Cross-Participant and Cross-Task Classification of Cognitive Load Based on Eye Tracking

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard-Karls-Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Tobias Appel
aus Marburg

Tübingen
2020

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Prüfung: 27.01.2021
Stellvertretender Dekan: Prof. Dr. József Fortágh
1. Berichterstatter: Prof. Dr. Enkelejda Kasneci
2. Berichterstatter: Prof. Dr. Katharina Scheiter

Acknowledgments

An erster Stelle möchte ich meinen Betreuern und Mentoren Prof. Dr. Enkelejda Kasneci und Prof. Dr. Peter Gerjets danken. Sie haben mich dazu ermutigt, mich beim Doktorandenprogram der LEAD Graduiertenschulen zu bewerben und haben mich während meiner Zeit als Doktorand immer unterstützt. In meinem akademischen Arbeiten haben sie mir einerseits die Freiheit gegeben, mich selbst zu entfalten, andererseits waren sie stets zur Stellen, wenn ich sie gebraucht habe. Für ihre wissenschaftlichen Anregungen und ihre konstruktive Kritik möchte ich mich herzlich bedanken.

Ich bedanke mich bei der Graduiertenschule und dem Forschungsnetzwerk LEAD, sowie dem Wilhelm-Schickard-Institut für Informatik, die mir durch Stellen als wissenschaftlicher Mitarbeiter die Möglichkeit eröffnet haben, diese Doktorarbeit zu schreiben. Besonders möchte ich mich bei Mareike Bierlich und Sophie Freitag für ihre Unterstützung im Alltag bei LEAD bedanken.

Mein Dank gilt auch Dr. Christian Scharinger, Prof. Dr. Korbinian Möller, Dr. Manuel Ninaus, Franz Wortha, Katerina Tsarava und Natalia Sevcenko dafür, dass sie Datensätze mit mir geteilt und mich beim Vorbereiten von Publikationen unterstützt haben. Ebenso bedanke ich mich bei Stefan Hoffmann von Promotion Software für die fruchtbare Zusammenarbeit.

Ich möchte mich auch bei meinen ehemaligen und aktuellen Mitdoktoranden und Kollegen sowohl bei LEAD als auch beim Wilhelm-Schickard-Institut bedanken, die meinen akademischen Horizont erweitert haben und mit denen ich Probleme und Erfolgen teilen durfte.

Mein herzlicher Dank gilt auch meinen Freunden und meiner Familie, die mich während meiner Promotion immer unterstützt haben und mir stets mit Verständnis und Geduld zur Seite gestanden sind.

Zuletzt gilt mein ganz besonderer Dank meiner Frau Luzia. Ich kann nicht in Worte fassen, wie sehr sie mir in den dreineinhalb Jahren meiner Promotion eine Stütze war und mir in hektischen Zeiten den Rücken freigehalten hat.

Summary

Cognitive load refers to the total amount of working memory resources a person is currently using. Successfully detecting the cognitive load a person is experiencing is the first important step towards applications that adapt to a user's current load. Provided that cognitive load is estimated correctly, a system can enhance a user's experience or increase its own efficiency by adapting to this detected load. Using digital learning environments as an example to illustrate this idea, a learning environment could tune the difficulty of presented exercises or learning material to match the learner's current load to not underwhelm them, but also to prevent overload and frustration.

Physiological sensors have great promise when cognitive load estimation is concerned as many physiological signals show distinctive signs of cognitive load. Eye tracking is an especially promising candidate as it does not require physical contact between sensor and user and is therefore very subtle. A major problem is the lack of general classifiers for cognitive load as classifiers are usually specific to a single person and do not generalize well. For adaptive interfaces based on a user's cognitive load to be viable, a classifier that is accurate and performs well independently of user and specific task would be needed. In the current doctoral thesis, I present four studies that successively build upon each other and build up towards an eye-tracking based classifier for cognitive load that is 1) accurate, 2) robust, 3) can generalize, and 4) can operate in real-time.

Each of the presented studies advances our approach's capability to generalize one step further. Along the way, different eye-tracking features are explored and evaluated for their suitability as predictors of cognitive load and the implications for the distinction between cognitive load and perceptual load are discussed. The resulting method demonstrates a degree of generalization that no other approach has achieved and combines it with low hardware requirements and high robustness into a method that has great promise for future applications. Overall, the results presented in this thesis may serve as a foundation for the use of eye tracking in adaptive interfaces that react to a user's cognitive load.

Zusammenfassung

Kognitive Belastung bezeichnet die Menge an Arbeitsgedächtnisresourcen, die eine Person gerade verwendet. Das erfolgreiche Erkennen von kognitiver Belastung stellt den ersten Schritt auf dem Weg zu Anwendungen dar, die sich der Belastung des Nutzers anpassen. Vorausgesetzt, dass die kognitive Belastung richtig eingeschätzt wurde, könnte ein System sich dieser Belastung anpassen und damit das Erlebnis des Nutzers verbessern oder seine eigene Effizienz steigern. Dies lässt sich gut am Beispiel von digitalen Lernumgebungen veranschaulichen. Die Lernumgebung könnte den Schwierigkeitsgrad der präsentierten Aufgaben und des Lernmaterials so anpassen, dass sie den kognitiven Anforderungen der lernenden Person ensprechen, sodass sie weder überwältigt ist, noch durch Überbeanspruchung frustriert wird.

Physiologische Sensoren sind vielversprechend, wenn es um die Einschätzung von kognitiver Belastung geht, da sich diese Belastung in vielen physiologischen Signalen widerspiegelt. Eye Tracking sticht dabei besonders heraus, da es keinen physischen Kotakt mit dem Anwender braucht und daher sehr subtil eingesetzt werden kann. Ein großes Problem ist jedoch, dass Klassifizierungsmethoden, die mit physiologischen Signalen arbeiten, schlecht verallgemeinern können und die resultierenden Klassifikatoren nur spezifisch für die Person funktionieren, für die sie trainiert wurden. Damit adaptive Interfaces, die sich der kognitiven Belastung des Nutzers anpassen können, realisierbar sind, benötigt man einen Klassifikator, der genau ist und unabhängig vom Nutzer und der konkreten Aufgabe gute Ergebnisse liefert. In der vorliegenden Doktorarbeit präsentiere ich vier Studien, die aufeinander aufbauen und sukzessive auf einen Eye-Tracking basierten Klassifikator für kognitive Belastung hinarbeiten, der 1) genau, 2) robust und 3) allgemeingültig ist, sowie in 4) Echtzeit verwendet werden kann.

Jede der vorgestellten Studien bringt uns der Generalisierbarkeit unserer Vorgehensweise einen Schritt näher. Dabei werden verschiedene Eye-Tracking Feature untersucht und auf ihre Eignung hin getestet, als Prädiktoren für kognitive Belastung zu fungieren. Außerdem werden die Implikationen, die diese Feature für die

Unterscheidung zwischen kognitiver Belastung und Belastung der Wahrnehmung haben, erörtert. Die Methode, die in dieser Arbeit präsentiert wird, erreicht einen Grad an Generalisierung, der von keinem anderen Verfahren erreicht wