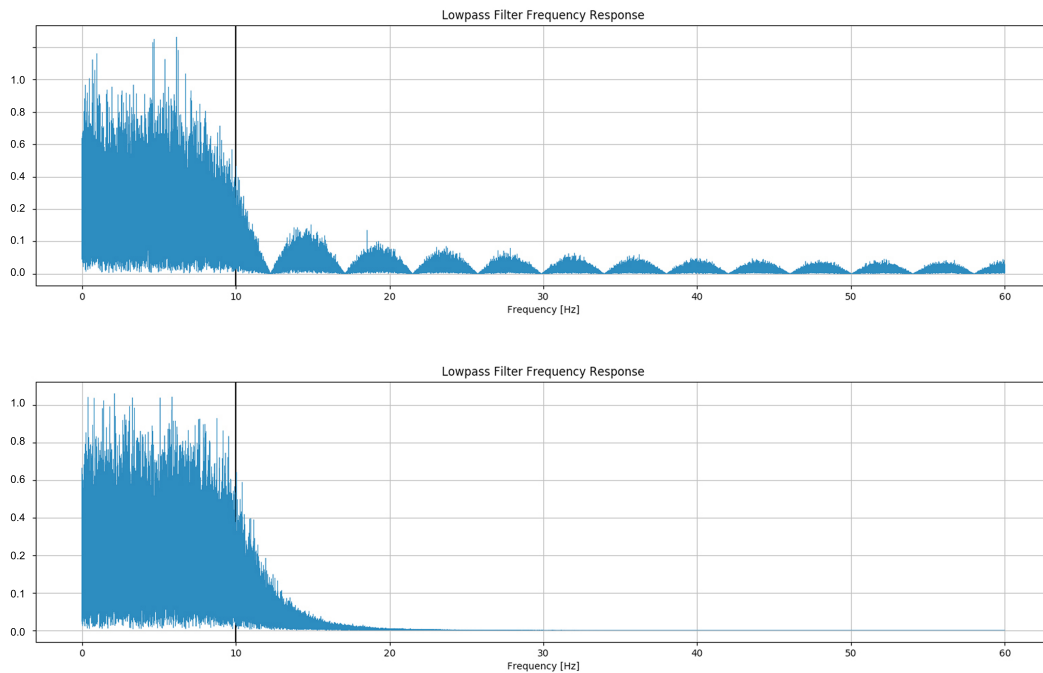
Figure 7: Example of the frequency responses of a six order Butterworth filter (bottom) and a Savitzky–Golay filter (top), both with a cutoff frequency of 10hz (applied on white noise sampled at 120hz). The resulting amplitude of the signal (normalised) after the application of the filters is visible on the vertical scale as it changes for a range of frequencies (0-60Hz) on the horizontal scale.

An alternative approach, that does not require physical interaction with the participant, is to consider the calculated pupil size as a reference measure of average pupil size in a given moment, or in other words as the pupil size that is to be expected with the average cognitive workload in a given task. Once this assumption is made, it is possible to scale the measured pupil size so that the mean of the measured pupil size will match the mean of the estimated pupil size. The mean value of both the expected and measured pupil sizes are calculated and used to proportionally scale the measured pupil size. The result of the two different paths is visible in figure 10. Clearly, there is a difference between the two plots; how well the estimate pupil size value will fit the definition of pupil size at a mean cognitive workload will depend on how different the mean workload of the task is in comparison to the underlining workload that originally affected the unified model data. It is therefore difficult to express workload as an absolute measure (e.g. Claim that "a 1mm dilatation corresponds to a specific workload"), it would be otherwise possible to observe relative changes in the workload and observe that a certain dilatation over or under the expected pupil size represents a significant change in the workload. At this stage, subjective data are still necessary to interpret and classify a change in workload.
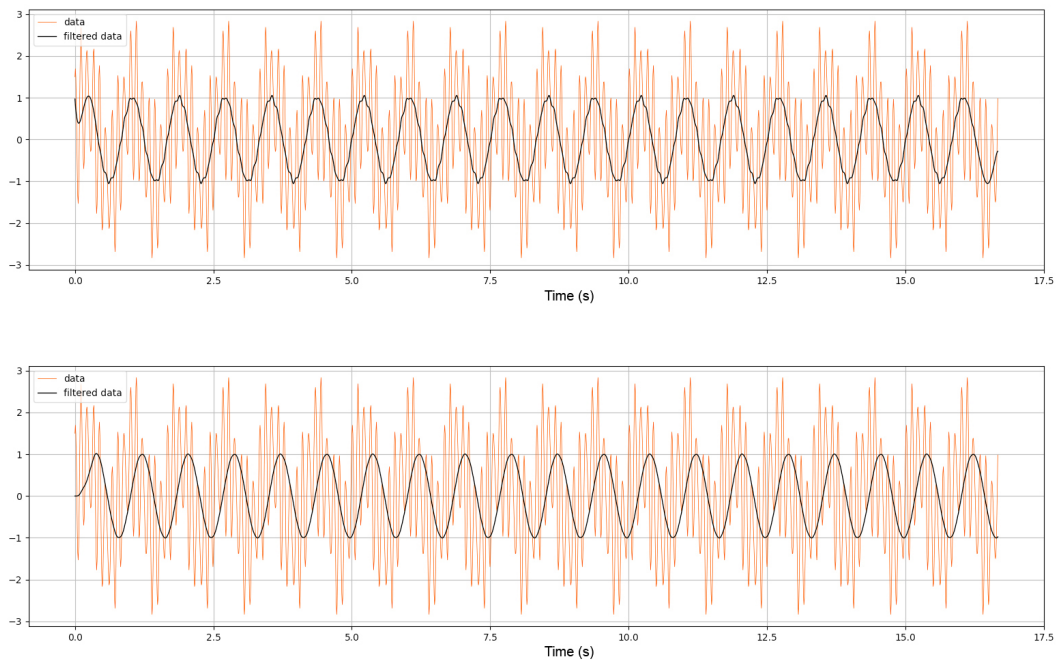
Figure 8: Example of two different filtering techniques: six order Butterworth low-pass filter (bottom) and a Savitzky–Golay filter (top). The resulting wave (black line) is the "recovered" 1.2 Hz sinusoidal extracted from a noisy signal (orange). It is visible how the Savitzky–Golay filter is removing less high frequency noise form the output (ripples in the wave) but correctly preserves the phase and amplitude of the sinusoidal. In comparison, significant phase shift is visible in the signal filtered trough the Butterworth filter (bottom). The amplitude (vertical scale) is expressed as normalised values and is the results of the sum of two (-1 +1) waves.

3.5.4 Distance

Concluding the chapter, a note on the way the output, representative of the cognitive workload, is being calculated. Once the expected pupil size is extracted, either from the luminance sensor or the video data and the pupil size has been scaled down to millimetres, it is necessary to quantify the difference between the two. Initial experimentation involved Euclidean distance or the more refined Dynamic time warping (DTW) as a technique to find an optimal alignment between the two time-dependent sequences (*Dynamic Time Warping* 2007). Dtw and in general measures of distance are very helpful in quantifying the difference between two data sets, and were used to evaluate and improve the use of the Unified Formula (e.g. how close pupil size calculated by the algorithm to the actual pupil size) but are not necessarily useful to extrapolate the cognitive workload effect on the pupil size. First of all, a measure of distance will tend to zero when the data-sets are overlapping, a deviation, be it positive or negative will result in a higher distance. If the expected pupil size is considered as the pupil size at the average cognitive workload, the resulting workload is expected to be a relative number oscillating around this average, as a negative number for workload below average and as a positive number for above
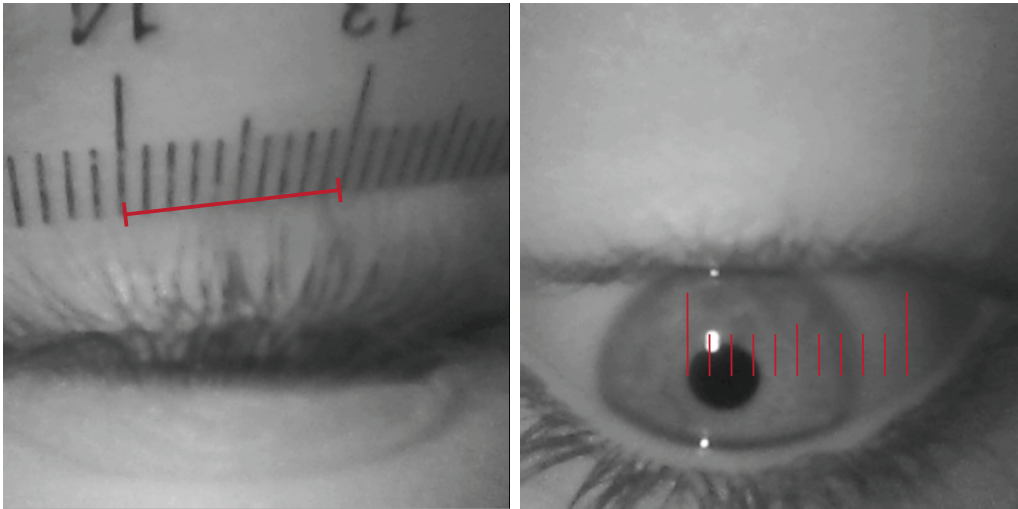
Figure 9: In order to calculate the ppmm of the camera it is possible to place an object with a known dimension in the frame, in this case, one centimetre on the ruler measured 167.6 px in the frame estimating a 16.7 ppmm pixel density.

average workload; therefore the sign has to be preserved. The measure of workload used for the two experiments is a simple subtraction between the two data sets, the expected pupil size is removed from the measured pupil size, in order to reduce the noise in the process a window can be defined so that the mean over a number of samples is used instead, the dimension of the window determines the frequency of the output data-set. The output data remains in mm.
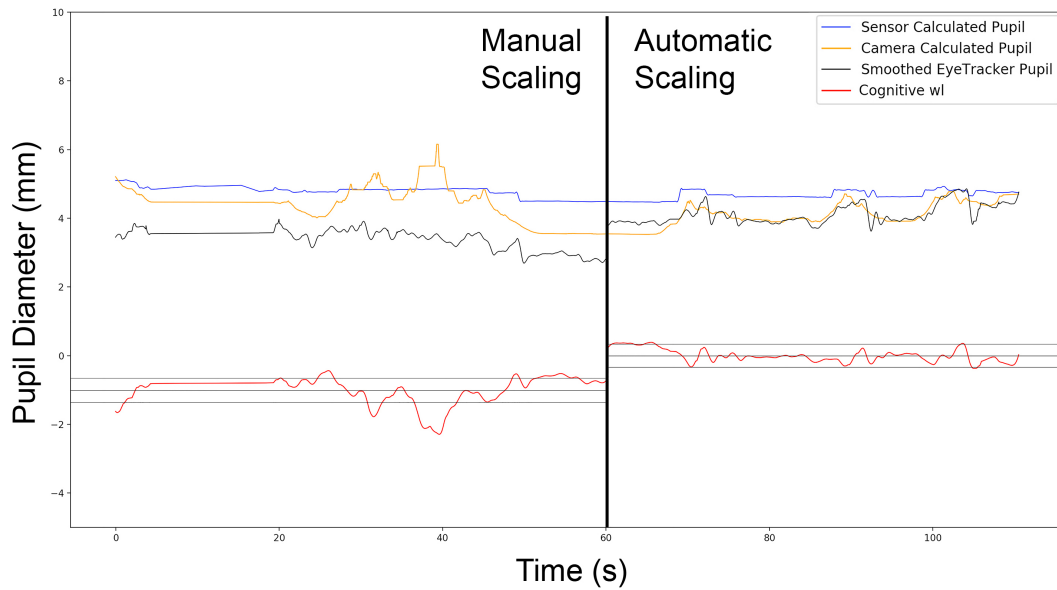
Figure 10: Examples of the output of the cognitive workload: the black line is pupil diameter over time (millimetres on the vertical scale) as measured by the eye tracker, yellow and blue lines are the expected pupil size as calculated by the system either using the luminance sensor alone or the luminance sensor and the camera. Red is the output value representing the cognitive workload induced pupil dilatation. As the measured pupil size is originally expressed in pixels it has to be scaled to millimetres. The manual scaling of the measured pupil value (left), is based on a manual measure of the camera resolution of 16.7 ppmm. The automatic scaling estimated a lower 12,7 ppmm density. The difference between the two results in a shift in the output (red line, representing the difference between estimated and calculated pupil size) and therefore a change in the absolute value of the output data, the relative value, visible looking at the position of the minus and plus one standard deviations marks (black horizontal lines) is not significantly affected.

4 Laboratory Validation

4.1 Introduction

The session in a controlled environment was organised to evaluate the performance of the eye tracker and psychophysical formula in a controlled manner.

Three main conditions are of interest for this scope:

- Variable cognitive workload with fixed visual stimuli (brightness).
- Fixed (rest) cognitive workload with variable visual stimuli.
- Variable cognitive workload with variable visual stimuli.

As highlighted in the literature review, different levels of cognitive workload can be induced by having the participants perform a variety of tasks. The choice, in this case, was guided by the need to separate the visual stimuli from the task and control them independently. This leads to the selection of four mental tasks of increasing complexity, see table 1. All the information on the task were given by voice to the participants in a short briefing session, during the experiment, just before they had to perform each task. Only little information was given to them in advance, to avoid them preparing for the task itself during the "rest" phase. The main instruction was to try to focus on the centre of the screen and keep their eyes open throughout the experiment (they were allowed to bilk). The main experiment involved variable visual stimuli, while the control experiment kept the visual stimuli fixed. The fixed (control) condition required fixating a spot in a neutral and uniform (grey) background, for the duration of the experiment, the stimuli chroma/luminance and ambient illuminance are stable. In the variable condition, the background changes luminance following a sinusoidal wave at 0.05Hz (e.g. 10s to shift from black to white). The change is obtained modifying the RGB value of the background from black to white. The range of luminance is therefore limited by the maximum values that can be produced by the mean used to create the stimuli, in this case, a projector as described in figure 13. In the same figure, it is visible how the head-mounted sensor and spectroradiometer present diverging (but linearly related) measurements; this is most likely the consequence of the two different fields of view that are being measured by the different instruments. The spectroradiometer measures a narrow field of view in the centre of the projection (two degrees), the brightest part of the projection; the head-mounted sensor instead measures a much wider filed (more than sixty degrees) that includes darker areas. Nevertheless, this measure can be used as a simple calibration as the relation between the two is linear.

Rest (baseline)	1 min
Briefing	~10sec
Count up 0 to 60	~30 sec
Recover	1 min
Briefing	~10sec
Count down 60 to 0	~30 sec
Recover	1 min
Briefing	~10sec
Count down 91 to 0 every 4	~1.5 min
Recover	1 min
Briefing	~30sec
Fibonacci sequence to over 100	~1.5 min

Table 1: Task sequence.

4.2   Methods

4.2.1   Participants and Methodology

Twenty-one (twelve females) students, with a mean age of 29, from the NTNU campus in Gjøvik, took part in the experiment. The participants were selected with a simple convenience sampling. It was deemed to be accepted as the test does not aim to measure the psychophysical response to an induced cognitive workload in the tested population. The scope of the test is to evaluate the performance of a tool in measuring a well known and documented psychophysical relation, as such, the subject of the research is the tool itself and not the participants. The research question for this phase is in fact: "Is the system, composed by the eye-tracker and baseline algorithm, capable of measuring the cognitive workload reliably independently to the visual stimuli?".

The independent variables are:

- Induced cognitive workload (task difficulty).
- Stimuli luminance cd/m2.

The dependent variables are:

- Measured pupil size (dependent on cognitive workload, stimuli luminance precision of the instrument, hippus).
- Calculated pupil size (dependent on the stimuli luminance).
- Calculated cognitive workload (dependent on Calculated pupil size and Measured pupil size).

Choice of Visual Stimuli

During the design of the experiment, it was decided not to repeat the same workload measurement (multiple light conditions) with the same subject. The memory effect on both the "count down every four" and "Fibonacci sequence" tasks could very well skew the data as the task difficulty reduces with repetition. The limited time available to perform the experiment required a

selection of the light conditions to be tested. The fixed "light" condition was therefore narrowed down to only one light level as there is no indication in the literature that different "stable" luminous stimuli can affect the cognitive workload. It can otherwise have an effect on how the workload is manifested through the pupil size, as the pupil has a physically limited range to dilate and contract. To model, the response of the pupil size in function of cognitive workload and luminance levels would require a large number of observations that were not achievable at the time. It was, therefore, deemed of interest to verify that the system is able to correctly track cognitive workload in a light condition. Considering the aforementioned limitation, the obtained value of "change in measured diameter" cannot be used as an absolute value of the cognitive workload. (e.g. it will not be possible to generalise that dilatation of x mm corresponds to a certain level of workload in any but the measured light condition). However, it will be possible to claim that a dilation (or contraction) corresponds to a change in cognitive workload in a broad range of light conditions. The variable light stimuli were selected with a similar criterion to find one relevant condition of interest, it does not necessarily constitute an extensive benchmark of the system, but it is thought out to be a useful index of how it can perform in evaluating the cognitive workload over several seconds. A high-frequency index of cognitive workload would require the model of pupillary response to light to consider also the "transitional states" (the response of the pupil as it transitions between two adapted pupil sizes). As this is not part of the adapted pupil size model that is being tested, high frequency (sudden) changes in light condition will introduce noise as the measured pupil size will lag behind the expected size. This effect can be partially removed by filtering the data at the cost of temporal resolution.

- An highly variable light condition (within the technical constraints of the adopted medium).
- Repeatable for each task/rest session.
- Slow enough change of light not to introduce (excessive) discomfort in the participants (flashing light)
- Slow enough change to allow for the eye to adapt to the light state as it changes.

A sinusoidal sequence was chosen as the carrier of the visual stimuli (0,05Hz or ten seconds between full black and full white). A sinusoidal fits well within the list of criteria and allows for a parallel method of removing the effect of light on pupil size. The spectrum of a simple sine wave has only energy at one frequency (simple signal, visible in figure 11), it is therefore possible to remove it, or its direct effects, from a signal with a narrow band-stop filter. Other periodic waveforms (e.g. saw-tooth, square) would require a more complex filter (affecting a wider frequency range) that could potentially mask other components of the "pupil size" signal. The "filtered" pupil size signal should correspond exactly in shape to the result of the model being tested, as all the effect of the light change on pupil size should occur with a frequency close to the one of the sinusoidal wave. This should remain true also considering the aforementioned distortion (transitional state) as long as the frequency of the stimuli is low enough. This allows to achieve the same result with two independent methods, one frequency independent, the system being tested and the other, frequency dependent, that is only applicable to this particular condition (known frequency of the luminous stimuli).
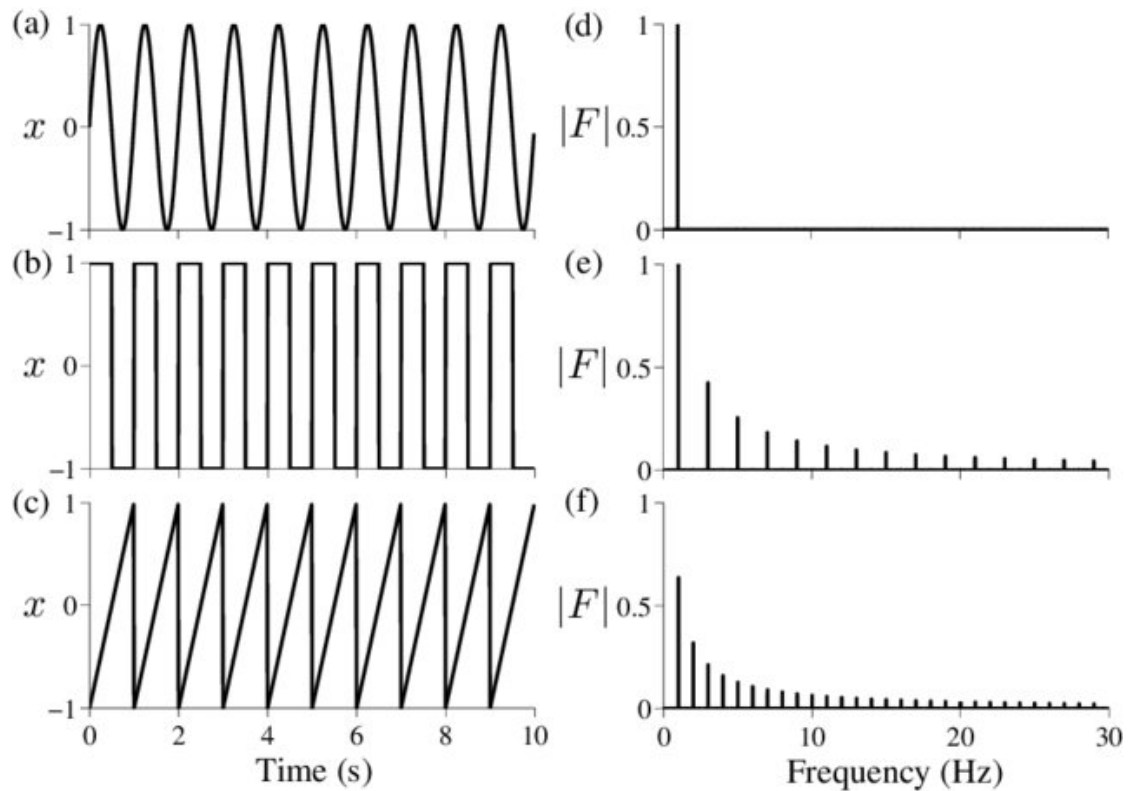
Figure 11: Time series of three differently shaped periodic functions: (a) A sine wave; (c) A square wave; (e) A saw-tooth wave; and the respective amplitudes of the one-sided Fourier transforms (b), (d) and (f) (Clemson & Stefanovska 2014).

4.2.2 Apparatus

The Pupil Pro eye tracker (Kassner et al. 2014) has been used for the infrared pupil detection, the pupil camera was set at 400x400px @120 Hz while the world video was recording at 1280x720 @60fps. The average luminance in front of the participants was logged at 10Hz using the tsl2591 Lux sensor (AMS 2013). The experiment used the open source software package "PsychoPy" (Peirce 2007), (Peirce 2018) for neuroscience and experimental psychology. The PsychoPy requires the definition of screen resolution, dimension and viewing distance in order to correctly render the stimuli elements independently to the device used allowing the use of an absolute measure of visual dimension (degrees of viewing angle). The chosen projector, EPSON EB-1776W, was mounted on the ceiling and projected on a white paper backdrop from a distance of 230cm. The measured brightness (luminance) in this setup was 105 cd/m^2 and the measured ambient contrast ratio (ACR)(Chen et al. 2017) was 160:1. The setup was measured using a Konica Minolta CS-2000 Spectroradiometer (Ltd 2018) at the viewing distance of 130 cm. The projector was the only light source in the room. The sitting position is described in figure 12
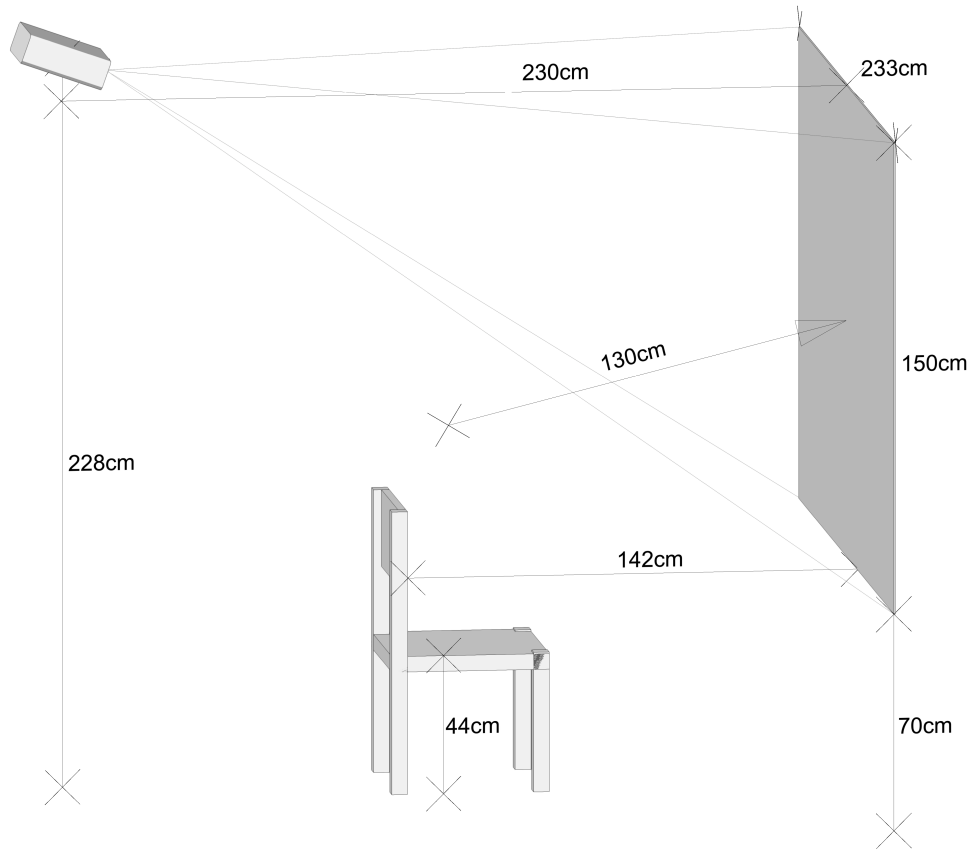
30

Figure 12: The experimental setup and sitting position.

RGB	cd/m2 (1 deg)	cd/m2 (sensor)
-1.0	0.64	0.35
-0.8	1.55	0.84
-0.6	3.64	2.2
-0.4	7.39	4.1
-0.2	12.7	7.1
0.0	20.27	11.28
0.2	29.31	16.44
0.4	42.06	23.82
0.6	58.40	33.92
0.8	77.9	44.52
1.0	105.65	62.32

Table 2: Projector measurement.

4.2.3 Procedure

The experiment was set up in one of the rooms available in the Norwegian Colour and Visual Computing Laboratory in Gjovik. The room selected has no windows and controlled illumination (kept dark for this experiment). The test sessions were arranged in advance in slots of forty-five minutes while the actual experiment required only twenty minutes. The participants were greeted with an introduction to the scope of the experiment, how the data was going to be processed and the signature of the informed consent. The actual content of the tasks was intentionally kept vague and described as mathematical sequences intended to generate different levels of cognitive workload, as such the result was not of interest and it was made clear that the performance of the participants was not tested. It was also specified that, during the "rest" periods, the participant should have tried to relax and avoid thinking back to the tasks. The participants were seated in front of the projection screen, as seen in figure 12 and were asked to wear the eye tracker. The test started with the calibration of the eye tracker using on-screen markers in the Pupil Capture software. Once that was completed successfully, the test software was initialised in Psycopy. The Psycopy script included a second "calibration" session in which the user was asked to follow a marker on the screen to verify the correct gaze tracking. The test starts with a sixty seconds "rest", then the briefing (voice instruction regarding the tasks) for the first task, followed by the task as per the sequence described in table 1. The Psycopy script took care of saving timestamps for the start and end of each step. The debriefing included a more accurate explanation of the project and, for those interested, a plot of the participant data.

4.2.4 Data Analysis

The processing of the data followed the scheme presented in the previous chapter (figure 6), only data from the luminance sensor was used, the onboard video feed and gaze position were not used. The visual output of the Python code is visible in figure 14. The measured pupil size (black) is compared to the expected pupil size (blue) to compute a distance between the two (red) that represent the cognitive workload, plus noise. As noted before this process attempts to remove the effect of light from the pupil size. The output(red line) is what is then saved in a separate file, timestamped and tagged for the different steps in the sequence for further data analysis. The

figure 15 shows an example of the verification process, that checks for the correct removal of the effect of light (sinusoidal wave pulsating at 0.05Hz) through the main algorithm (red line), a band-pass filter is applied to the measured pupil size suppressing a narrow band of frequencies around 0.05Hz, the result is visible as the orange line. This procedure should correctly remove the effect of light, but it is only applicable to regular pulsations such as the artificial condition of the experiment.

The data has been imported in IBM SPSS (IBM 2017) for further analysis. The data was aggregated as a single file with the following header:

- Participant number: code assigned to the participants.
- Age.
- Light Condition: Variable or Fixed.
- Task: rest, count up, count down, count down four, Fibonacci.
- Distance: the measure of Cognitive workload as defined in the previous chapter obtained by removing the expected pupil size from the measured pupil size, expressed in mm.
- Timestamp: the distance was computed as a rolling average every 100ms; this is the resulting timestamp for each sample.

Each row is a sample; each participant has several samples. From this data, a new database was

produced computing the mean distance and standard deviation for each task, now intended as a sequence of measures, all the rest sessions were aggregated.

- Participant number: code assigned to the participants.
- Light Condition: Variable or Fixed.
- Rest mean value.
- Count Up mean value.
- Count Down mean value.
- Count Every four mean value.
- Fibonacci mean value.

Each row is a participant.

The primary statistical technique used to process the data is a General Linear Model (GLM) Repeated Measures. The GLM Repeated Measures procedure provides analysis of variance of the results of a repeated (same) measurement for each subject; in this case, the within-subjects factor is the measure of cognitive workload (distance). The variable or fixed light condition is used as the between-subjects factor dividing the population into two groups. The imbalance between the number of samples for the two conditions and a relatively low number (six) of participants for the control (fixed light) experiment could reduce the effectiveness of the analysis for the control group. Null hypotheses have been tested for both the between-subjects factors and the within-subjects factors.

4.3 Results

The processing on eye tracking and luminance data produces for each participant two data-sets (.csv files), the first includes the measured (scaled) and the calculated pupil size alongside timestamps and a variety of accessory information, the second includes the measure of "workload" or the distance between the measured and expected pupil size, again accompanied with tags and timestamps. The figure 14 shows the visualisation of these files for two participants; the red line represents the cognitive workload or distance, accompanied by three lines, marking plus and minus one standard deviation. The blue line is the measured pupil size, scaled to fit over the expected pupil size (blue line). The background colour is used to highlight the different steps in the experiment: light blue for rest, red for briefing.

IBM SPSS (IBM 2017) has been used to produce the rest of the analysis. The independent sample t-test is used to check the equality of variance and equality of means between the control group and experiment group, none of the combinations in figure 16 produced a significant result, proving that the groups can indeed be treated as one and that the residual effect of illumination in the measure of cognitive workload can be considered as non significant noise.

The General linear model (GML), again produced through IBM SPSS (IBM 2017), is used to evaluate significance in the effects of the two independent variables and their interaction (Task, Light Condition and Task*Light Condition). The output (17 and 18), shows that the within-subject factor "Task" has a significant effect on the dependent variable of Measured Workload (Pupil dilatation mm). It also shows that the interaction between the independent variables does not have a significant effect on the dependent variable. The Relation between the five levels of cognitive workload (Task) has a nearly linear relationship with the dependent variable of Mean Measured Workload (Pupil dilatation in mm), visible in figure19).
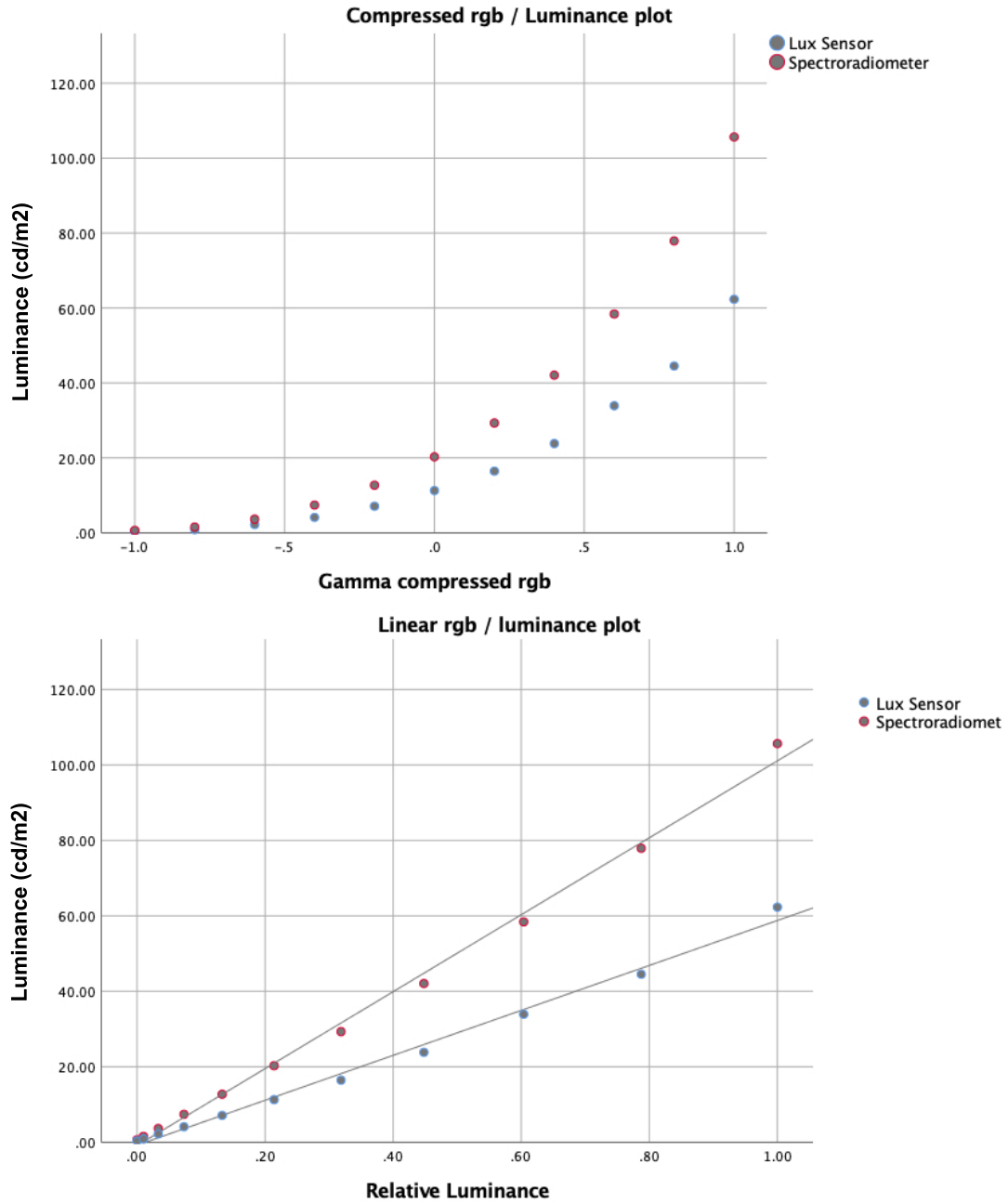
Figure 13: Experimental setup luminance measurement measured by the Spectroradiometer and the head mounted sensor. Represented (top) for eleven RGB values (non-linear, gamma-compressed), expressed as normalised values from -1 black to +1 white and for the respective relative luminance (bottom) as the normalised value of luminance computed from linearised RGB converted with the luminosity function (average spectral sensitivity of human vision).
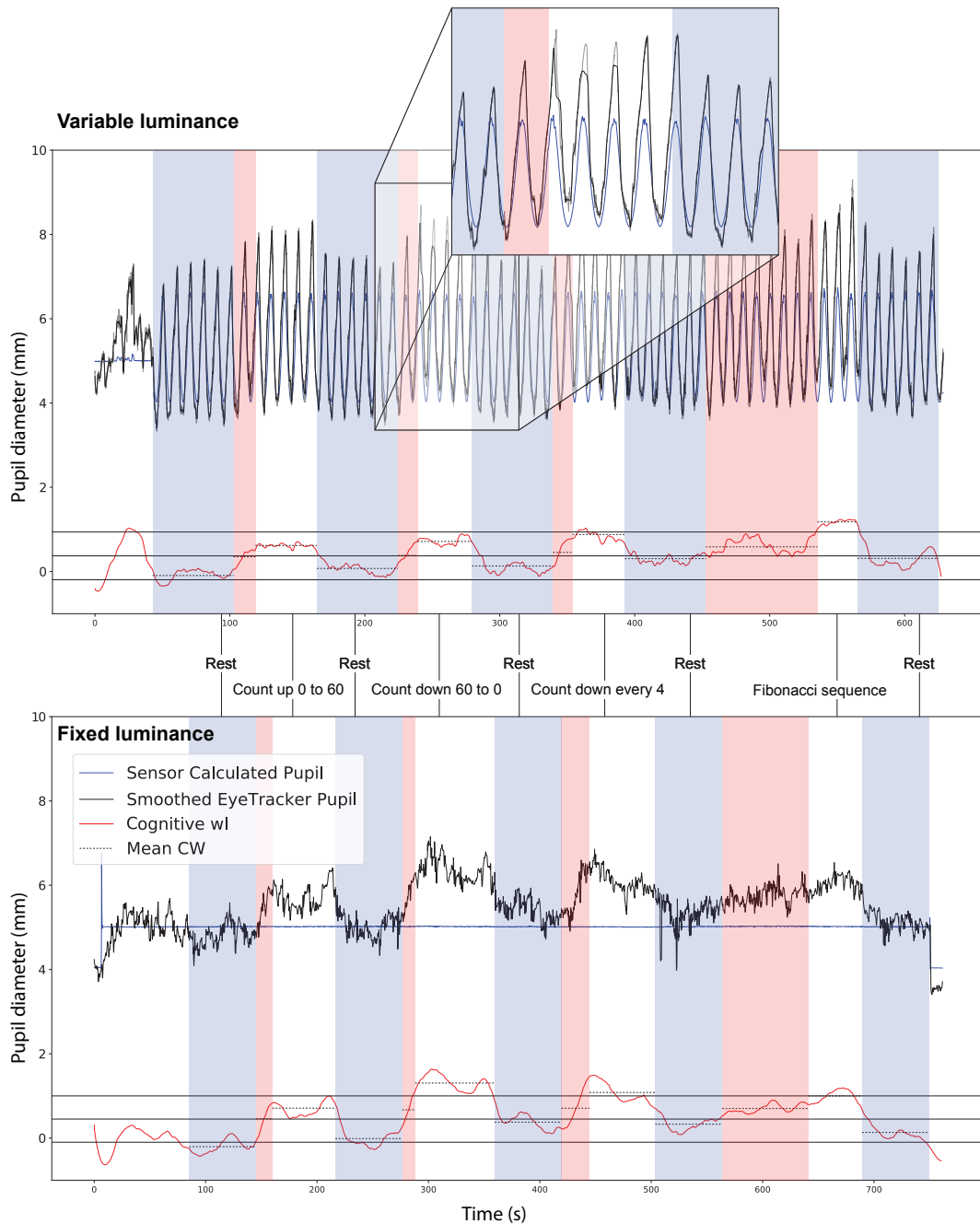
Figure 14: Two examples of the recorded data, on the top a variable light experiment, on the bottom a fixed light experiment. The black line represent the measured pupil size over time (how dilated was the pupil during the experiment). This value is compared to the expected pupil size (blue line) describing how dilated the pupil should be for the current visual stimuli, to extract the remaining pupil dilatation relative to cognitive workload (red).

Figure 15: To check the correct removal of the effect of light (sinusoidal wave pulsating at 0.05Hz) through the main algorithm (red line) a band-pass filter is used to produce similar result (orange line), suppressing a narrow band of frequencies around 0.05Hz. The orange and red line are closely overlapping (the horizontal alignment is effected by the phase shift result of the band-pass filter).

Group Statistics

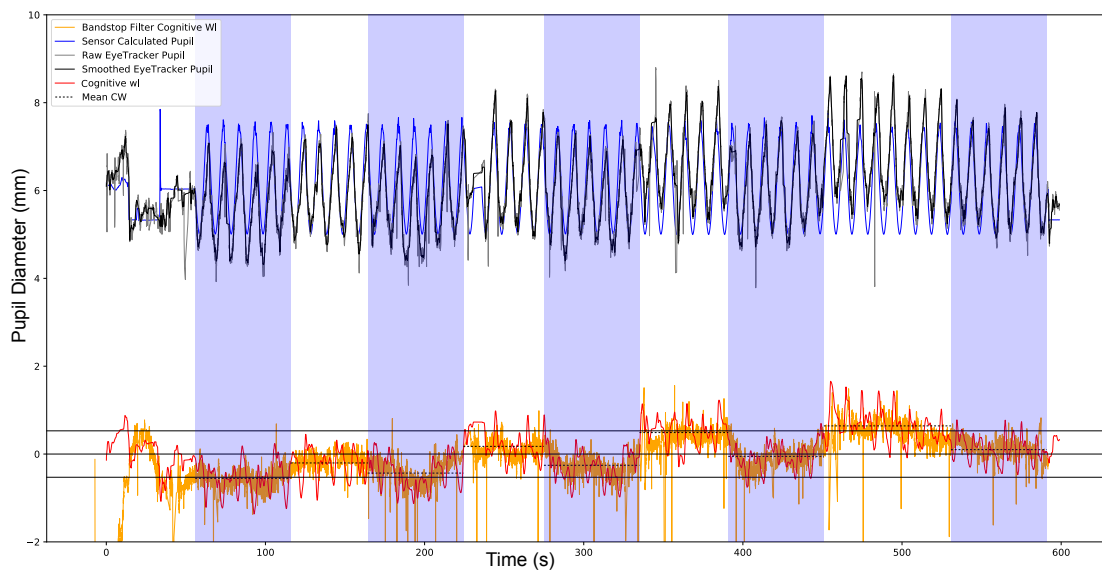|	Ligt	N	Mean	Std. Deviation	Std. Error Mean
Rest	.00	6	-.2300533	.12558335	.05126919
	1.00	15	-.2809293	.08033905	.02074345
Count Up	.00	6	.1475907	.18327162	.07482032
	1.00	15	.0833825	.16748862	.04324538
Count Down	.00	6	.4228236	.32699266	.13349419
	1.00	15	.1702164	.18570748	.04794946
Count -4	.00	6	.4279963	.38110701	.15558628
	1.00	15	.3930646	.20602513	.05319546
Fibonacci	.00	6	.4815192	.27587210	.11262431
	1.00	15	.4807614	.16655513	.04300435

Independent Samples Test

		Levene's Test Eq. of Variances		t-test for Equality of Means				Std. Error Diff.	95% Diff. Confidence Interval	
		F	Sig.	t	df	Sig.	Mean Diff.	Std. Error Diff.	Lower	Upper
Rest	Equal variances assumed	3.278	.086	1.116	19	.278	.0509	.0456	-.0445	.1463
	Equal variances not assumed			.920	6.707	.390	.0509	.0553	-.0811	.1828
Count Up	Equal variances assumed	.010	.922	.774	19	.449	.0642	.0830	-.1095	.2379
	Equal variances not assumed			.743	8.558	.477	.0642	.0864	-.1328	.2613
Count Down	Equal variances assumed	2.291	.147	2.260	19	.036	.2526	.1118	.0186	.4866
	Equal variances not assumed			1.781	6.336	.123	.2526	.1418	-.0901	.5953
Count -4	Equal variances assumed	3.111	.094	.274	19	.787	.0349	.1273	-.2316	.3015
	Equal variances not assumed			.212	6.207	.839	.0349	.1644	-.3642	.4340
Fibonacci	Equal variances assumed	2.103	.163	.008	19	.994	.0008	.0972	-.2026	.2041
	Equal variances not assumed			.006	6.515	.995	.0008	.1206	-.2887	.2902

Figure 16: Independent sample t-test result, output by IBM SPSS (IBM 2017).

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Task	Sphericity Assumed	5.946	4	1.486	57.801	.000
	Greenhouse-Geisser	5.946	2.533	2.347	57.801	.000
	Huynh-Feldt	5.946	3.109	1.913	57.801	.000
	Lower-bound	5.946	1.000	5.946	57.801	.000
Task * Light Condition	Sphericity Assumed	.168	4	.042	1.633	.175
	Greenhouse-Geisser	.168	2.533	.066	1.633	.200
	Huynh-Feldt	.168	3.109	.054	1.633	.190
	Lower-bound	.168	1.000	.168	1.633	.217
Error(Task)	Sphericity Assumed	1.954	76	.026		
	Greenhouse-Geisser	1.954	48.124	.041		
	Huynh-Feldt	1.954	59.063	.033		
	Lower-bound	1.954	19.000	.103		

Tests of Within-Subjects Contrasts

Source	factor1	Type III Sum of Squares	df	Mean Square	F	Sig.
Task	Linear	5.360	1	5.360	136.124	.000
	Quadratic	.546	1	.546	17.658	.000
	Cubic	.037	1	.037	1.444	.244
	Order 4	.002	1	.002	.342	.566
Task * Light Condition	Linear	.007	1	.007	.183	.674
	Quadratic	.077	1	.077	2.486	.131
	Cubic	3.049E-5	1	3.049E-5	.001	.973
	Order 4	.084	1	.084	11.873	.003
Error(Task)	Linear	.748	19	.039		
	Quadratic	.587	19	.031		
	Cubic	.484	19	.025		
	Order 4	.134	19	.007		

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	3.767	1	3.767	38.024	.000
Light Condition	.139	1	.139	1.408	.250
Error	1.882	19	.099		

Figure 17: GLM Repeated Measures data output generated by IBM SPSS (IBM 2017).

Within-Subjects Factors

Task	Dependent Variable
1	Rest
2	Count Up
3	Count Down
4	Count -4
5	Fibonacci

Between-Subjects Factors

		Value Label	N
Light Condition	1	Fixed	6
	2	Variable	15

Multivariate Tests

Effect		Value	F	Hypothesis df	Error df	Sig.
Task	Pillai's Trace	.914	42.667 [b]	4.000	16.000	.000
	Wilks' Lambda	.086	42.667 [b]	4.000	16.000	.000
	Hotelling's Trace	10.667	42.667 [b]	4.000	16.000	.000
	Roy's Largest Root	10.667	42.667 [b]	4.000	16.000	.000
Task * Light Condition	Pillai's Trace	.415	2.832 [b]	4.000	16.000	.060
	Wilks' Lambda	.585	2.832 [b]	4.000	16.000	.060
	Hotelling's Trace	.708	2.832 [b]	4.000	16.000	.060
	Roy's Largest Root	.708	2.832 [b]	4.000	16.000	.060

Mauchly's Test of Sphericity

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[b] Greenhouse-Geisser	Epsilon[b] Huynh-Feldt	Lower-bound
Task	.221	26.271	9	.002	.633	.777	.250

Figure 18: GLM Repeated Measures data output generated by IBM SPSS (IBM 2017).
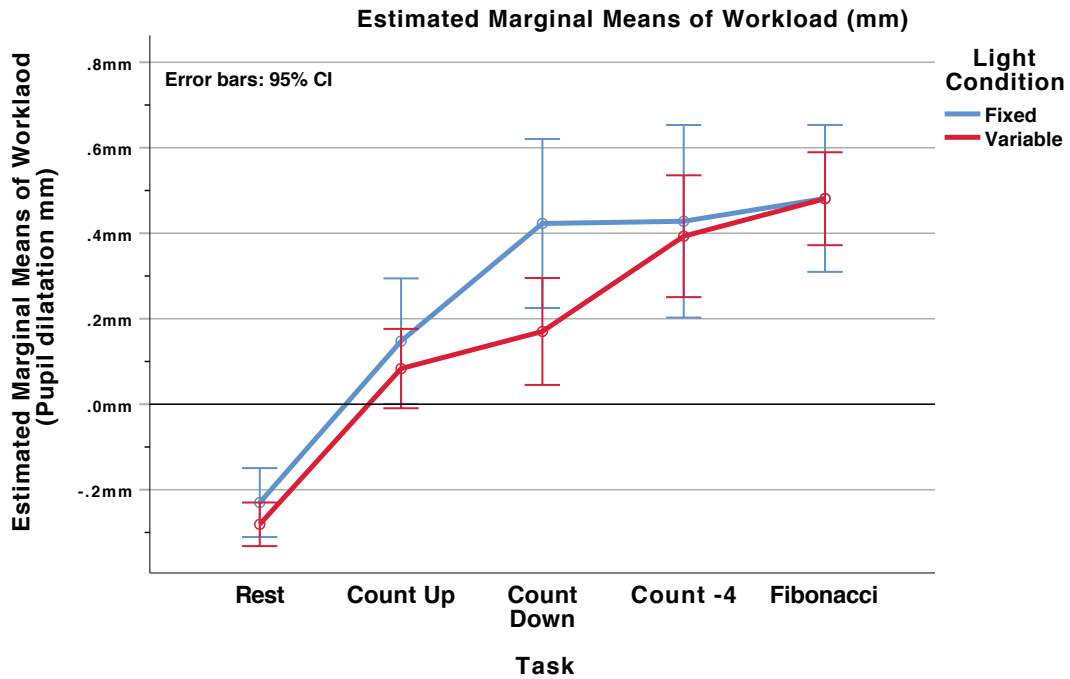
Figure 19: GLM Repeated Measures data output generated by IBM SPSS (IBM 2017).
The vertical scale represents the average pupil dilatation (change in diameter) across all the participants calculated from the mean diameter of each test. The 0mm point is only relative to this experiment and marks the average workload (or average pupil dilatation as consequence of cognitive workload) specific to this series of tasks. It tends toward the rest condition mean as the rest is over-represented (five repetitions) in comparison to the other tasks. The correct way to read the chart is to evaluate the change between the tasks (i.e. between a "count up" and a "Fibonacci" task it can be expected to observe a 0.4mm dilatation) and not as absolute values (i.e. assume that a -0.2mm contraction below the average of a recording indicates that the subject was resting).

4.4 Discussion

The two examples in figure 14 are of the main(top) and control(bottom) experiments, these are plots of two different participants and are a good representation of what most of the data looks like: the difference between the measured pupil size and expected pupil size increases (pupil dilate) during the tasks and decreases during the "rest". In the main experiment expected and measured pupil size follow the sinusoidal sequence, the effect of the light change is therefore removed (partially) from the measured pupil size. The two horizontal lines surrounding the red line (figure 14) mark plus and minus one standard deviation from the mean "cognitive workload". Consistently among the participants, the Fibonacci sequence produces dilatation of above one SD and the initial rest measures often below one SD.

4.4.1 GLM Repeated Measures

Figure 19 substantiate the relation of cognitive workload and pupil size showing dilatation of the pupil as the cognitive workload increases. It also shows that the different light condition (between-subjects factor) does not have an effect on the measured cognitive workload confirming that the data processing is removing a significant portion of the light effect on pupil size.

4.4.2 Independent Sample t-test

On the matter of the influence of light, comparing groups statistics and the output of an independent sample t-test between variable and fixed light condition (figure 16) highlights that, as visible in the figure 19, the mean and variance of the control group (fixed light) and experiment group (variable light) do not have a significant difference. The only meaningful deviation is for the countdown task that is characterised by a 0.25 mm difference. The deviation in countdown task can be explained by the high variability between the participants for this particular task, as some were seemingly struggling with it while others were doing it with ease,

4.4.3 Research Question Follow-up

> "Is the system, composed by the eye-tracker and baseline algorithm, capable of measuring the cognitive workload reliably independently to the visual stimuli?"

The laboratory test confirms that it is possible to remove the effect of light from the pupillary response using the development system and sensor and that the resulting signal represents changes in cognitive workload.

5   Field test

5.1   Introduction

A second test session was scheduled early in the process to include a field test in the project. The development was carried out in such a way that would account for multiple conditions of use, with a variety of test runs conditions: day/night, outdoor and indoor, driving, walking, sitting at a desk, to evaluate how the system would cope with outside and intrinsic interference. Initial contacts were held with two main test partners, a helicopter pilot operating in the Oslo area and a representative of the Royal Norwegian Naval Academy (RNoNA) Odd Sveinung Hareide. Once the system has proven sufficiently promising to justify further exploration, the session with the RNoNA was finalised and the participants recruited among the students of the academy. The scope of a field test is not to validate the instrument, as the confounding variables are almost impossible to control, and it is, therefore, impractical to prove causation or correlation between variables. The shift between a controlled laboratory condition to a real scenario in field conditions has two consequences to consider:

- the noise produced by the ever-changing light and gaze position dramatically increases;
- the cognitive workload distribution will variate between subjects unpredictably and will not show clear steps as in the previous test as it is not composed (exclusively) of well-defined sub-task with constant workload;
- the cognitive workload of the task could be either too stable or too little to have a significant effect on the pupil, as explained by Kahneman (2012) they could not observe a significant dilation of the pupil comparing a subject at rest with a subject doing "small-talk" and could only observe dilatation as the subject started performing mathematics tasks. This would imply that highly "automated" tasks such as driving and walking will not produce a visible change in pupil size.

Therefore the scope of a field study is to answer two alternative questions:

- Is this method usable in field condition (where luminance variates in an unpredictable manner), how does it compare with more traditional methods of assessing the workload (in either practicality or quality of the data)?
- Does the result correlate with subjective data or can be used in conjunction with such data?

5.2   Method

5.2.1   Participants and Methodology

The study included five cadets of the Royal Norwegian Naval Academy (RNoNA), with on average three sailing years experience, all males. When recruiting subjects for the data collection, several challenges with the availability of relevant personnel were identified. The workload on personnel is high, and as data collection is not an operational service, this makes recruiting participants

Figure 20: The vessel (Kvarven, on the right) used for the test session, in the RNoNA harbour.

challenging and only convenience sapling is to be expected. Across the rest of the chapter the participants will be referred to as Nx Ax, N stands for navigator, A for assistant, followed by the sequential number of test rounds (e.g. N2 and A2 were recorded at the same time), if referring to only one of the two halves of a test run the naming is extended with a letter (N2A N2B).

5.2.2 Apparatus

The vessel (Kvarven), figure 20, used for the test session is a training vessel often used by the RNoNA students, it is a small (15m) vessel equipped with a standard navigation system, the speed limit for such vessels is 30 knots, the target speed for the test was set to 25knots. A navigator generally operates the boat alongside with an assistant to the navigator and the helmsman/throttle-man, a training instructor is usually present and sitting behind the crew. Two Pupil Pro eye tracker (Kassner et al. 2014), infrared pupil trackers were used at a sampling frequency of 120 Hz @400x400px resolution. The average luminance in front of the participants was logged at 10Hz using the tsl2591 Lux sensor (AMS 2013). The data was saved directly on two different laptops, one for each participant figure 21. Gps data was recorded using a smartphone while the rest of the subjective data, NASA TLX (Hart & Staveland 1988) and perceived workload on a map were done with pen and paper A.1 and A.5.

Figure 21: One of the study participants wearing the eye tracker and luminance sensor.

5.2.3 Procedure

The test session, scheduled for the 24th of April 2019, started with a long pretest to test and set up all the equipment as well as define the stopping point on the course. Afterwards, the five participants were embarked and briefed inside the vessel. They were introduced to the route 22 and the stopping point, halfway through the route. They were given the time to read through the instruction to correctly compile the Nasa TLX A.1 and time to read and sign the consent form A.3.

The five sessions required three participants each time, the navigator, sitting in the centre, the assistant on the left and the helmsman on the right, the data was recorded from the assistant and navigator only. The number of participants and the number required each run means that each participant had to experience the run multiple times, in different roles. Each session was dived in two parts, each requiring around 18 minutes of navigation. At the end of the first half, the participants were asked to compile both the pairwise score for the task (Hart & Staveland 1988) and the rating scale. At the end of the second half, as the next group were getting in position, they were asked to compile a second rating scale as well as draw on a map of the route (A.5 and figure 23) with three markers, defining green as below average workload, yellow as average workload and red/purple as above average workload. The pairwise score was not repeated as the two halves of the tasks were considered as very similar in the distribution of the workload.

5.2.4 Data Analysis

The data recorded during the five-session includes a faceted measure of workload:

- the psychophysical measure derived from the pupil size, representing the relative changes of cognitive workload over time, and in this case, thanks to the GPS data, over the course;
- the Nasa TLX as a fine grain self-reported and absolute (not relative) measures of workload for each half of the course;
- and finally the map drawings, a "low grain" tool, as it allows for only three levels of self-reported cognitive workload but with the addition of the spatial dimension.

The pen and paper Nasa TLX forms were digitised in a spreadsheet to calculate the final score for each session, this data produces a standardised index of load and allows to compare the perceived load of the first half versus the second half of each run. The pairwise score for each of the six workload metrics can also be used to check how the same task is perceived by the different participants (figure 24). The manual (self-reported) map data has been divided into forty-one sections (figure 25) of around 500mt (2.7 Nautical miles) and scored accordingly to colour (1-green, 2-yellow and 3-red) the mean perceived workload is visible in figure 26.

The pupillometry data has been processed similarly to what was recorded in the laboratory study. In this case the video data from the camera is fundamental to the process of removing the effect of light (see the chapter 3 and in particular figure 4) as gaze position has a significant impact on the pupil size in a field condition, the improvement over the sole use of the luminance sensor is dramatic. In figure 28 it is visible how the yellow (pupil size calculated with the gaze and camera data) follows more consistently the black line (measured pupil size) than the blue line (pupil size based on the sole luminance sensor). The pupillometry data is then paired with the GPS data producing two outputs:

- a data-set of the average measured workload for each of the forty-one sections, a summary is visible in figure 27, to be used for quantitative analysis;
- a series of plots with high spatial resolution and time marks to be used for qualitative analysis (visual comparison with the self-report maps and comparison with the video recordings), figure 29.

5.3 Results

The digitised map data (figure 26, self reported workload over the course) correlates (figure 30) with the measured workload (figure 26). The Workload variation over the recording is quite limited (see figure 28) compared with the results of the lab experiment (figure 14) indicating a small overall change of the workload over the course.

The qualitative analysis includes a broad comparison between the output plot and the self-reported workload maps (figure 29), on both visualisations, the workload is higher in more confined waters (tight passages with obstacles such as around sections 7/8 and section 34/35) and before and after course alterations. Also, with more traffic density, the workload is expected to increase (around 24/30 the commercial and civil traffic is higher, and the course was often altered).
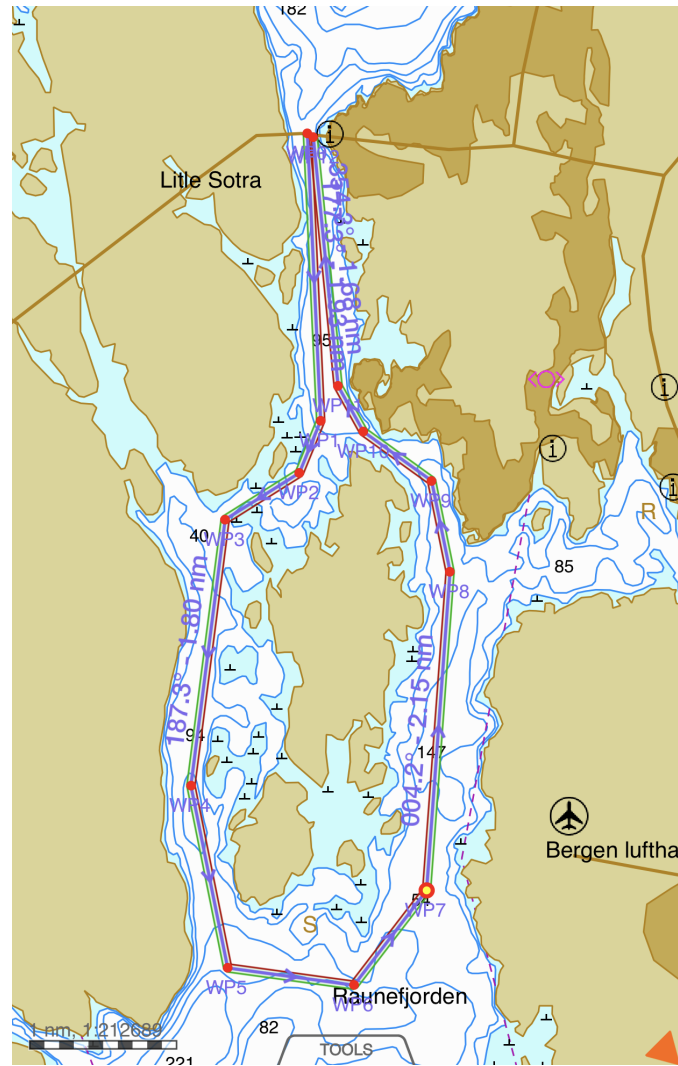
Figure 22: The route, courtesy of Odd Sveinung Hareide, starts and ends under the Sotra Bridge near the RNoNA harbour and runs clockwise around the Bjorøyhavn island.
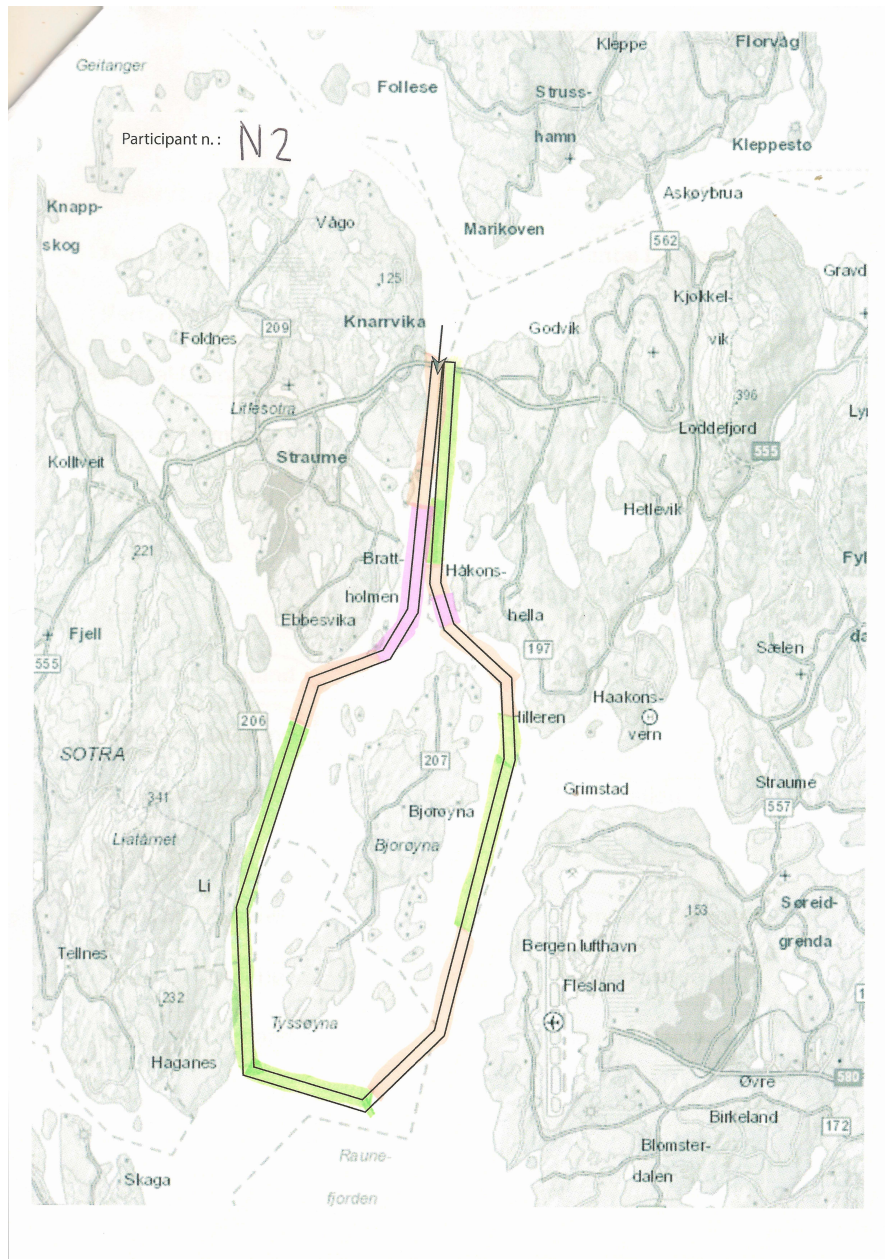
Figure 23: Each participant complied a self-report map of the cognitive workload after each run. They were instructed to use three markers, green to mark below average workload, yellow for average workload and red/purple for above average workload, unfortunately, this was not completely clear for them and often interpreted as high medium and low workload, for this reason, the last few maps are almost entirely green.

Figure 24: The aggregated Nasa TLX (Hart & Staveland 1988) score for all the participants A(first half) B(second half). The TLX score is an ordinal scale (from 0 to 100) in which an higher score corresponds to higher workload.
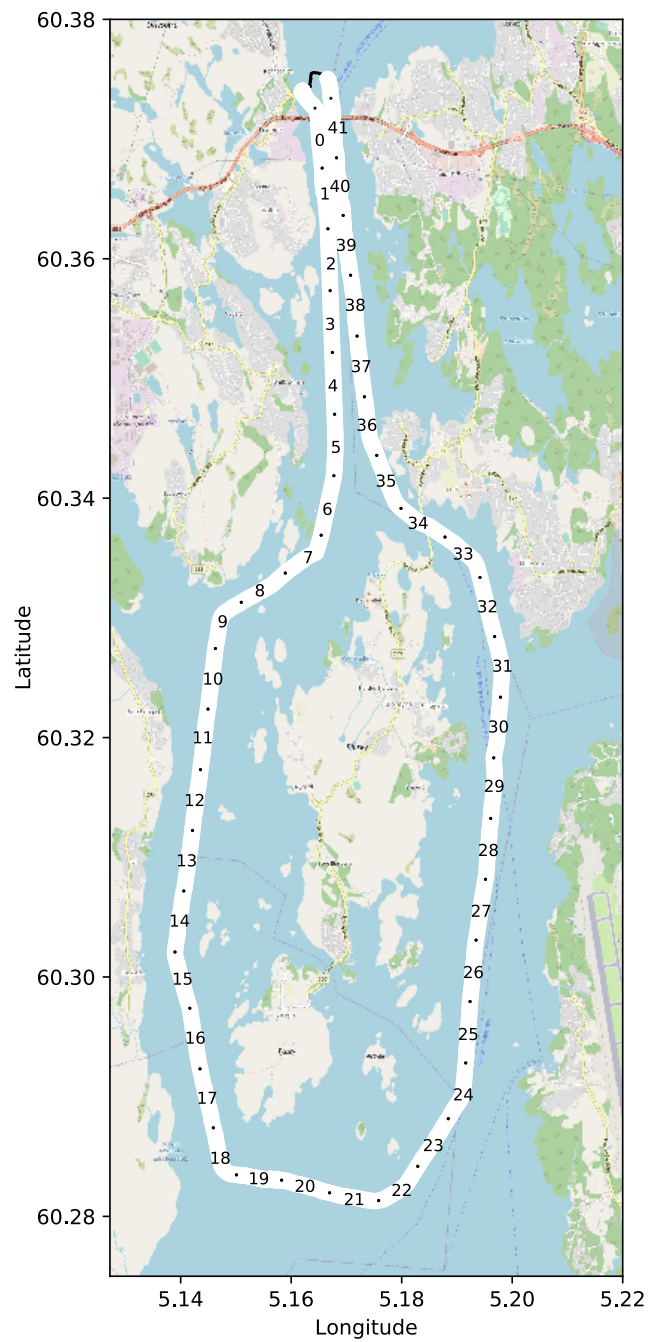
Figure 25: Eye tracking data and self reported map data are divided in forty one 500mt sections to be comparable. The Sotra Bridge is used as start line for each recording to position the first section.
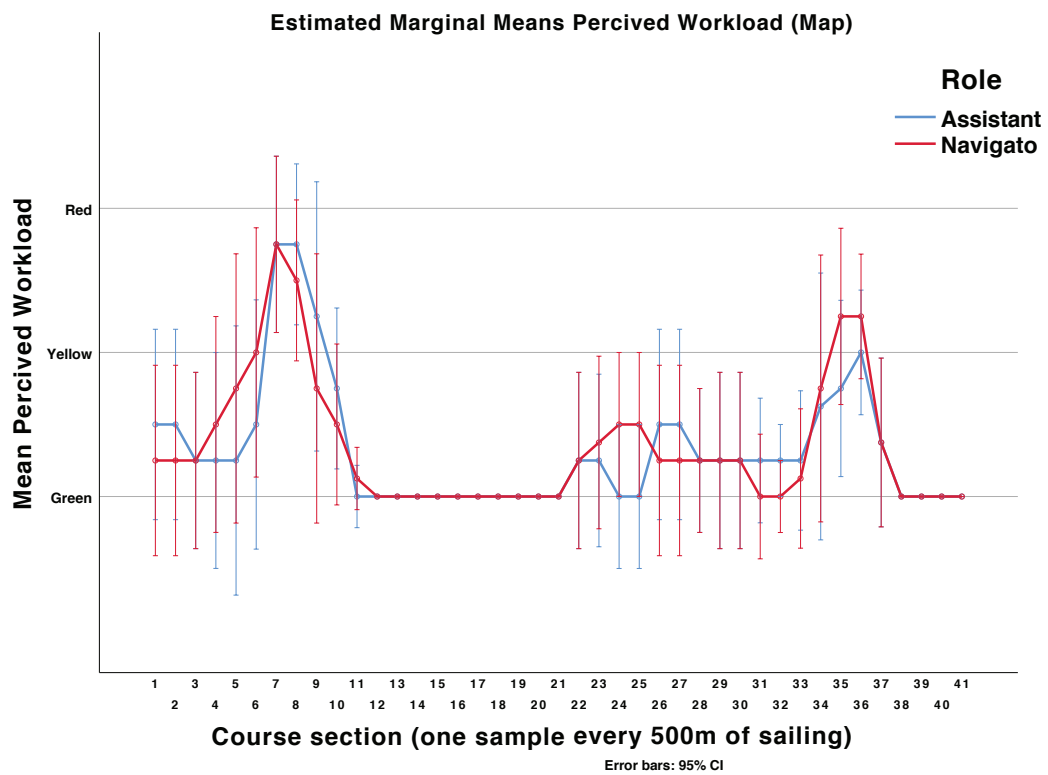
Figure 26: The mean perceived workload for the forty one sections (figure 25). Green is described as below average workload, yellow as average workload and red/purple as above average workload. Output generated by IBM SPSS (IBM 2017).
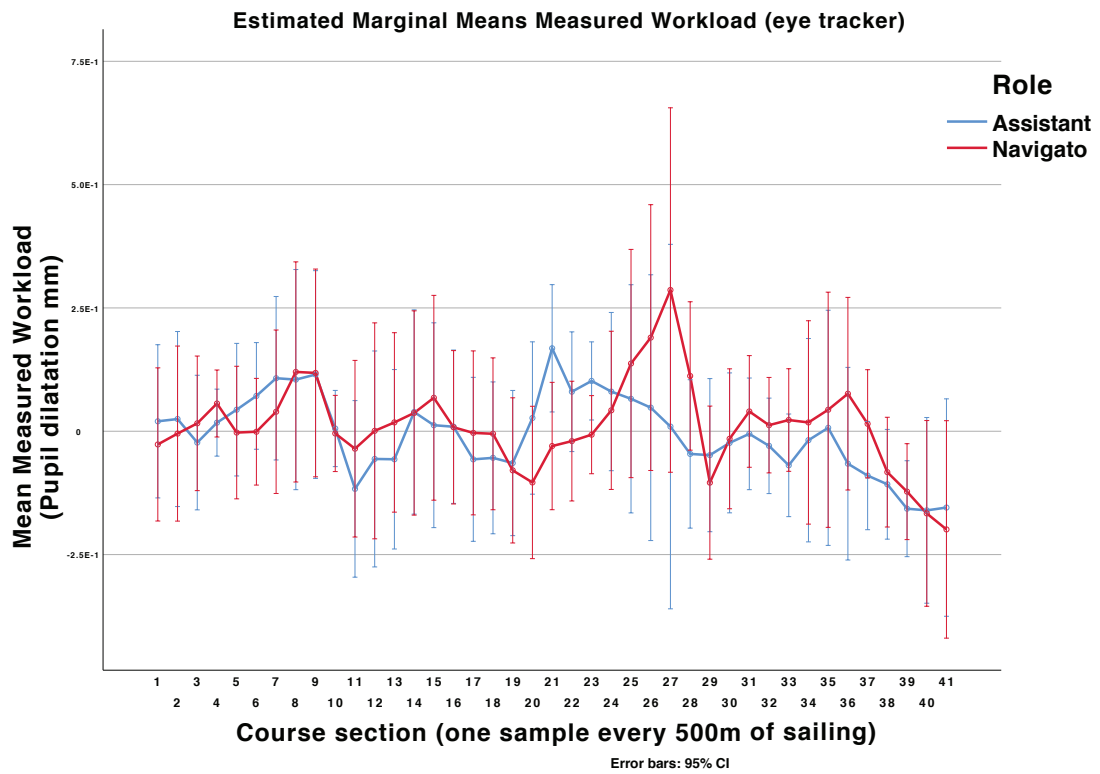
Figure 27: The mean measured workload for the forty one sections (figure 25). The 0mm point is only relative to this experiment and marks the average workload (or average pupil dilatation as consequence of cognitive workload) specific to this tasks. Output generated by IBM SPSS (IBM 2017).
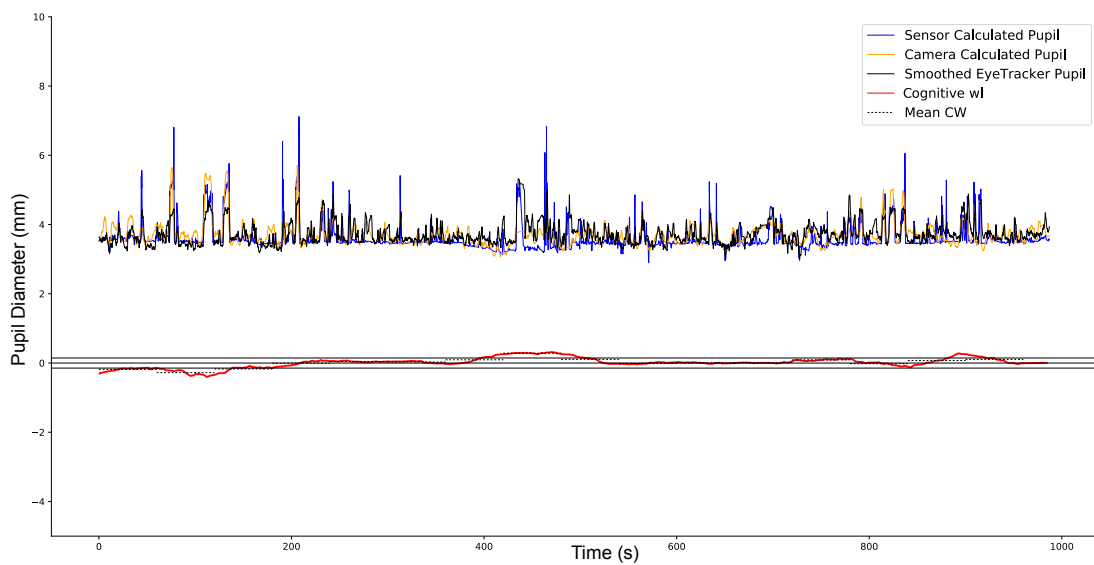
Figure 28: An example of the output data for one recording (half course one participant), the filtering is quite more aggressive than what the laboratory experiment, to remove the quite higher noise floor consequence of the field condition. The use of camera data is fundamental as it reduces the effect of the high contrast between the differed areas of interest (outside and digital instruments), see the chapter 3 and in particular figure 4.

Navigator I
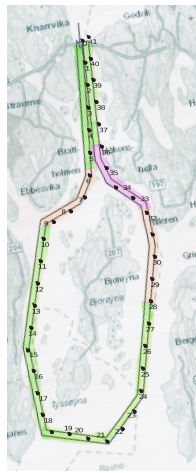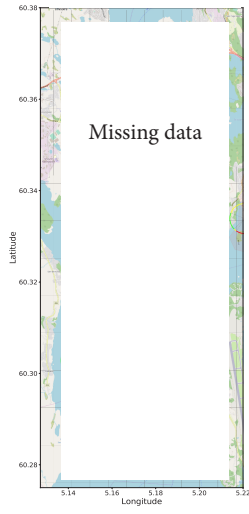
Navigator II
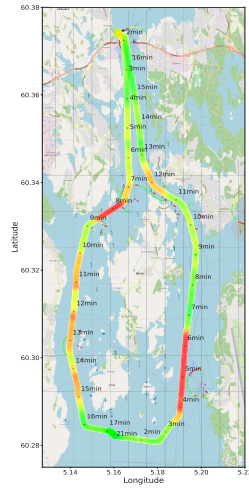
Navigator III
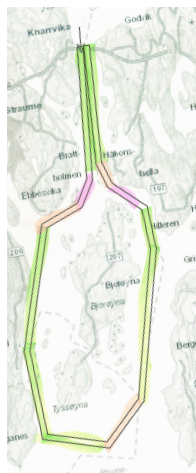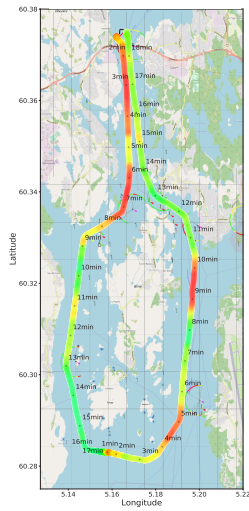
Navigator IV

Navigator V

Workload

-1.5 SD   Mean   +1.5 SD
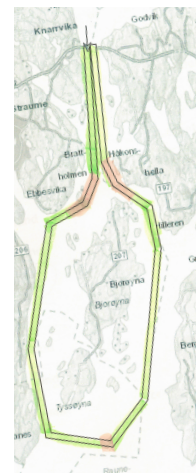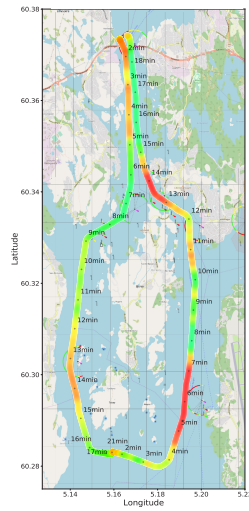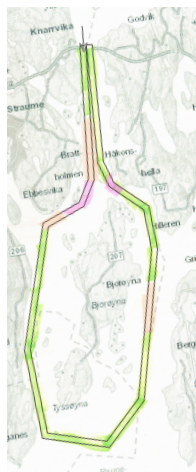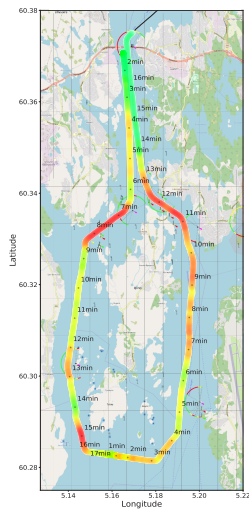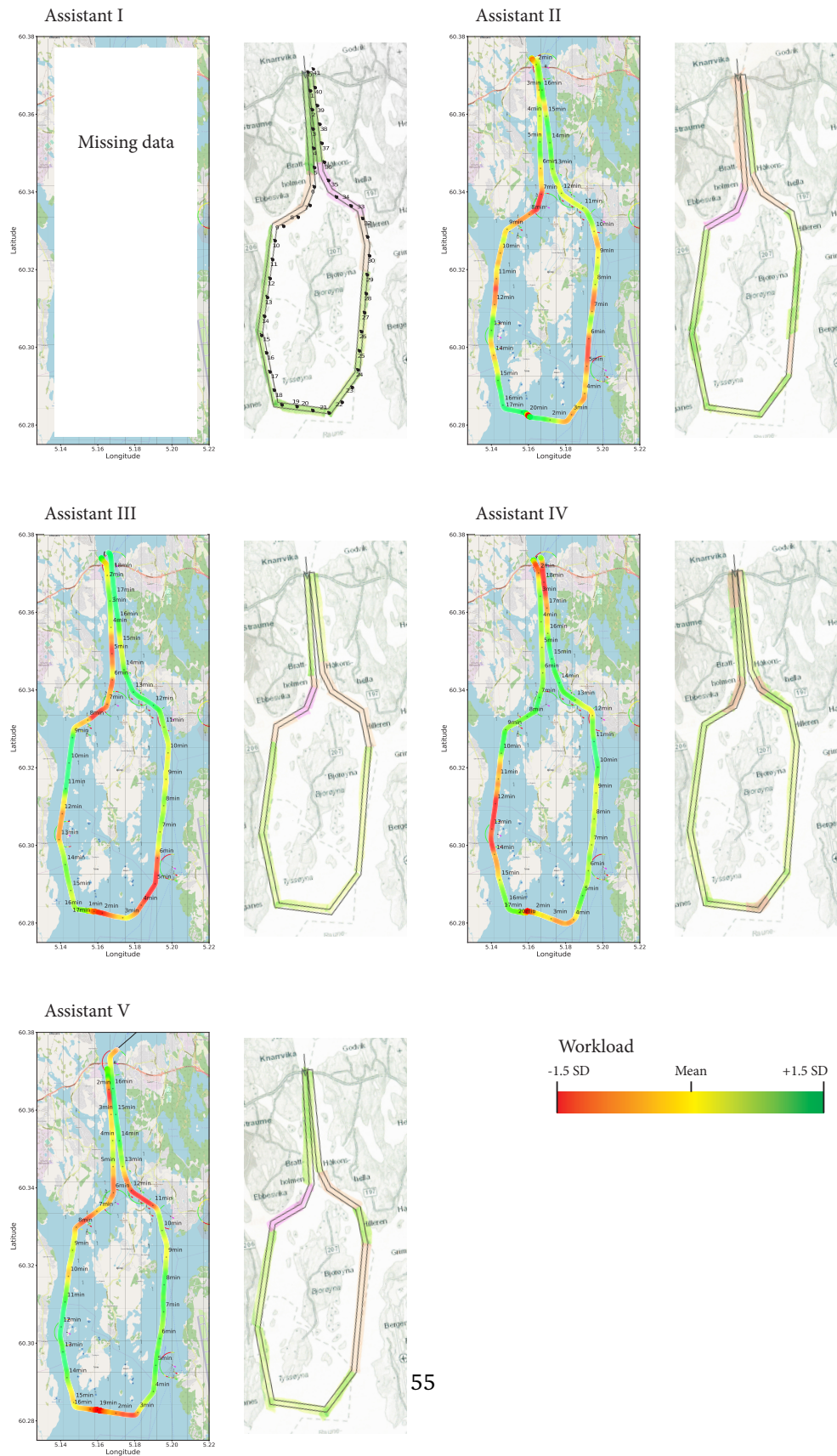
Figure 29: Output data for qualitative analysis of workload.

Correlations

			Mean Measu red Workload	Mean Percived Workload
Spearman's rho	Mean Measu red Workload	Correlation Coefficient	1.000	.538**
		Sig. (2-tailed)	.	.00024
		N	42	42
	Mean Percived Workload	Correlation Coefficient	.538**	1.000
		Sig. (2-tailed)	.00024	.
		N	42	42

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 30: Spearman's correlation between the Measured Cognitive workload (pupillometry, Scale data) and Perceived Cognitive Workload (Self Reported Map, Ordinal Data). The correlation is positive (as one value grows the other does too). Output generated by IBM SPSS (IBM 2017).

5.4   Discussion

The quantitative analysis of such faceted data is complex, more so considering the small number of subjects. The Spearman correlation (figure 30) between the digitised map data (figure 26, self-reported workload over the course) and the measured workload (figure 26), suggests that the eye tracking data indeed represents the cognitive workload albeit with some noise. The limited workload variation over the recording (see figure 28) compared with the results of the lab experiment (figure 14), can be traced back to the small variation and low overall workload that is experienced by either the Assistant and Navigator compared to the large difference between rest and a rather demanding task such as the "Fibonacci" sequence.

The Nasa TLX data (figure 24) has a limited temporal resolution but is nevertheless of use as it is the only absolute measure of cognitive workload collected for the experiment. Of interest is the apparent accord between navigator and assistant, as in every case the value changes between the first and second half (A and B) in the same direction. Overall, it is noticeable how repeating the activity, albeit in a different role, reduces the perceived load.

The information provided by the TLX scoring can be used to weight the distribution of the load and evaluate the meaning of the other self-reported and psychophysical data (figure 29). In this instance the TLX data reports that the task did not include extreme conditions of overload or under-load, the average yellow/orange (in figure 29) can be considered as a comfortable working condition with fluctuations (red/green) representing a normal attention cycle.

A more detailed qualitative analysis, which involves going through the videos to identify the source of the workload peaks in the output, highlights how the precision of the workload data is highly connected to the quality of the eye tracking data. Reflection in the pupil, incorrect framing of the eye and eye anatomy can unpredictably compromise the tracking either with noise or false positives. Two examples are in Navigator II and IV (figure 29). Around the 6th minute, a false peak is present on both, but for different reasons, in the Navigator II recording, poor framing compromises the tracking disrupting the recognition algorithm in what would be otherwise a sound recording. Instead, the Navigator IV recording is limited by the reflection of the bright windshield on the dark pupil, making it poorly recognisable. This condition is to be expected in an environment with high infrared light content (sunlight) but requires manual analysis of the data to be recognised. When the recording has good quality, proper tracking of the eye and pupil size as well as tracking of the gaze,the resulting estimation shows a clear connection between the actions of the participants and the level of workload. This state is, unfortunately, difficult to achieve as the eye-tracking technology does not seem to be robust nor flexible enough to account for the variability between subjects and require significant effort and experience on the operator's side. Variation in the eye appearance such as a pronounced Epicanthic Fold or lower contrast between the iris and pupil can also have a significant effect on the quality of the eye recognition, as the computer vision algorithm struggles to identify the eye's features in the video.

5.4.1   Research Questions Follow-up

"Is this method usable in field condition (where luminance varieties in an unpredictable manner), how does it compare with more traditional methods of assessing the workload (in either practicality or quality of the data)?"

The developed system has proven to be usable in a variety of field conditions such as maritime and automotive. Extensive training of the operator is still necessary to correctly set up the eye tracker and design the experiment, as the quality of the workload data is highly dependant on the quality of the recordings. Compared to more traditional subjective data, either qualitative or quantitative, this approach allows for a higher temporal resolution but still lacks the ability to measure workload on an absolute scale (only recording changes in workload).

> "Does the result correlate with subjective data or can be used in conjunction with such data?"

The self-reported workload, collected in the forms of colour-able maps of the sailing course, correlates with the result of the pupillometry and is a useful method to support the eye-tracking data and verify the quality of the recordings. The NASA TLX scores can be used alongside eye-tracking to evaluate the cognitive workload as an absolute measure and correctly interpret the pupillometry data (variation of workload).

6   Conclusion

This research project aimed to push the use of psychophysical metrics in usability from the "usability lab" to the field. In its small way, it was successful in showing how this can be achieved for pupillometry. By using inexpensive consumer parts and the already present onboard camera, the developed system can dynamically analyse the visual stimuli the user is subjected to and remove most of its effect from the pupillometry data. The laboratory test session further indicates that the output signal generated by the system retains the effects of cognitive workload on pupil dilatation and only a limited amount of noise and can be used for both qualitative and quantitative analysis of workload quality and distribution. Hopefully, this work will encourage a pragmatic use of technology, in which qualitative and quantitative data can be used together to represent the complex relations between users and products, without being restricted to the artificial boundaries of the laboratory condition or obligated to stand on the unstable foundation that often subjective data is.

6.1   Future Work

The data processing developed in this project has proven its effectiveness in removing the effects of light from the measure of pupil size, but it remains in an immature state of development. Future work on the subject could include:

- Implement the ability to read exposure data over time from the camera and eliminate the necessity to use the external sensor.
- Revisit the "area of interest" detection algorithm in the video analysis to represent better the way the human eye adapts to different luminous stimuli. The video analysis was developed in a purely empirical manner, comparing the results of a few different algorithms (figure 4) and deserve more in-depth research, both on the side of Computer vision and the psychophysical response of the pupil.
- Automatise the process, hiding and automating some of the parameters to reduce the threshold of the use of this software.
- Investigate how to evaluate Cognitive Workload as an absolute metric: at the moment the results have to be interpreted and can not be used as an absolute value as it is expressed as a change in pupil diameter not as an absolute metric of cognitive workload.

6.2   Repeatability

The code used during the field and laboratory study has been iterated upon multiple times to experiment and expand its functionalities. As it often happens, it was at the end difficult to use and overly complicated for everyone apart from the author. In order to encourage the reuse and possibly improvement of the software, a final revision was made to create a "public" version. This version retains only the main functionalities necessary to extract the cognitive workload measure
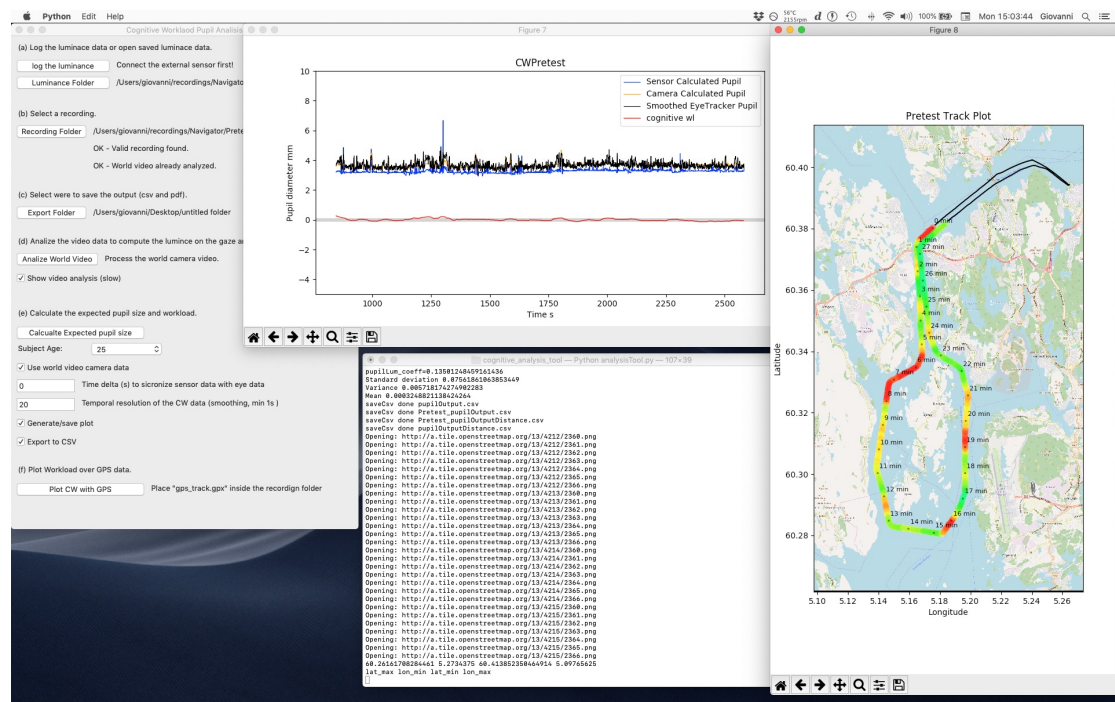
Figure 31: The GUI has been added to allow the use of the software to a broader group of researchers It includes all the main functionalities of the original code and allows the export of output data for further analysis with other statistical tools.

form eye tracking recording:

- Record luminosity data from the external sensor.
- Analysis of the World video to calculate the luminance around the gaze.
- Calculation and export of the expected pupil size for the recorded visual stimuli.
- Calculation and export of the cognitive workload related pupil dilatation.
- Visualisation of workload over time or over geographical coordinates (using GPS data).

The script is now accompanied by a basic GUI, that allows the input of all the main parameters without having to modify the code directly (figure 31). The code is made available as a GitHub Repository including extensive documentation to get started, assemble the luminance sensor, and analyse the data. The documentation is visible online or in the appendix A.6. It includes a bill of material to assemble all the necessary parts for less than a hundred dollars (excluding the Eye Tracker). One assembled sensor kit was sent for evaluation to Luciano Perondi, Associate Professor at Iuav University in Venice, to be used in an automotive setting and is hopefully only the first to experiment and reuse this system.

Bibliography

Afergan, D., Peck, E. M., Solovey, E. T., Jenkins, A., Hincks, S. W., Brown, E. T., Chang, R. & Jacob, R. J. (2014), Dynamic difficulty using brain metrics of workload, *in* 'Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14', ACM Press, Toronto, Ontario, Canada, pp. 3797–3806.
URL: *http: // dl. acm. org/ citation. cfm? doid= 2556288. 2557230*

AMS (2013), *TSL25911_Datasheet_EN_v1.pdf ams163.5*, AMS.
URL: *https: // cdn-shop. adafruit. com/ datasheets/ TSL25911_ Datasheet_ EN_ v1. pdf*

Boles, D. B. & Adair, L. P. (2001), 'Validity of the Multiple Resources Questionnaire (MRQ)', *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* **45**(25), 1795–1799.
URL: *http: // journals. sagepub. com/ doi/ 10. 1177/ 154193120104502508*

Cacioppo, J. T., Tassinary, L. G. & Berntson, G., eds (2007), *Handbook of Psychophysiology*, 3 edn, Cambridge University Press, Cambridge.
URL: *http: // ebooks. cambridge. org/ ref/ id/ CBO9780511546396*

Chen, H., Tan, G. & Wu, S.-T. (2017), 'Ambient contrast ratio of LCDs and OLED displays', *Optics Express* **25**(26), 33643.

Clemson, P. T. & Stefanovska, A. (2014), 'Discerning non-autonomous dynamics', *Physics Reports* **542**(4), 297–368.
URL: *https: // linkinghub. elsevier. com/ retrieve/ pii/ S0370157314001318*

community, T. S. (2014), *Butterworth digital and analog filter design.*, The Scipy community.
URL: *https: // docs. scipy. org/ doc/ scipy-0. 14. 0/ reference/ generated/ scipy. signal. butter. html*

community, T. S. (2015), *Apply a Savitzky-Golay filter.*, The Scipy community.
URL: *https: // docs. scipy. org/ doc/ scipy-0. 16. 1/ reference/ generated/ scipy. signal. savgol_ filter. html*

Council, U. N. R., ed. (1997), *Tactical display for soldiers: human factors considerations*, National Academy Press, Washington, D.C.

Di Stasi, L., Marchitto, M., Antolí, A. & Cañas, J. (2013), 'Saccadic peak velocity as an alternative index of operator attention: A short review', *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology* **63**(6), 335–343.
URL: *https: // linkinghub. elsevier. com/ retrieve/ pii/ S1162908813000741*

Dynamic Time Warping (2007), *in* 'Information Retrieval for Music and Motion', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 69–84.
  URL: *http://link.springer.com/10.1007/978-3-540-74048-3_4*

Gama, J., Foutch, B., Grundland, M. & Dodgson, N. (n.d.), 'CVD documentation'.
  URL: *https://rdrr.io/rforge/CVD/man/*

Hart, S. G. & Staveland, L. E. (1988), Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research, *in* 'Advances in Psychology', Vol. 52, Elsevier, pp. 139–183.
  URL: *http://linkinghub.elsevier.com/retrieve/pii/S0166411508623869*

Hess, E. H. & Polt, J. M. (1964), 'Pupil Size in Relation to Mental Activity during Simple Problem-Solving', *Science* **143**(3611), 1190–1192.
  URL: *http://www.sciencemag.org/cgi/doi/10.1126/science.143.3611.1190*

Hiscocks, P. D. & Eng, P. (2014), 'Measuring Luminance with a Digital Camera', *Syscomp Electronic Design Limited* p. 27.

Holmqvist, K. & Andersson, R. (2017), *Eye tracking: a comprehensive guide to methods and measures*. OCLC: 1057373387.

Human Performance Research Group, N. A. R. C. (n.d.), 'TASK LOAD INDEX NASA-TLX Paper and Pencil Pakage'.
  URL: *https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20000021488.pdf*

IBM, C. (2017), 'IBM SPSS Statistics'.

Industries, A. (2019), *Adafruit Industries, Unique & fun DIY electronics and kits*.
  URL: *https://www.adafruit.com*

ISO (2016), 'Road vehicles – Transport information and control systems – Detection- Response Task (DRT) for assessing attentional effects of cognitive load in driving.', *ISO 17488:2016.* .

Jung, H. S. & Jung, H.-S. (2001), 'Establishment of overall workload assessment technique for various tasks and workplaces', *International Journal of Industrial Ergonomics* **28**(6), 341–353.
  URL: *http://linkinghub.elsevier.com/retrieve/pii/S0169814101000403*

Kahneman, D. (2012), *Thinking, fast and slow*, Penguin Books, London. OCLC: 798805166.

Kahneman, D. & Beatty, J. (1966), 'Pupil Diameter and Load on Memory', *Science* **154**(3756), 1583–1585.
  URL: *http://www.sciencemag.org/cgi/doi/10.1126/science.154.3756.1583*

Kahneman, D., Beatty, J. & Pollack, I. (1967), 'Perceptual Deficit during a Mental Task', *Science* **157**(3785), 218–219.
  URL: *http://www.sciencemag.org/cgi/doi/10.1126/science.157.3785.218*

Kassner, M., Patera, W. & Bulling, A. (2014), Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction, *in* 'Adjunct Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing', UbiComp '14 Adjunct, ACM, New York, NY, USA, pp. 1151–1160.
URL: *http://doi.acm.org/10.1145/2638728.2641695*

Ltd, K. M. S. S. P. (2018), *CS-2000 SPECTRORADIOMETER*, Konica Minolta Sensing Singapore Pte Ltd.
URL: *http://sensing.konicaminolta.asia/products/cs-2000-spectroradiometer/*

Marshall, S. P. (1999), 'Method and apparatus for eye tracking and monitoring pupil dilation to evaluate cognitive activity'.
URL: *https://patents.google.com/patent/US6090051A/en*

Mehler, B., Reimer, B., Coughlin, J. F. & Dusek, J. A. (2009), 'Impact of Incremental Increases in Cognitive Workload on Physiological Arousal and Performance in Young Adult Drivers', *Transportation Research Record: Journal of the Transportation Research Board* **2138**(1), 6–12.
URL: *http://journals.sagepub.com/doi/10.3141/2138-02*

OpenCV (2019), *About - OpenCV*, OpenCV team.
URL: *https://opencv.org/about.html*

Palinko, O. & Kun, A. (2011), Exploring the Influence of Light and Cognitive Load on Pupil Diameter in Driving Simulator Studies, *in* 'Proceedings of the 6th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design : driving assessment 2011', University of Iowa, Olympic Valley-Lake Tahoe, California, USA>, pp. 329–336.
URL: *http://ir.uiowa.edu/drivingassessment/2011/papers/48*

Palinko, O., Kun, A. L., Shyrokov, A. & Heeman, P. (2010), Estimating cognitive load using remote eye tracking in a driving simulator, *in* 'Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications - ETRA '10', ACM Press, Austin, Texas, p. 141.
URL: *http://portal.acm.org/citation.cfm?doid=1743666.1743701*

Peirce, J. (2018), *PsychoPy*, Jonathan Peirce.
URL: *http://www.psychopy.org*

Peirce, J. W. (2007), 'PsychoPy—Psychophysics software in Python', *Journal of Neuroscience Methods* **162**(1-2), 8–13.

Pickup, L., Wilson, J. R., Norris, B. J., Mitchell, L. & Morrisroe, G. (2005), 'The Integrated Workload Scale (IWS): A new self-report tool to assess railway signaller workload', *Applied Ergonomics* **36**(6), 681–693.
URL: *http://linkinghub.elsevier.com/retrieve/pii/S0003687005000918*

Pignoni, G. (2018), Effects of ambient illumination on text recognition for UI development, *in* '2018 Colour and Visual Computing Symposium (CVCS)', IEEE, Gjøvik, pp. 1–6.
URL: *https://ieeexplore.ieee.org/document/8496507/*

Pignoni, G. & Komandur, G. (2019), Development of a quantitative evaluation tool of cognitive workload in field studies through eye tracking, *in* 'Lecture Notes in Computer Science', Springer.

Recarte, M. A. & Nunes, L. M. (2000), 'Effects of verbal and spatial-imagery tasks on eye fixations while driving.', *Journal of Experimental Psychology: Applied* **6**(1), 31–43.
URL: *http://doi.apa.org/getdoi.cfm?doi=10.1037/1076-898X.6.1.31*

Reimer, B., Mehler, B., Coughlin, J. F., Godfrey, K. M. & Tan, C. (2009), An on-road assessment of the impact of cognitive workload on physiological arousal in young adult drivers, *in* 'Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '09', ACM Press, Essen, Germany, p. 115.
URL: *http://portal.acm.org/citation.cfm?doid=1620509.1620531*

Rubio, S., Diaz, E., Martin, J. & Puente, J. M. (2004), 'Evaluation of Subjective Mental Workload: A Comparison of SWAT, NASA-TLX, and Workload Profile Methods', *Applied Psychology* **53**(1), 61–86.
URL: *http://doi.wiley.com/10.1111/j.1464-0597.2004.00161.x*

Rusnock, C. F. & Borghetti, B. J. (2018), 'Workload profiles: A continuous measure of mental workload', *International Journal of Industrial Ergonomics* **63**, 49–64.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0169814116301287*

Savitzky, A. & Golay, M. J. E. (1964), 'Smoothing and Differentiation of Data by Simplified Least Squares Procedures.', *Analytical Chemistry* **36**(8), 1627–1639.
URL: *http://pubs.acs.org/doi/abs/10.1021/ac60214a047*

Schafer, R. (2011), 'What Is a Savitzky-Golay Filter? [Lecture Notes]', *IEEE Signal Processing Magazine* **28**(4), 111–117.
URL: *http://ieeexplore.ieee.org/document/5888646/*

Stanley, P. (1995), 'The effect of field of view size on steady-state pupil diameter', *Ophthalmic and Physiological Optics* **15**(6), 601–603.
URL: *http://linkinghub.elsevier.com/retrieve/pii/027554089400019V*

Tossel, K. (n.d.), *Libuvc*.
URL: *https://ken.tossell.net/libuvc/doc/index.html*

Tsang, P. S. & Velazquez, V. L. (1996), 'Diagnosticity and multidimensional subjective workload ratings', *Ergonomics* **39**(3), 358–381.
URL: *http://www.tandfonline.com/doi/abs/10.1080/00140139608964470*

W3C (2018), *G17: Ensuring that a contrast ratio of at least 7:1 exists between text (and images of text) and background behind the text*, W3C.
URL: *https://www.w3.org/TR/WCAG20-TECHS/G17.html*

Watson, A. B. & Yellott, J. I. (2012), 'A unified formula for light-adapted pupil size', *Journal of Vision* **12**(10), 12–12.
  URL: *http://www.journalofvision.org/lookup/doi/10.1167/12.10.12*

Wierda, S. M., van Rijn, H., Taatgen, N. A. & Martens, S. (2012), 'Pupil dilation deconvolution reveals the dynamics of attention at high temporal resolution', *Proceedings of the National Academy of Sciences* **109**(22), 8456–8460.
  URL: *http://www.pnas.org/cgi/doi/10.1073/pnas.1201858109*

Wüller, D. & Gabele, H. (2007), The usage of digital cameras as luminance meters, San Jose, CA, USA, p. 65020U.
  URL: *http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.703205*

Čegovnik, T., Stojmenova, K., Jakus, G. & Sodnik, J. (2018), 'An analysis of the suitability of a low-cost eye tracker for assessing the cognitive load of drivers', *Applied Ergonomics* **68**, 1–11.
  URL: *https://linkinghub.elsevier.com/retrieve/pii/S0003687017302326*

A   Appendix

A.1 Appendix A (Nasa TLX)

1. **Subject Instructions: Rating Scales**

We are not interested in assessing your performance but the **experience** you had during the task conditions. In the most general sense we are examining the **"Workload"** you experienced.

The factors that influence your experience of workload may come from the task itself, your feelings about your own performance, how much effort you put in, or the stress and frustration you felt.

Physical components of workload are relatively easy to conceptualise and evaluate. However, the mental components of workload may be more difficult to measure.

Workload is an individual experience, caused by many different factors, we would like you to evaluate several of these factors individually.

This set of six rating scales was developed for you to use in evaluating your experiences during different tasks. Please read the descriptions of the scales carefully.

On the rating scales sheet evaluate the task by putting an "X" on each of the six scales at the point which matches your experience.

Consider each scale individually. Your ratings will play an important role in the evaluation being conducted, thus. your active participation is essential to the success of this experiment and is greatly appreciated by all of us.

2. **Subject Instructions: Sources-Of-Workload Evaluation**

Through out this experiment rating scales are used to assess the **experience** you had during the task, first it is necessary to assess the relative importance of different factors determining the experienced workload.

The procedure is simple: You will be presented with a series of pairs (for example. Effort vs. Mental Demands) and asked to choose which of the items was more important to your experience of workload in the task{s) that you just performed.

Circle the component that represents the more important contributor to workload for the specific task you performed.

Don't think that there is any correct pattern; we are only interested in your opinions.
If you have any questions, please ask them now. Otherwise. start whenever you are ready. Thank you for your participation.

RATING SCALE DEFINITIONS

Title	Endpoints	Descriptions
MENTAL DEMAND	*Low/High*	How much mental and perceptual activity was required (e.g.. thinking. deciding. calculating. remembering. looking. searching. etc.)? Was the task easy or demanding. simple or complex. exacting or forgiving?
PHYSICAL DEMAND	*Low/High*	How much physical activity was required (e.g.. pushing. pulling. turning. controlling. activating. etc.)? Was the task easy or demanding. slow or brisk. slack or strenuous. restful or laborious?
TEMPORAL DEMAND	*Low/High*	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
PERFORMANCE	*good/poor*	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
EFFORT	*Low/High*	How hard did you have to work (mentally and physically) to accomplish your level of performance?
FRUSTRATION LEVEL	*Low/High*	How insecure. discouraged. irritated. stressed and annoyed versus secure. gratified. content. relaxed and complacent did you feel during the task?

Frustration	or	**Effort**
Performance	or	**Mental Demand**
Performance	or	**Temporal Demand**
Mental Demand	or	**Effort**
Mental Demand	or	**Physical Demand**
Effort	or	**Physical Demand**
Frustration	or	**Mental Demand**
Effort	or	**Performance**
Temporal Demand	or	**Frustration**
Temporal Demand	or	**Effort**
Physical Demand	or	**Frustration**
Performance	or	**Frustration**
Physical Demand	or	**Temporal Demand**
Physical Demand	or	**Performance**
Temporal Demand	or	**Mental Demand**

Subject ID: _____ Task ID: _____

RATING SHEET

MENTAL DEMAND

Low   High

PHYSICAL DEMAND

Low   High

TEMPORAL DEMAND

Low   High

PERFORMANCE

Good   Poor

EFFORT

Low   High

FRUSTRATION

Low   High
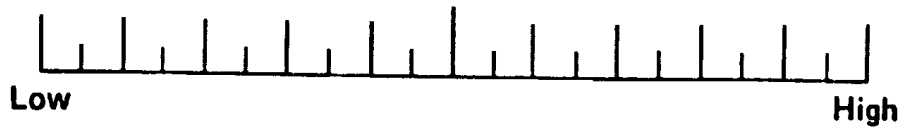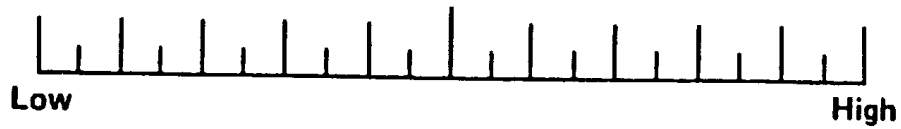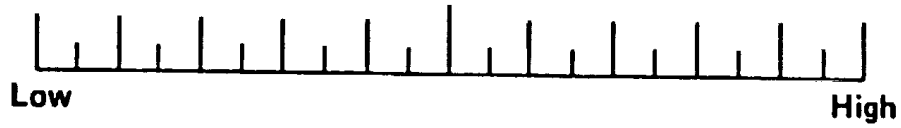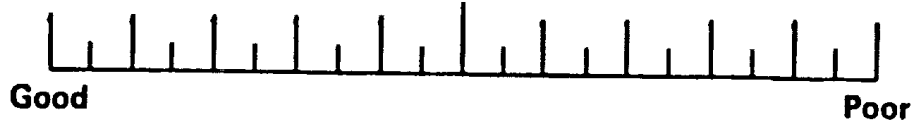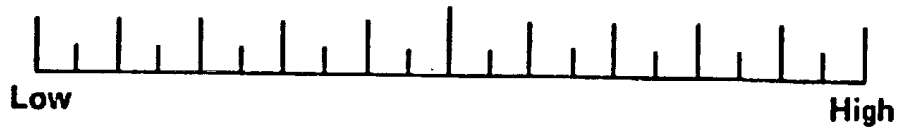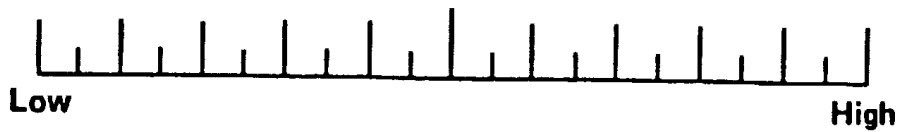
Subject ID: _____   Task ID: _____

WEIGHTED RATING WORKSHEET			
Scale Title	Weight	Raw Rating	Adjusted Rating (Weight X Raw)
MENTAL DEMAND			
PHYSICAL DEMAND			
TEMPORAL DEMAND			
PERFORMANCE			
EFFORT			
FRUSTRATION			

Sum of "Adjusted Rating" Column = _____

WEIGHTED RATING =
[i.e., (Sum of Adjusted Ratings)/15]

19

Subject ID: _____   **Date:** _____

SOURCES-OF-WORKLOAD TALLY SHEET		
Scale Title	*Tally*	*Weight*
MENTAL DEMAND		
PHYSICAL DEMAND		
TEMPORAL DEMAND		
PERFORMANCE		
EFFORT		
FRUSTRATION		

Total count = _____

(NOTE - The total count is included as a check. If
the total count is not equal to 15, then something has
been miscounted. Also, no weight can have a value
greater than 5.)

A.2   Appendix B (Publication)

Development of a quantitative evaluation tool of cognitive workload in field studies through eye tracking

Giovanni Pignoni[1][0000−0003−0730−371X] and Sashidharan Komandur[1]

Department of Design, Norwegian University of Science and Technology, NTNU in Gjøvik, NO-2802, Norway `postmottak@ntnu.no`
https://www.ntnu.edu/design

Abstract. Eye tracking is mainly employed as mean of tracking visual attention of an observer/operator. Still, eye tracking is also capable of recording a wider variety of data such as traces of mental workload. Pupil diameter have been validated as such measure. Most of the studies that have validated this are in laboratory conditions, where the perceived luminance (measured in candela per square meter) can be controlled. Luminance affects the pupil diameter as well; this means if the pupil diameter varies for an operator/observer in field conditions it cannot be accurately determined if the change in the pupil diameter is due to mental workload alone. Although there are some studies, which have attempted to simultaneously account for the contribution of the change in pupil diameter due to luminance and mental workload, not many have attempted to account for this in field conditions for safety-critical systems such as a helicopter or a maritime ship bridge. In this study as a first step, we define a method to measure luminance while tracking the gaze point. We will record eye-tracking data simultaneously recording the video feed of the field of view of the operator/observer. We will use the video feed to estimate the luminous flux from the point of view of the subject. We will be collecting this data from a helicopter pilot and his co-pilot during an actual operation (e.g. transportation of personnel and carrying a payload for an electrical power provider company in Norway or Sweden). We will also be collecting data from a navigator and his first officer in a high-speed marine craft of the Norwegian navy. We will also be collecting subjective data using paper-based tools such as NASA-TLX in addition to a conventional video recording of the scene of activity and handwritten notes of observation for validation purposes. We will also capture mental workload data from a few other objective sources such as heart rate variability (ECG). We expect to clearly define an approach to separately account for the effect of mental workload independent of the impact of changing light conditions in field situations for safety-critical systems. This includes a mathematical model that we innovate based on other mathematical models that are already available in the literature.

Keywords: Eye Tracking· Pupillometry· Cognitive Workload· Field Study· Maritime Usability

1   Introduction

Pupil dilation is an important metric for assessment of mental workload [13] [19], especially in a safety critical system [**?**] such as a helicopter (or) ship's bridge or driving [**?**]. In these environments, the light condition fluctuates dramatically. Since changing light conditions can also impact pupil dilation, it is necessary to separate the effect of the mental workload from the effect of the changing light conditions to be able to utilise it reliably to evaluate the risk profile.

As discovered through the literature review, currently, there is no open and validated method to measure cognitive workload in a field condition trough pupillometry. The only commercially available method is proprietary [24] ad thus closed source, it disconnects the researcher from the ability to adapt the tools to specific research questions and have a deep understanding of the variables at play [15]. Moreover, no methods have been validated for use with low-cost eye trackers, which would enable affordable data gathering, including collaborative studies with multiple eye trackers.

The relation between cognitive workload and pupillary responses has been assessed as back as 1964 [13], here Hess and Polt measured changes in the pupil size of a subject during the resolution of "simple multiplication problems" and observed a link between the pupillary response and the difficulty level. Likewise, D.Kahneman et al. [19] investigated the correlation between task-evoked pupillary diameter and memory intensive tasks; reporting different pupillary responses from the learning and report/recollect phases as well as variations directly related to the task difficulty. These initials results were validated for a variety of "intensive cognitive tasks", including language, writing, listening, speech and the solution of mathematical problems [20]. The psycho-physiological studies on workload and the pupillary response are often limited by three main factors:

- Perceived luminance, as the variable with the more significant influence on the pupil diameter, it can mask the influence of cognitive workload. It is common to keep luminance as controlled conditions to isolate its effect.
- Real-time tracking is only possible tough a limited number of the reviewed ed methods. Online tracking of cognitive workload, [38] and [24] requires high-temporal-resolution, as well as the control of the environmental variables, including the estimate of a continually changing baseline value. Lack of such features limits the study to the evaluation of well defined/separated tasks.
- Open source. The only method that currently provides a solution to both the previous limitation is tied to patented technology, only limited documentation is available on the underlying method [24], and this makes it impossible for independent researchers to reuse, adapt and improve on such systems.

1.1   Research questions

This paper is part of an ongoing, work-in-progress, research and will present only preliminary results to the following research questions:

- How to reliably measure luminance from the POV of a subject using a small calibrated video camera?
- How to calculate the baseline pupil size for a visual stimulus and use it to infer the cognitive state of the subject then?
- How to scale this method in field conditions (where luminance variate in an unpredictable manner)?

1.2   Planned contributions

- Open source code, available in a public repository that can work on files generated by at least one common eye-tracker vendor.
- Thesis report and paper with validation data.

1.3   Existing research

Cognitive workload. Mental workload can be evaluated through a variety of methodologies [4] [33]:

- Subjective-empirical measures of perceived effort as rated by the subject.
- Performance of the subject in a controlled task.
- Physiological indices of cognitive state.

Subjective reports such as questionnaires and multidimensional ratings (e.g. the NASA-TLX [12]) are indirect means of evaluation of the perceived workload. They are usually easy and cheap to administer but have several limitations [36]. As post-facto evaluations, relying on the personal impression and memory of the subject, they do not track a change over time and are therefore difficult to use for the identification of specific peaks on cognitive workload. The NASA-TLX questionnaire is a standardised assessment tool of cognitive workload. It employees a "multi-dimensional rating scale" measuring six parameters to give an estimate of the overall task workload: mental, physical, temporal, frustration, performance and effort. Like other forms of self-report, it doesn't record changes in cognitive load over time (multiple questionaries can be used to assist a complex task if divisible in subtasks). Other subjective workload measures are: Multiple Resources Questionnaire (MRQ) [2], Subjective Workload Assessment Technique (SWAT) [32] , Overall Workload Level (OWL) [18] and Integrated Workload Scale (IWS) [29].

Performance-based measures of mental workload indirectly measure the cognitive state of a subject through the execution of a standardised task. Changes (speed, accuracy, response time) in the execution of the secondary task can be interpreted as a difference in cognitive/visual workload. The ISO defined Detection Response Task (DRT) [17] is an example of a performance-based method, T.Cegovnik et al. [40] used a tactile DRT to validate the use of a Tribe eye tracker to assess changes in the cognitive load of the subjects using oculography and pupillometry. The DTR estimates the cognitive load trough response rate and miss-rate of the response task: a stimulus is delivered through a vibrator

attached to the subject in a random sequence; the test measures the response time (time needed to press a button attached to the steering wheel in a driving simulator). The use of this class of methodologies is to be planned considering the effect of the controls tasks on the main task as well as the low temporal resolution of the events that can be measured. Moreover, the relationship between cognitive workload and task performance is not linear and follows an inverted U-shape as defined in the Hebb-Yerkes-Dodson Law [40]. Both overload under-load can, therefore, result in decreased performances making the measure potentially unreliable if not paired with subjective data.

Physiological indices indirectly connect a measurable psychophysical parameter to an expected mental workload. Heart rate, respiratory rate, galvanic skin response, brain activity (EEG, fRMITCD), as well as eye activity [3] are parameters that over the years have been used to measure the mental workload. Modern eye tracking has an intrinsic advantage of being unobtrusive, and less impending that most of the other aforementioned techniques and could be a reliable instrument for over time monitoring of the mental workload [4]. Still, psychophysiological measures have several limitations [40] when applied in field conditions. In most of the studies mentioned beforehand, the experimental design included two or more tasks with different levels of difficulty indirectly estimating the workload required for each of the tasks. The physiological and neurological models employed in the psychophysiological methods have to be specially designed and trained to fit a particular task-evoked neural activity. It is, therefore, difficult to compare the results to a generalised measure of workload.

Blinks Eyeblink duration and rate have been identified as an alternative metric for visual workload [1] and [31]. Unfortunately, this metric reliability is limited, as the blink rate can be influenced, in an opposite manner, by both the mental workload and visual workload.

Pupil In the Handbook of Psychophysiology [3, p. 443], J.T Cacioppo et al. defines the pupillary system as a "dually innervated organ". The pupil size is determined by the concurring action of a parasympathetically innervated constricting muscles and sympathetically innervated radial dilator muscles. The parasympathetic activity is dominant, responding to light reflexes, and determine the varying pupil size baseline, the sympathetic activity is instead connected to behavioural and stress contexts and can be used as a psychophysiological parameter of cognitive activity.

Task-evoked pupillary response. Palinko et al. [28] estimated the driver's cognitive load from pupil size measurements finding that it the pupillary response correlates with the measured driving performances, and this as similar studies seems to confirm the reliability of pupillometry as a measure of cognitive workload. However, the analysis is limited to a simulated task with low variability between target luminance. Palinko et al. [28] introduced a pupillo-

metric cognitive load measure for real-time cognitive load changes (every several seconds).

Light and cognitive load effects on pupil diameter. Palinko et al. [27] follows up the previous study with a proof of concept of the possible separation between cognitive and ambient light components of pupil dilatation. The study was conducted using a driving simulator as a controlled environment for both an Aural Vigilance Task an Illumination Task (with different brightness targets) and a combined validation task. The study results show that it should be theoretically possible to model the psychophysical functions of the pupillary response over time to light stimuli and shows the measured trend over time. It also notes how the transitions bright/dark/bright are not equal as different muscle groups are involved in the contraction and dilation movements. The bright light reaction is quick "to protect the retina from overexposure", while the reaction to darkness is slower and gradual. The psychophysical function to predict an expected baseline pupil diameter should, therefore, take into account multiple parameters, current light level, previous light level, the rate of change, as well as age, and target. The study concludes that it is possible to discern the effects of luminance and cognitive load on pupil diameter and that the "proof of concept" predictor works in the limited experimental setting.

Discern between mentally and visually workload. M.A. Recarte et al.[30] validated the use of pupillary response as workload index in a field scenario ignoring the effect of illumination changes as the variable was impossible to control. The data they collected shows consistent results across the different driving task (no task, verbal task and visual task) collected during multiple driving session, to such a degree that cannot be explained by the sole different lighting conditions.

Unified formula for light-adapted pupil size. Since the pupil diameter can be deconstructed as the result of multiple concurring factors, in order to correctly differentiate the cognitive workload from the pupillary light response, it can be useful to compute the expected pupil diameter for a given brightness condition and use the resulting value as a baseline value upon which calculate the cognitive-driven component of the measured pupil size. In a recent paper, A.B. Watson et al. [37] (NASA Ames Research Center and University of California) have reviewed seven different published psychophysical functions defining the relation between target luminance (cd/m2) and expected pupil diameter. In the same paper, they also published a newly developed unified formula. The calculated baseline would work in the range of 2 to 8 mm, the reliability of the unified formula have to be tested to ensure that the little variability rage of the pupil size provoked by cognitive workload is preserved ($<+$-1mm)[27]. The unified formula [37] is valid only for a light-adapted condition with stable illuminant and point of view (PoV) as it doesn't account for the adaptation state

or the "pupillary unrest" (low-frequency random fluctuation in the range of 0.02 to 2.0 Hz and amplitude within +-0.25 mm).

Independent variables in the unified equation:

- Luminance.
- Age (The maximum pupil diameter, as well as total range, declines as the age grows).
- Field diameter (deg).
- Number of eyes stimulated, the final diameter is dependent on the number of eyes that are adapted to the light condition they defined the "Effective corneal flux density" (the variable controlling the effective pupil diameter) as dependent on the number of eyes (attenuated by a factor of 10 for one eye).

$$F = LaM(e)$$

(F)Flux density as the product of (L)luminance, (a)area, and (M(e))monocular effect.

Wavelet analysis. S.P. Marshall et al. [24] describes a technique to identify the origin of a recorded pupillary response that works independently from the target luminance. The procedure employs wavelet analysis to identify the dilation reflexes of the subject's pupil. She explains how the reflexes can be differentiated as the pupil have different responses to light and psychosensorial stimulus. In a steady light, the pupil shows an irregular pulsation (light reflex) provoked by the interaction of the circular contracting muscles (agonist) and antagonist radial muscles act as the antagonist and are inhibited from dilating the pupil. A cognitive workload provokes a different waveform as both circular and radial muscles dilate the pupil creating a brief peak. This would imply that the cognitive workload is measured as the frequency and intensity of such events and not as a steady dilatation of the pupil (for the duration of the load), but it is unclear how this method would perform in a field condition, with highly variable ambient luminance.

An application of this technology is explained in a study conducted by S.P. Marshall for the US Navy. She applied the patented metric of Index of Cognitive Activity (ICA) [25] thanks to a networked system that is set up to record the cognitive workload for multiple team members during a collaborative task. The study assessed the performance of a three-person team in a simulation system, and the effort to overcome mission-related problems. A similar study, a collaboration between NASA Ames and EyeTracking, Inc. used the ICA and eye metrics to detect the difference between low and high fatigue states [26].

Machine learning for pupillometry. S.M. Wierda et al. [38] and the related work of A. Ferscha et al. [7] represent a different approach to the problem of the indirect assessment of mental workload. As the response time of the pupil to a mental workload event is too slow (several seconds) to be used as a real-time measure, it can be used directly only as an average over time. This makes it suitable to evaluate lengthy tasks that have a reasonably constant load in cognitive workload (at least several seconds). These two studies show a proof of concept of how to obtain an high-temporal-resolution (c.a. 10 Hz) tracking of the cognitive processes through deconvolution. The aim of real-time cognitive workload measurement gains value in the context of the implementation of a real-time feedback loop in the interaction design of a system (e.g. a system able to respond to different cognitive states of the user).

S.M. Wierda et al. [38] fixed the distribution of "attention impulses" every 100ms defining the output's temporal resolution. Employing a model of the "Task-evoked pupil impulse response," it reconstructs the intensity of the attention impulse that provoked the measured pupillary response. A. Ferscha et al. [7] further developed the concept through machine learning for better performances without the need of a fixed temporal resolution of the cognitive impulses. To reduce the effect of incident light A. Ferscha et al. [7] used the average illumination in the subject's field of view analysing the eye tracker camera stream. The technique they used is possibly still insufficient to adapt the technology to a field study with a highly variable illumination. In the described implementation a luminance change more significant than the set threshold would trigger a suspension of the tracking, this state is then maintained until the condition is stable again and a new baseline can be calculated. A similar solution was implemented to filter out blinks. The dynamic baseline is computed through a series of threshold and doesn't adjust for small changes in target luminance.

Illuminance measurement using a digital camera. Luminance as a measure needed to dynamically estimate the pupil size of a subject, candela per square meter cd/m2, is the quantity of light radiating from a source. An illuminance meter is an expensive and bulky device, in more than one instance this has resulted in attempts to use a camera as cheaper and more flexible alternative [14] and [39].

A digital sensor is at its core an array of Illuminance sensors. Each pixel measures the number of photons hitting the photoelectric surface. The presence of a Bayer filter for colour photography makes it so that to reconstruct the information form the entire visible spectrum multiple pixels have to be analysed at the same time. Each pixel in the final image has reconstructed values from the neighbouring pixels for all the RGB channels and in itself would be sufficient to reconstruct the illuminance of the scene, in order to reduce noise and gain reliability multiple pixels should be used, the number of pixels used is effectively the field of view of the instrument [14]. The formula proposed by P.D Hiscocks [14] has been optimised for a DSLR camera, not all the parameters are accessible when using an embedded digital video camera.

Parameters:

- Pixel value (0-255 for an 8bit monochrome image).
- Shutter Speed (In a video camera, this is limited by the frame rate, e.g. 1/30s for a 30fps camera) and aperture or focal ratio.
- Iso or film speed.
- Camera Constant (The calibration constant for a specific camera model that has to be determined with a known instrument).

D. Wuller [39] suggest a different model, reversing the colour processing of the camera, first from gamma-compressed RGB to linear RGB and then from linear RGB to CIE XYZ, extrapolating then $y(\lambda)$ as the relative luminance (luminance as defined by the luminosity function, reproducing the spectral luminous efficiency of the human eye). The author notes that to access the linear response of an image sensor the correct inverse gamma has to be applied and that this could deviate from the standard 2.2 (sRGB), the relative luminance can then be converted to luminance through a linear relation specific for a particular sensor/camera settings combination.

Maritime Usability and SA Endsley et al. [5] defines situational awareness (SA) as "the perception of the elements in the environment within a volume of space and time, the comprehension of their meaning and the projection of their status in the near future". Low SA has been found to be one of the primary sources of human error in safety-critical systems [34]. Real-time monitoring of SA seems possible through the analysis of the subject visual attention aided by a variety of eye tracking data such as:

- Fixation duration: length of fixations (e.g. time spent on a single target without movement).
- Fixation rate: average number of fixations in a unit of time.
- Dwell time: the sum of all the fixation time in a single area of interest.
- Saccadic main sequence: the relation between the saccadic duration and magnitude and between peak velocity (PV) and magnitude [4], as both PV and duration increase with the magnitude.
  - Saccadic duration: the period between two positions of the fovea.
  - Saccadic magnitude: the magnitude of the saccadic movement (angle).
  - Peak saccadic velocity: highest velocity reached during saccades deg/sec.

Use of SAGAT The Situation Awareness Global Assessment Technique (SAGAT), is a global tool developed to assess SA [6]. "A simulation employing a system of interest is frozen at randomly selected times, and operators are queried as to their perceptions of the situation at that time. The system displays are blanked, and the simulation is suspended while subjects quickly answer questions about their current perceptions of the situation. As a global measure, SAGAT includes queries about all operator SA requirements, including Level 1 (perception of data), Level 2 (comprehension of the meaning) and Level 3 (projection of the

near future) components. This includes a consideration of system functioning and status, as well as relevant features of the external environment. SAGAT queries allow for detailed information about subject SA to be collected on an element by element basis that can be evaluated against reality, thus providing an objective assessment of operator SA."

Ikuma et al. [16] compared different standard human factors measurement tools: workload ratings (SWAT and NASA-TLX) and Situation Awareness Global Assessment Technique (SAGAT). Eye tracking was also used to analyse the gaze path of the participants during the simulation, "the percentage of time spent looking at different areas of the screen during steady-state periods differed among workload levels". This study only looks at a small number of areas of interest (AOIs) on the interface, to infer the visual attention of the subjects for different areas of the interface. The usability of on-board interfaces on High-Speed Craft (HSC) has been assessed through the application of eye tracking technology [9]. The cognitive workload and SA of the crew of a military HSC in littoral waters were selected as of interest because of the combination of high-speed navigation and the need to navigate outside established routes. With particular interest on the role of the navigator [8] and its use of the onboard interface "Route Monitor Window". Hareide et al. [9] collected data from both field and simulator activities using the Tobii Pro Glasses 2 Eye Tracker. The methodology of the study tried to account for the difference in the environment and datasets between the simulator and field conditions. The Author followed up a mid-life update of the interface [11] [10] , with further validation of the redesigned interface for the primary objective of increased navigator attention dedicated to the "outside" Area of interest opposed to the various interfaces. Eye trackers were in this case used as to indirectly evaluate situational awareness of the navigator through quantisation of the time spent on the interface rather than observing the environment. Hareide et al. [10] apply the concept of dwell time, look-backs and Backtracks to the analysis of AOIs:

- Look-backs (returns, refixation) are saccades landing in an AOIs already visited. The analysis of a look-back can point to a variety of concurring factors: memory failure, confusion on the function of a command/element, the difficulty of content understanding and intrinsic importance of the information present in an AOI.
- Backtracks are calculated on the specific sequence of saccades and are a sudden (inverted gaze direction) rapid eye movement back to a just visited AOI. Confusion or uncertainty, changes in goals, a mismatch between the users' expectation and interface layout.

The author shows how eye tracking data can be used to guide the development of a GUI through the analysis of areas of interest and gaze behaviour, but also notes several limitations in the use of the eye tracker that need to be considered not to influence the behaviour of the user group. This includes the thickness of the eyepiece frame, creating a visible "frame of vision", unwanted reflection and glares on the protective glass, difficulty using the binoculars in conjunction with the trackers and unfavourable lighting conditions.

2   Methods

2.1   Measure of luminance from POV

This is a work in progress. his research starts with the development of the necessary tools. Even though it is theoretically possible to use a video camera as luminance meter, the reliability and accuracy of this technique will depend significantly on the software and hardware. The calibration and validation of the equipment will be done with a know good instrument in the Norwegian Colour and Visual Computing Laboratory. A colour checker will be measured with both the Konica Minolta CS-2000 [23] spectroradiometer and the World Camera mounted on the Pupil lab eye tracker (Pupil Pro). The data can then be used to calibrate the World Camera as a rudimentary luminance meter. A colour checker illuminated by a diffuse light at a variable intensity will be measured through both the Pupil Pro and the spectroradiometer. Different combinations of illumination and exposure settings on the software are required to model the sensor response.

The Pupil lab [21] software in his current version (1.10.20) does not dynamically save the exposure settings during the recording. Libuvc [?] (cross-platform library for USB video devices) is used to receive the video stream and communicate with the two cameras. Libuvc supports either getting or setting the exposure value and should allow retrieving the current exposure data during the recording. The lack of support of these functionalities in the Pupil lab software makes it impossible to use automatic exposure as the calibration values would change during the recording in an unpredictable manner. Using a fixed manual exposure is possible but severely limits the maximum dynamic range of the light meter.

Alternatives to the use of the camera, to simulate a scenario in which the aforementioned limitation does not apply, which would be easily reachable with some interest from the developers, is to use an external light meter mounted on the eye-tracker. This would provide a measurement that is not bound to the limited dynamic range of the camera with the drawback of having two disconnected data streams.

Instrumentation The Konica Minolta CS-2000 spectroradiometer [23] is a high precision polychromatortype spectroradiometer, it will allow measurement on a vast range of luminance (0.3 to 500,000 cd/m2) with a $\pm 2\%$ accuracy. The Pupil Pro [21] and [22] World Camera is mounted just above the subject eye, facing outward. The camera offers different combinations of resolution and frame-rate 1920x1080 @30fps, 1280x720 @60fps, 640x480 @120fps covering a FoV of 60 or 100 degrees diagonally (depending on the lens).

2.2   Calculate the baseline pupil size

The measure of luminance from the POV will be connected to the unified formula for light-adapted pupil size developed by [37]. Two elements will need validation:

- the accuracy and precision of the unified formula.
- the different possible methods to convert the input from the camera to the correct input values for the formula.

The [37] unified formula is based on a standard procedure involving a defined stimulus: the observer is shown a bright circle on a dark background. The size (degrees of field of view) and luminance (cd/m2) of the circle determine the corneal flux density (i.e. the product of luminance and subtended area) as defined by [35]: $D = 7.75 - 5.75[(F/846)0.41/((F/846)0.41 + 2)]$ "where D is the pupil diameter (mm), and F is the corneal flux density (cdm-2deg2)".

The model implies that at its core the pupil control mechanism reacts as a 'flux integrator', following an S-shaped curve.

An image from the camera has to be used to evaluate the flux density or to indirectly convert the image into a corresponding standardised stimulus (i.e. a circle on a dark background). The most promising approach is to consider the average luminance on the camera sensor (or external light sensor) as equal to the luminance of a standardised stimulus (bright circle) as wide as the entire field of view (FoV) (120-190°).

This assumption ties the precision of the calculated pupil size to how well the camera FoV matches the user FoV). To test the quality of the model NTNU students and staff (age from 20 to 50) will be recruited for validation in a controlled environment.

Validation The procedure will refer to the methods used by Palinko et al. [27]. It will be divided into three parts:

- No-load - variable light-adapted state. The participants will be sitting in a dark room looking at a selection of projected images; the sitting position will be adjusted to maintain a constant field of view and distance from the screen. The projected images will include standardised stimuli as well as more complex images (e.g. outdoor naturalistic scenery). Each image will be represented for several seconds to let the pupil reach an adapted state (c.a. 15s). Each image will include a focal point ad the participants will be asked to stare at the focal point.
- Load - static light-adapted state. With a constant ambient luminance (e.g. grey image projected), the participants will be asked to perform an Aural Vigilance Task (AVT), as in [27]. The task involves listening to a voice counting from 1 to 18, repeated multiple times. Every 6th number (6, 12, and 18) might contain errors (i.e. another number is replaced to the correct sequence). The participants would have to perform an action such as pressing a button when they detect an error. The task should induce an increased cognitive workload near every 6th number. The location of the error should be randomly selected for each session.
- Load - variable light-adapted state.
  The same AVT task is repeated but with variable standardised images being projected.

2.3   Measure of luminance from POV

Subjects The subjects for the final session will be recruited as cadets of the Royal Norwegian Naval Academy (RNoNA) and will require access to the training vessel (Kvarven). The crew of a training vessel includes navigator, assistant, helmsman and training instructor. During a study session, the vessel would also include one to three researchers to set up and record the data. Depending on the availability of multiple eye trackers both the navigator and helmsman could participate in the experiment for each session.

Procedure The setup for each session will include:

- Introduction to the research and signature of the informed consent.
- Application of the eye tracker (glasses and recording device).
- Calibration of the eye tracker.
- Reference measurements of ambient illumination.
- The debriefing will include a short interview and the NASA-TLX questionnaire as a further reference of the cognitive workload.

Two researchers should be present on board at any time. Between each session, up to 30 minutes will be required for cleaning of the instrumentation, download of the test data and recharge of the various batteries.

The task aims to highlight different levels of cognitive workload. The navigation task should be repeatable Fig. 1, and it should last less than one hour (not including the setup and debriefing) and should include a mix of low and high workload for the subjects: E.g. Steady navigation - change of course - steady navigation.
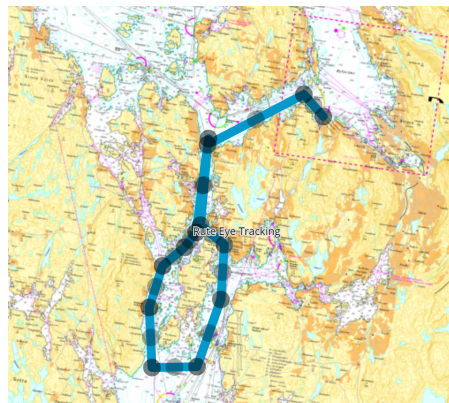

Fig. 1. The course suggested by Odd Sveinung Hareid from Laksevåg (Bergen)

3   Expected Results

3.1   Proof of concept

This is a work in progress, at the time of writing, the research is still in the initial exploratory phase and should be completed by the end of May 2019. The literature review helped in defining a path to follow in order to develop the necessary tools (the pupil baseline calculation), but there are inherent risks and challenges in the generalisation of findings that were initially only meant for controlled condition/laboratory study. A series of non validated tests is being carried out to determine:

- weather the camera-based luminance meter works well enough for a reliable field application (limited dynamic range of the camera, limited bit depth, the difference in the camera FOV compared to the subject FOV).
- Weather pupil baseline calculation is precise to such a degree not to mask the cognitive workload.
- Weather the chosen eye tracker can operate in field conditions.
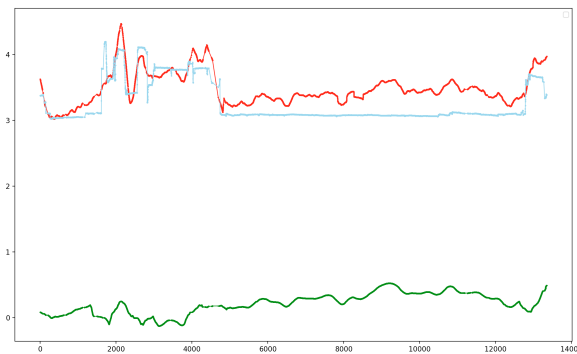
Fig. 2. This sample data output was recorded from a user sitting in front of a laptop in a dark room. The green line is the pupil size (mm) as measured by the eye tracker; the red line is the baseline as calculated from the video data and the blue line is the difference between the two. The blue will ultimately represent the cognitive workload. In this sample, the middle "steady" part has been measured on a subject performing an IQ test. The horizontal axe, time, is expressed in video frames at 30fps.

A series of artefacts have been identified in the sample data collected:

- Pupillary overshoot, the model of the pupil size is specific to adapted state and doesn't account for the pupil natural overshoot that can be observed when a rapid change in luminance occurs Fig. 2.

- Pupillary unrest, in the form of low-frequency random fluctuation in the range of 0.02 to 2.0 Hz and amplitude within +-0.25 mm Fig. 2 and Fig. 3.
- Incorrect measured pupil range; the pupil size is calculated from the video image in pixels to an estimated mm by the 3d model. different calibration of the pupil camera (distance from the eye) can bring the measured range outside the unified formula unified formula [37] range (c.a. more than 2mm and less than 8mm).
- Luminance outside the camera dynamic range, this is visible in the second plot, in this case, an outdoor recording ended inside a building, the completely black image from the video produces an unreliable luminance reading Fig. 3.
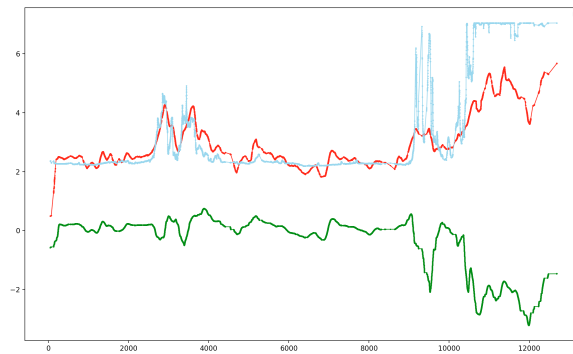
Fig. 3. This outdoor session shows the limitations of the camera if used with a fixed exposure.

During the test multiple data sources will be combined including: Heart rate variability, POV camera, eye tracking camera all will be used to substantiate the measure of cognitive workload. Dates and travel The number of sessions depends on the availability of cadets, five to ten subjects would be a good result. Given the nature of the experiment weather and ambient illumination conditions should be kept constant within a reasonable range. This could require the spread of the study over multiple days.

To account for the limitations of the eye tracker, it would be advisable to plan a portion of the sessions after dawn, (lower the contrast between the user interface inside the cabin and the outside).

4   Conclusion

The development of the necessary tools is in progress and will hopefully end as a refined proof of concept and validation of the method with the intent of attracting the interest of developers to consolidate the application. The validation that will be attempted as part of the research will be by no mean be exhaustive, it is expected that the interest surrounding the measure of the cognitive workload will result in a variety of experiments on the topic, to further explore the benefits and limitations of the developed methods.

The choice of a camera as a luminance meter could severely limit the accuracy of the method but would allow a variety of head-mounted eye trackers to be used for cognitive workload studies without the need of any extra hardware.

References

1. Benedetto, S., Pedrotti, M., Minin, L., Baccino, T., Re, A., Montanari, R.: Driver workload and eye blink duration. Transportation Research Part F: Traffic Psychology and Behaviour **14**(3), 199–208 (May 2011). https://doi.org/10.1016/j.trf.2010.12.001, http://linkinghub.elsevier.com/retrieve/pii/S136984781000094X

2. Boles, D.B., Adair, L.P.: Validity of the Multiple Resources Questionnaire (MRQ). Proceedings of the Human Factors and Ergonomics Society Annual Meeting **45**(25), 1795–1799 (Oct 2001). https://doi.org/10.1177/154193120104502508, http://journals.sagepub.com/doi/10.1177/154193120104502508

3. Cacioppo, J.T., Tassinary, L.G., Berntson, G. (eds.): Handbook of Psychophysiology. Cambridge University Press, Cambridge, 3 edn. (2007). https://doi.org/10.1017/CBO9780511546396, http://ebooks.cambridge.org/ref/id/CBO9780511546396

4. Di Stasi, L., Marchitto, M., Antolí, A., Cañas, J.: Saccadic peak velocity as an alternative index of operator attention: A short review. Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology **63**(6), 335–343 (Nov 2013). https://doi.org/10.1016/j.erap.2013.09.001, https://linkinghub.elsevier.com/retrieve/pii/S1162908813000741

5. Endsley, M.R.: Design and Evaluation for Situation Awareness Enhancement. Proceedings of the Human Factors Society Annual Meeting **32**(2), 97–101 (Oct 1988). https://doi.org/10.1177/154193128803200221, http://journals.sagepub.com/doi/10.1177/154193128803200221

6. Endsley, M.R.: Direct measurement of situation awareness: Validity and use of SAGAT. In: Situation awareness analysis and measurement., pp. 147–173. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US (2000)

7. Ferscha, A., Gollan, B.: Modeling Pupil Dilation as Online Input for Estimation of Cognitive Load in non-laboratory Attention-Aware Systems. COGNITIVE 2016: the Eighth International Conference on Advanced Cognitive Technologies and Applications: March 20-24, 2016, Rome, Italy (2016)

8. Gould, K.S., Røed, B.K., Saus, E.R., Koefoed, V.F., Bridger, R.S., Moen, B.E.: Effects of navigation method on workload and performance in simulated high-speed ship navigation. Applied Ergonomics **40**(1), 103–114 (Jan 2009). https://doi.org/10.1016/j.apergo.2008.01.001, http://linkinghub.elsevier.com/retrieve/pii/S0003687008000094

9. Hareide, O.S., Mjelde, F.V., Glomsvoll, O., Ostnes, R.: Developing a High-Speed Craft Route Monitor Window. vol. 10285, pp. 461–473. Springer International Publishing, Cham (2017). https://doi.org/10.1007/978-3-319-58625-0$_3$3, $http://link.springer.com/10.1007/978-3-319-58625-0_3$3

10. Hareide, O.S., Ostnes, R.: Maritime Usability Study by Analysing Eye Tracking Data. Journal of Navigation **70**(05), 927–943 (Sep 2017). https://doi.org/10.1017/S0373463317000182, https://www.cambridge.org/core/product/identifier/S0373463317000182/type/journal$_a$$rticle$

11. Hareide, O.S., Ostnes, R.: Validation of a Maritime Usability Study with Eye Tracking Data. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) Augmented Cognition: Users and Contexts, vol. 10916, pp. 273–292. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-91467-1$_2$2, $http://link.springer.com/10.1007/978-3-319-91467-1_2$2

12. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In: Advances in Psychology, vol. 52, pp. 139–183. Elsevier (1988). https://doi.org/10.1016/S0166-4115(08)62386-9, http://linkinghub.elsevier.com/retrieve/pii/S0166411508623869

13. Hess, E.H., Polt, J.M.: Pupil Size in Relation to Mental Activity during Simple Problem-Solving. Science **143**(3611), 1190–1192 (Mar 1964). https://doi.org/10.1126/science.143.3611.1190, http://www.sciencemag.org/cgi/doi/10.1126/science.143.3611.1190

14. Hiscocks, P.D., Eng, P.: Measuring Luminance with a Digital Camera. Syscomp Electronic Design Limited p. 27 (2014)

15. Holmqvist, K., Andersson, R.: Eye tracking: a comprehensive guide to methods and measures (2017), oCLC: 1057373387

16. Ikuma, L.H., Harvey, C., Taylor, C.F., Handal, C.: A guide for assessing control room operator performance using speed and accuracy, perceived workload, situation awareness, and eye tracking. Journal of Loss Prevention in the Process Industries **32**, 454–465 (Nov 2014). https://doi.org/10.1016/j.jlp.2014.11.001, https://linkinghub.elsevier.com/retrieve/pii/S095042301400179X

17. ISO: Road vehicles – Transport information and control systems – Detection- Response Task (DRT) for assessing attentional effects of cognitive load in driving. ISO 17488:2016. (2016)

18. Jung, H.S., Jung, H.S.: Establishment of overall workload assessment technique for various tasks and workplaces. International Journal of Industrial Ergonomics **28**(6), 341–353 (Dec 2001). https://doi.org/10.1016/S0169-8141(01)00040-3, http://linkinghub.elsevier.com/retrieve/pii/S0169814101000403

19. Kahneman, D., Beatty, J.: Pupil Diameter and Load on Memory. Science **154**(3756), 1583–1585 (Dec 1966). https://doi.org/10.1126/science.154.3756.1583, http://www.sciencemag.org/cgi/doi/10.1126/science.154.3756.1583

20. Kahneman, D., Beatty, J., Pollack, I.: Perceptual Deficit during a Mental Task. Science **157**(3785), 218–219 (Jul 1967). https://doi.org/10.1126/science.157.3785.218, http://www.sciencemag.org/cgi/doi/10.1126/science.157.3785.218

21. Kassner, M., Patera, W., Bulling, A.: Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction. In: Adjunct Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. pp. 1151–1160. UbiComp '14 Adjunct, ACM, New York, NY, USA (2014). https://doi.org/10.1145/2638728.2641695, http://doi.acm.org/10.1145/2638728.2641695

22. Kassner, M., Patera, W., Bulling, A.: Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction (Apr 2014), http://arxiv.org/abs/1405.0006

23. Ltd, K.M.S.S.P.: CS-2000 SPECTRORADIOMETER. Konica Minolta Sensing Singapore Pte Ltd (Mar 2018), http://sensing.konicaminolta.asia/products/cs-2000-spectroradiometer/

24. Marshall, S.P.: Method and apparatus for eye tracking and monitoring pupil dilation to evaluate cognitive activity (Mar 1999), https://patents.google.com/patent/US6090051A/en

25. Marshall, S.P.: What the Eyes Reveal: Measuring the Cognitive Workload of Teams. In: Duffy, V.G. (ed.) Digital Human Modeling, vol. 5620, pp. 265–274. Springer Berlin Heidelberg, Berlin, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02809-0_29, $http://link.springer.com/10.1007/978-3-642-02809-0_29$

26. Marshall, S.: The Index of Cognitive Activity: measuring cognitive workload. In: Proceedings of the IEEE 7th Conference on Human Factors and Power Plants. pp. 7–5–7–9. IEEE, Scottsdale, AZ, USA (2002). https://doi.org/10.1109/HFPP.2002.1042860, http://ieeexplore.ieee.org/document/1042860/

27. Palinko, O., Kun, A.: Exploring the Influence of Light and Cognitive Load on Pupil Diameter in Driving Simulator Studies. In: Proceedings of the 6th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design : driving assessment 2011. pp. 329–336. University of Iowa, Olympic Valley-Lake Tahoe, California, USA> (2011). https://doi.org/10.17077/drivingassessment.1416, http://ir.uiowa.edu/drivingassessment/2011/papers/48

28. Palinko, O., Kun, A.L., Shyrokov, A., Heeman, P.: Estimating cognitive load using remote eye tracking in a driving simulator. In: Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications - ETRA '10. p. 141. ACM Press, Austin, Texas (2010). https://doi.org/10.1145/1743666.1743701, http://portal.acm.org/citation.cfm?doid=1743666.1743701

29. Pickup, L., Wilson, J.R., Norris, B.J., Mitchell, L., Morrisroe, G.: The Integrated Workload Scale (IWS): A new self-report tool to assess railway signaller workload. Applied Ergonomics **36**(6), 681–693 (Nov 2005). https://doi.org/10.1016/j.apergo.2005.05.004, http://linkinghub.elsevier.com/retrieve/pii/S0003687005000918

30. Recarte, M.A., Nunes, L.M.: Effects of verbal and spatial-imagery tasks on eye fixations while driving. Journal of Experimental Psychology: Applied **6**(1), 31–43 (2000). https://doi.org/10.1037/1076-898X.6.1.31, http://doi.apa.org/getdoi.cfm?doi=10.1037/1076-898X.6.1.31

31. Recarte, M., Pérez, E., Conchillo, , Nunes, L.M.: Mental Workload and Visual Impairment: Differences between Pupil, Blink, and Subjective Rating. The Spanish Journal of Psychology **11**(2), 374–385 (2008). https://doi.org/10.1017/S1138741600004406

32. Rubio, S., Diaz, E., Martin, J., Puente, J.M.: Evaluation of Subjective Mental Workload: A Comparison of SWAT, NASA-TLX, and Workload Profile Methods. Applied Psychology **53**(1), 61–86 (Jan 2004). https://doi.org/10.1111/j.1464-0597.2004.00161.x, http://doi.wiley.com/10.1111/j.1464-0597.2004.00161.x

33. Rusnock, C.F., Borghetti, B.J.: Workload profiles: A continuous measure of mental workload. International Journal of Industrial Ergonomics **63**, 49–64 (Jan 2018). https://doi.org/10.1016/j.ergon.2016.09.003, https://linkinghub.elsevier.com/retrieve/pii/S0169814116301287

34. Sharma, C., Bhavsar, P., Srinivasan, B., Srinivasan, R.: Eye gaze movement studies of control room operators: A novel approach to improve process safety. Computers & Chemical Engineering **85**, 43–57 (Feb 2016). https://doi.org/10.1016/j.compchemeng.2015.09.012, https://linkinghub.elsevier.com/retrieve/pii/S0098135415003075

35. Stanley, P.: The effect of field of view size on steady-state pupil diameter. Ophthalmic and Physiological Optics **15**(6), 601–603 (Nov 1995). https://doi.org/10.1016/0275-5408(94)00019-V, http://linkinghub.elsevier.com/retrieve/pii/027554089400019V

36. Tsang, P.S., Velazquez, V.L.: Diagnosticity and multidimensional subjective workload ratings. Ergonomics **39**(3), 358–381 (Mar 1996). https://doi.org/10.1080/00140139608964470, http://www.tandfonline.com/doi/abs/10.1080/00140139608964470

37. Watson, A.B., Yellott, J.I.: A unified formula for light-adapted pupil size. Journal of Vision **12**(10), 12–12 (Sep 2012). https://doi.org/10.1167/12.10.12, http://www.journalofvision.org/lookup/doi/10.1167/12.10.12

38. Wierda, S.M., van Rijn, H., Taatgen, N.A., Martens, S.: Pupil dilation deconvolution reveals the dynamics of attention at high temporal resolution. Proceedings of the National Academy of Sciences **109**(22), 8456–8460 (May 2012). https://doi.org/10.1073/pnas.1201858109, http://www.pnas.org/cgi/doi/10.1073/pnas.1201858109
39. Wüller, D., Gabele, H.: The usage of digital cameras as luminance meters. p. 65020U. San Jose, CA, USA (Feb 2007). https://doi.org/10.1117/12.703205, http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.703205
40. Čegovnik, T., Stojmenova, K., Jakus, G., Sodnik, J.: An analysis of the suitability of a low-cost eye tracker for assessing the cognitive load of drivers. Applied Ergonomics **68**, 1–11 (Apr 2018). https://doi.org/10.1016/j.apergo.2017.10.011, https://linkinghub.elsevier.com/retrieve/pii/S0003687017302326

A.3   Appendix C (Consent Form 1)

Are you interested in taking part in the research project

"Development of a quantitative evaluation tool of cognitive workload in field studies through eye tracking."?

This is an inquiry about participation in a research project where the main purpose is investigate the use of eye-tracking technology in human factors engineering and usability testing with a specific focus on field studies. In this letter we will give you information about the purpose of the project and what your participation will involve.

Purpose of the project
The proposed project aims to investigate the use of eye-tracking technology in human factors engineering and usability testing with a specific focus on field studies.
The data we are collecting in this phase will help us validate a method to
estimate the pupil size as it changes over time for a given lumiose stimuli.
The research should start with the study and validation of such techniques outside a controlled environment and could eventually branch off in two different paths. - Usability testing of the current state of the art technology to evaluate the limits of such technology in a selection of case studies, e.g. ship bridge operations, driving a car. - Usability testing based on a single instance of the current technology to adapt it to a particular research case I.e.obtain an open source eye tracker from pupil labs and adapt the ergonomics and implementation to better fit a specific scenario such as helicopter pilot to eventually increase both the reliability and quantity of data reducing the impact of the instrument on the actual operation.

Who is responsible for the research project?
Researcher (student) Giovanni Pignoni
  giovanpi@stud.ntnu.no
  tel 46904106
Supervisor Sashidharan Komandur
  sashidharan.komandur@ntnu.no
  NTNU Norges teknisk-naturvitenskapelige universitet / Institutt for design

Why are you being asked to participate?
You have been selected as a student of NTNU Gjøvik with normal vision.

What does participation involve for you?
Participation to the study will require the subject to wear an eye tracking device, the participant will be asked to look on specific points on a screen while performing other mental tasks. The test will last approximately 15 minutes, after the test, the participant will be provided with the opportunity for a debriefing in which he/she will be able to ask further questions to the team regarding the research.
The eye tracker will record a video feed off the participant right eye.
The researcher will perform a pre-test screening to evaluate if the participant respects the selected criteria and reserve the possibility to interrupt the interview at any moment (e.g. if unable to calibrate

the instrumentation).

Participation is voluntary
Participation in the project is voluntary. If you chose to participate, you can withdraw your consent at any time without giving a reason. All information about you will then be made anonymous. There will be no negative consequences for you if you chose not to participate or later decide to withdraw.

Your personal privacy – how we will store and use your personal data
We will only use your personal data for the purpose(s) specified in this information letter. We will process your personal data confidentially and in accordance with data protection legislation (the General Data Protection Regulation and Personal Data Act).
You will not be identified in any reports on this study. This study is anonymous. We will not be collecting or retaining any information about your identity or that could indirectly identify you. The records of this study will be kept strictly confidential. The data will be stored until the end of the research (fall semester 2019), and eliminated afterwards. It is voluntary to participate, and you can at any time withdraw your consent without starting the reason. Contact giovanpi@stud.ntnu.no to ask for the removal of your personal data before the end of the study.

What will happen to your personal data at the end of the research project?
The received data will be deleted after the completion of the final report for the master thesis project connected to this research, at the latest 1st December 2019.
The data will be stored on an encrypted external hard drive, only the Researcher and Supervisor will have access to the data.

Your rights
So long as you can be identified in the collected data, you have the right to:
- access the personal data that is being processed about you
- request that your personal data is deleted
- request that incorrect personal data about you is corrected/rectified
- receive a copy of your personal data (data portability), and
- send a complaint to the Data Protection Officer or The Norwegian Data Protection Authority regarding the processing of your personal data

Given the nature of the anonymised data, it will be impossible to request a correction of the samples, it will be possible to ask for deletion of the data, contact the Principal Investigator for further information regarding the handling of your data.
Giovanni Pignoni giovanpi@stud.ntnu.no tel 46904106

What gives us the right to process your personal data?
We will process your personal data based on your consent.

Based on an agreement with NTNU, NSD – The Norwegian Centre for Research Data AS has assessed that the processing of personal data in this project is in accordance with data protection legislation.

Where can I find out more?
If you have questions about the project, or want to exercise your rights, contact:

- NTNU Norges teknisk-naturvitenskapelige universitet / Institutt for design via "Sashidharan Komandur".
- Our Data Protection Officer: "Thomas Helgesen"
- NSD – The Norwegian Centre for Research Data AS, by email: (personverntjenester@nsd.no) or by telephone: +47 55 58 21 17.

Yours sincerely,

Sashidharan Komandur   Giovanni Pignoni
(Researcher/supervisor)   (Student)

Consent form

I have received and understood information about the project "Development of a quantitative evaluation tool of cognitive workload in field studies through eye tracking." and have been given the opportunity to ask questions. I give consent:

- ☐ to participate in *the eye tracking experiment,*
- ☐ *for my personal data to be processed un Norway.*

I give consent for my personal data to be processed until the end date of the project, approx 1st December 2019.

(Signed by participant, date)

A.4   Appendix D (Consent Form 2)

Are you interested in taking part in the research project

"Development of a quantitative evaluation tool of cognitive workload in field studies through eye tracking."?

This is an inquiry about participation in a research project where the main purpose is investigate the use of eye-tracking technology in human factors engineering and usability testing with a specific focus on field studies. In this letter we will give you information about the purpose of the project and what your participation will involve.

Purpose of the project
The proposed project aims to investigate the use of eye-tracking technology in human factors engineering and usability testing with a specific focus on field studies. The data we are collecting in this phase will help us validate a method that estimates the pupil size as it changes over time for a given luminose stimuli.

Who is responsible for the research project?
Researcher (student) Giovanni Pignoni
        giovanpi@stud.ntnu.no
        tel 46904106
Supervisor Sashidharan Komandur
        sashidharan.komandur@ntnu.no
        NTNU Norges teknisk-naturvitenskapelige universitet / Institutt for design

Why are you being asked to participate?
You have been selected as a cadet of the Royal Norwegian Naval Academy

What does participation involve for you?
Participation to the study will require the subject to wear an eye tracking device while performing normal operation on board off the vessel. After the test, the participant will be asked to answer a questionnaire on the cognitive workload they experienced. The subjects will be provided with the opportunity for a debriefing in which he/she will be able to ask further questions to the team regarding the research. The eye tracker will record a video feed off the participant right eye. The researcher will perform a pre-test screening to evaluate if the participant respects the selected criteria and reserve the possibility to interrupt the interview at any moment.

Participation is voluntary
Participation in the project is voluntary. If you chose to participate, you can withdraw your consent at any time without giving a reason. All information about you will then be made anonymous. There will be no negative consequences for you if you chose not to participate or later decide to withdraw.

Your personal privacy – how we will store and use your personal data

We will only use your personal data for the purpose(s) specified in this information letter. We will process your personal data confidentially and in accordance with data protection legislation (the General Data Protection Regulation and Personal Data Act).

You will not be identified in any reports on this study. This study is anonymous. We will not be collecting or retaining any information about your identity or that could indirectly identify you. The records of this study will be kept strictly confidential. The data will be stored until the end of the research (fall semester 2019), and eliminated afterwards. It is voluntary to participate, and you can at any time withdraw your consent without starting the reason. Contact giovanpi@stud.ntnu.no to ask for the removal of your personal data before the end of the study.

What will happen to your personal data at the end of the research project?
The received data will be deleted after the completion of the final report for the master thesis project connected to this research, at the latest 1st December 2019.
The data will be stored on an encrypted external hard drive, only the Researcher and Supervisor will have access to the data.

Your rights
So long as you can be identified in the collected data, you have the right to:
- access the personal data that is being processed about you
- request that your personal data is deleted
- request that incorrect personal data about you is corrected/rectified
- receive a copy of your personal data (data portability), and
- send a complaint to the Data Protection Officer or The Norwegian Data Protection Authority regarding the processing of your personal data

Given the nature of the anonymised data, it will be impossible to request a correction of the samples, it will be possible to ask for deletion of the data, contact the Principal Investigator for further information regarding the handling of your data.
Giovanni Pignoni giovanpi@stud.ntnu.no tel 46904106

What gives us the right to process your personal data?
We will process your personal data based on your consent.

Based on an agreement with NTNU, NSD – The Norwegian Centre for Research Data AS has assessed that the processing of personal data in this project is in accordance with data protection legislation.

Where can I find out more?
If you have questions about the project, or want to exercise your rights, contact:
- NTNU Norges teknisk-naturvitenskapelige universitet / Institutt for design via "Sashidharan Komandur".
- Our Data Protection Officer: "Thomas Helgesen"
- NSD – The Norwegian Centre for Research Data AS, by email: (personverntjenester@nsd.no) or by telephone: +47 55 58 21 17.

Yours sincerely,

Sashidharan Komandur   Giovanni Pignoni

(Researcher/supervisor)   (Student)

Consent form

I have received and understood information about the project "Development of a quantitative evaluation tool of cognitive workload in field studies through eye tracking." and have been given the opportunity to ask questions. I give consent:

- ☐ to participate in *the eye tracking experiment,*
- ☐ *for my personal data to be processed un Norway.*

I give consent for my personal data to be processed until the end date of the project, approx 1st December 2019.

(Signed by participant, date)

A.5 Appendix E (Map)

Participant n. :

A.6   Appendix F (Read me)

COGNITIVE WORKLOAD TOOL FOR THE PUPIL EYE TRACKER

This script/application is part of my master's thesis project in MIXD developed at NTNU Gjøvik in the fall 2018 and spring 2019 semesters. The project aimed to experiment on the processing of eye tracking data, using an affordable eye tracker form Pupil Labs, for the measure of cognitive workload.
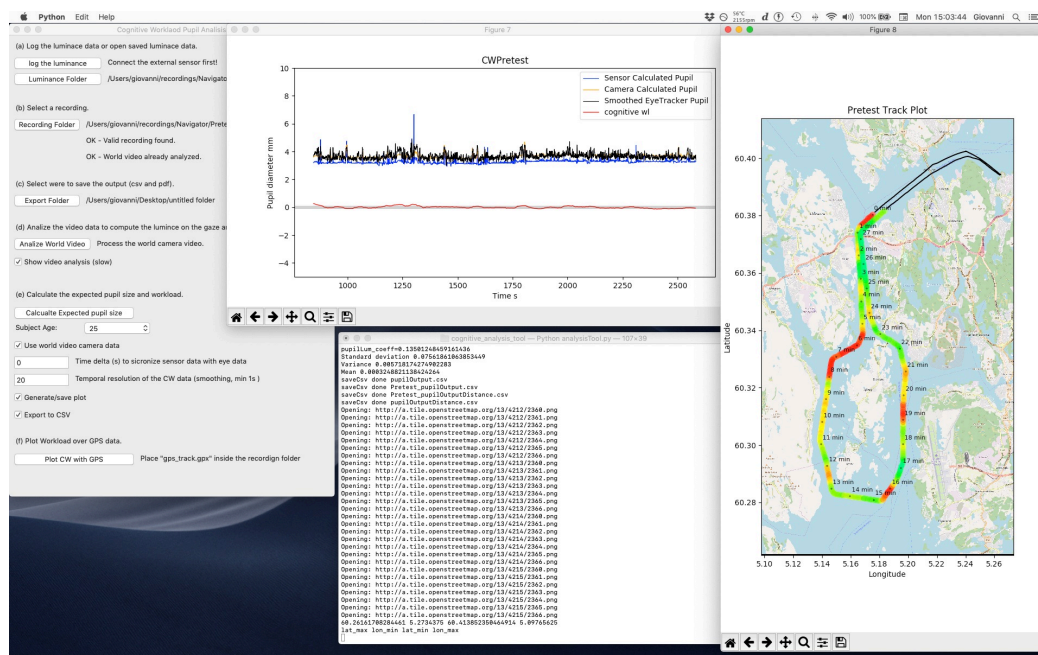
The tool user Interface.

The pupillary response (size of the pupil over time) is effected by the instantaneous cognitive workload level of the subject, higher workload results in a dilated pupil. Light intensity around the eye also has an effect on the pupil diameter as the pupil adjust to different luminosity.

This has historically limited the application of pupillometry to controlled laboratory study.

The scope of this software is to quantify the visual stimuli that the subject is receiving to subtract the effect of light from the pupillary response and allow the use of pupillometry in a filed condition with variable brightness.

This is done combining the gaze data and world video from the Pupil Mobile Eye Tracking Headset and calibrating it on the go with absolute luminance data from an external sensor Adafruit TSL2591.

*A participant wearing the eye trackers.*

Getting Started

- You should be familiar with the Pupil Mobile Eye Tracking Headset and the software suite that comes with it, be able to calibrate the gaze, make a recording and export it using the Pupil Player application.

- Visit the Pupil Labs Docs for more information about the eye tracker and eye tracking software.

Hardware
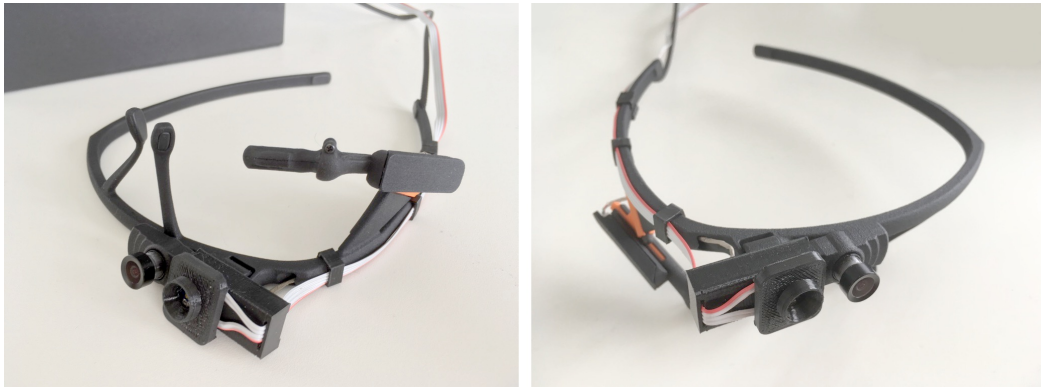
The assembled eye tracker.

The workload analysis requires the use of an external Luminosity Sensor, the Adafruit TSL2591 board.

To be able to compute the workload you need to make sure to record the output of the light sensor during the recording. This can be done in two ways either by connecting the Microcontroller to your computer or saving the data on the SD card directly.

Saving to the computer ensures that the timing of the luminance data is in sync with the eye tracking data as both are recorded on the same machine (same clock). The RTC on the Arduino will drift several seconds every day compared with the time on your computer or smartphone. If you decide to save the luminosity data on the SD card, you will also have to manually re-sync the two, specifying how much the Arduino clock is ahead or behind the computer clock.

The luminance value is timestamped with a Unix epoch time. Unfortunately, the Arduino code doesn't handle timezones. The best way to handle this is to set the time of the RTC to your local time (if your pc says it's 11:30 the Arduino should also say 11:30) and then include the difference (e.g. for Europe CET you should include the UTC Offset: UTC +1 and add 3600 seconds to compensate).
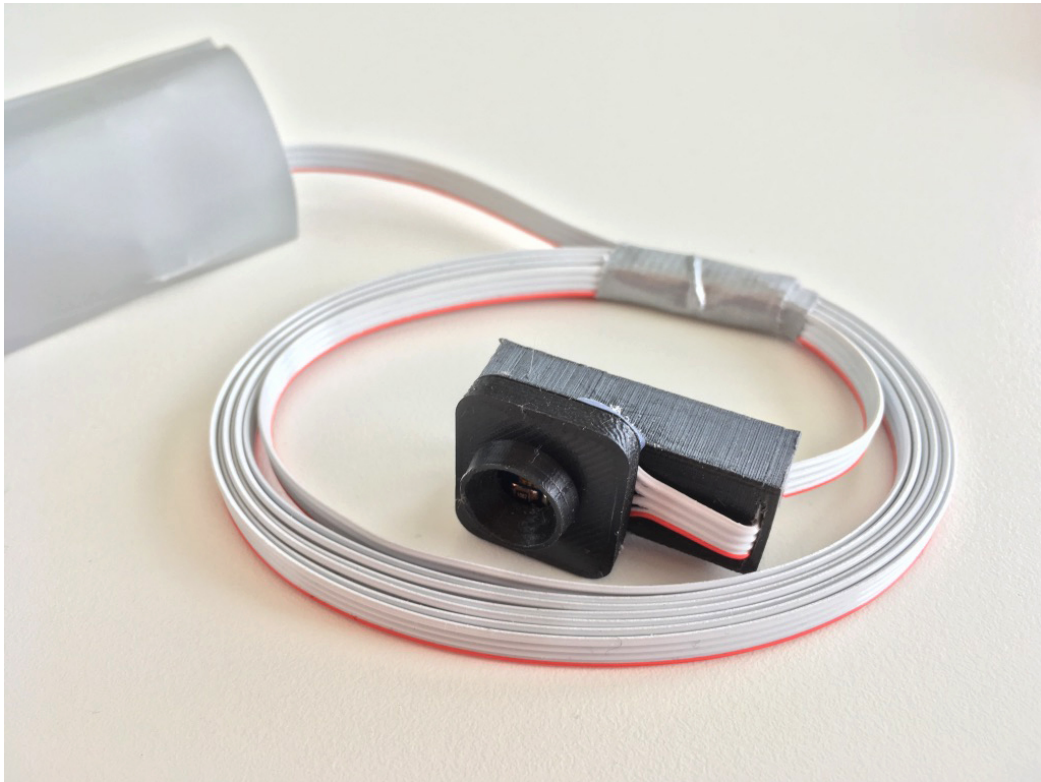
The Sensor and Logger.

- To log on the Microcontroller, you only need to power it up.

- To log on a laptop, you need to select a folder and press the "log the luminance" button.

- In both cases, the data will be divided into one new .csv per hour, containing the luminance values saved at around 10Hz.

To mount the board on the eye-tracker, you will need a 3d printed hardware kit; it can be ordered from Shapeways or downloaded and print it with a generic filament printer. The .stl 3d model is inside the "Lux Sensor" folder; it includes the hardware to secure the sensor to the workload camera on the Pupil Headset and clips to do some basic cable management.

To assemble the data logger, you will need a four conductors flat cable and a microcontroller, code is included to use both an Arduino Nano or the more powerful Adafruit Feather M4 Express. The Adafruit Feather is quite handy because it can be connected directly to a lithium battery and be used as a portable data logger.

If you plan always to use a computer for the recordings you only need the Microcontroller, if otherwise, you wish to use The Pupil Mobile bundle and record on a smartphone you will also need a data logging shield for the Microcontroller.

Example of BOM:

- Adafruit Feather M4, a powerful microcontroller with an included battery charger circuit.

- Adalogger FeatherWing, shield with Real-time clock and sd-card reader.

- 3.7v Lithium Battery.

- CR1220 backup battery for the real-time clock.

- Micro SD card 4-8gb is suitable for many days of logging.

- Ribbon Cable, at least 3-4 meters if you use a computer, 1.5m if you use the battery and the data logger says on the participant.

Alternatives to the Adafruit boards is to use:

- Arduino Nano as a microcontroller.

- Nano Data Logging Shield Deek-Robot ID 8105 as a data logging shield.

Prepare the luminance sensor

You can flash your Microcontroller with the Arduino IDE and the respective Arduino code included in the "Lux Sensor" folder.

The Adalogger FeatherWing code requires the Adafruit_Sensor and Adafruit_TSL2591 libraries for the sensor and uses the RTC_PCF8523 RTC.

See the Adafruit documentation to get started: - Feather M4, how to use it with the Arduino IDE.

- Adaloggher RTC Remember to set the time with one of the examples.

- tsl2591, wiring of the sensor and libraries.

The Nano Data Logging Shield code requires the Adafruit_Sensor and Adafruit_TSL2591 libraries for the sensor and uses the DS1307RTC RTC.

Use the example code in the respective libraries to initialise the RTC.

The sensor connects to the Microcontroller through the SDA SCL, and it is powered with up to 5v.

Pinout of the tsl2591:

- **Vin**, will take 3-5VDC safely, connect it to the 3V pin on the Feather

- **GND** - common ground for power and logic, connect it to the ground pin on the Feather

- **SCL** - I2C clock pin, connect to your micro-controllers I2C clock line (Labeled SCL on the Feather).

- **SDA** - I2C data pin, connect to your micro-controllers I2C data line (Labeled SDA on the Feather).

Software
The code is written in Python, but the GUI is in Cocoa, so it works only on Mac OS.

Installing
You need Python 3 installed on your machine:

To install Python, you should install Homebrew Open a new Terminal window and start typing:

```
/usr/bin/ruby -e "$(curl -fsSL
https://raw.githubusercontent.com/Homebrew/install/master/install)"
```

This will also install the Command line Tools for Xcode. Then use Homebrew to install python 3 and Open CV, be patient, it will take a while.

```
brew install python3
brew install opencv
```

Measurements, calculations and plots are made with the following tools:

```
python3.x -m pip install matplotlib
python3.x -m pip install Pillow
python3.x -m pip install numpy
python3.x -m pip install adjustText
python3.x -m pip install scipy
python3.x -m pip install DateTime
python3.x -m pip install pyserial
```

This project uses pyObjC and vanilla for the user interface on macOS. You can install them with the following commands:

```
python3.x -m pip install pyobjc
python3.x -m pip git+https://github.com/typesupply/vanilla
python3.x -m pip git+https://github.com/typesupply/defconAppKit
```

Running the application

To run the software launch the python application Download the software folder "cognitive_analysis_tool" on your computer and execute the "analysisTool.py" file. This will assume that you have placed the folder in your home directory.

```
cd /Users/YOURUSER/cognitive_analysis_tool
python3 analysisTool.py
```

Record Luminance data on Mac OS

To record the luminance data select a folder to save the data by pressing **Luminance Folder** step (a). Press **Log the luminance** to start saving. The interface will freeze, but you should see the luminance value change in the terminal window. Press Ctrl + z to top the logging.

Record Luminance without a computer

Once the Real Time Clock is set and the micro-controller is flashed with the correct software, and you have inserted a formatted (FAT32) micro SD card, you only need to power op the Microcontroller to start logging. Once finished disconnect the power and connect the SD card to your computer to retrieve the data.

Prepare a Pupil Recording

Visit the Pupil Labs Docs for instruction on how successfully make an eye tracking recording. Only a few parameters should be adjusted:

- **World camera**
  - Use the wide-angle camera at 1280x720px to avoid excessive vignetting 30 or 60 fps at your discretion.
  - Use automatic exposure.
  - Keep the standard post-processing settings.
- **Eye Camera**
  - Preferably use the higher resolution setting 400x400px and highest frequency 120Hz.
  - Use automatic exposure.
  - Keep the standard post-processing settings.
- **Recording**
  - Gaze calibration is important if you wish to use the more advanced algorithm that uses the world video in conjunction with the luminance sensor, for static images or a diffuse field (e.g. fixed luminance in a laboratory condition) the sole luminance sensor will be sufficient.

Once you have made a recording, either with the mobile software or Pupil Capture process it with the Pupil Player to export the raw data.

- In Pupil Player Activate the "Raw Data Exporter" module and select all three options. Press the Export Button on the left side.
- It is not necessary to export the World Video.
- Pupil Player will create a "exports" folder inside your recording folder, and an incremental subfolder inside every time you press the export button, at this moment the Cognitive workload tool will always look at the **first** (000) Folder.

Record GPS data

If you wish to plot the cognitive workload on a map inside the Cognitive workload tool, you can place a "gps_track.gpx" file inside the recording folder to be read by the software. Multiple applications can produce a gpx file, we tested myTracks for iOS and gpslogger for Android.

Pre-process the Wold video (optional)

Pre-processing of the world video is required to improve the analysis of the visual stimuli; the script will go through the footage to identify the area around the gaze of the subject so that the algorithm can evaluate to what the eye was adapting (as opposed to a vaguer average luminance in front of the participant).

To pre-process the world video, select the **Recording folder** pressing the relative button, step(b) in the interface. Then press ** Analyse World Video** to start. If you wish to see how the algorithm is interpreting the area around the gaze select "Show Video Analysis", but consider that the process will run slower. Press "q" at any moment to abort the analysis.

An outputFromVideo.csv file will be saved inside the recording folder.

Workload Analysis (Finally)

Make sure you select all the necessary folders:

- **Luminance Folder**, containing the luminance data either saved with the computer or a microcontroller and then transferred over via an SD card.

- **Recording Folder**, containing a recording that has been pre-processed with the Pupil Player app.

- **Export Folder**, a folder to save the exported data, a copy will always be saved inside the recording folder, but it is helpful to have another in a single folder when working with multiple files.

- Input the subject age.

- Select whether you want to use the world video data (you need to pre-process the Wold video first).

- Select a time delta if you need to re-synchronise the Luminance data (set to 0 if recording directly on the computer).

- Select the temporal resolution; this is an indicative value of the usable temporal resolution of the output data, 1s is the minimum, and is to be used in a situation with low variability of luminance, in field condition values around 30s are a more reasonable compromise.

- Decide whether to generate only a visual output or to save the output data-sheet.

Output data-sheet

RecordingNAME_pupilOutputDistance

- **relative_wl**, the linear distance between the expected pupil size and the

measured pupil size, it represents the cognitive workload, 0 is the average workload for the given task, positive values are above average workload negative values are below average workload.

- **timestamp_relative**, Timestamp in seconds from the beginning of the recording.
- **recording_name**
- **age**
- **timestamp_unix**, Timestamp in UNIX Epoch.

RecordingNAME_pupilOutput

- **timestamp_unix**, Timestamp in UNIX Epoch
- **timestamp_relative**, Timestamp in seconds from the beginning of the recording.
- **frame_n** Relative frame number of the World Video.
- **confidence** Confidence of the pupil Algorithm
- **mm_pupil_diameter_scaled** Pupil Diameter in mm scaled so that the mean matches with the calculated pupil size.
- **mm_pupil_diameter_calc_lux** Pupil Diameter in mm Calculated using the Luminance sensor only.
- **px_pupil_diameter_raw** Pupil Diameter in px.
- **recording_name**
- **age**
- **mm_pupil_diameter_calc_camera** Pupil Diameter in mm Calculated using the Luminance sensor and the world camera.

GPS Plot

To plot the GPS, you need to calculate the workload first and a "gps_track.gpx" file needs to be present inside the recording folder; an internet connection is necessary to download the background map from Open Street Maps.

License

This project is licensed under the MIT License - see the LICENSE.md file for details

Acknowledgments

The magicwand.py script has been adapted from Alexander Reynolds work magicwand, A Python+OpenCV implementation similar to Adobe Photoshop's magic wand selection tool.

The "Light Adapted Pupil Size" code has been ported from R to python and was originally written by Jose Gama CVD. And it is based on the publication by Watson A. B., Yellott J. I. (2012). A unified formula for light-adapted pupil size. Journal of Vision, 12(10):12, 1–16. http://journalofvision.org/12/10/12/, doi:10.1167/5.9.6.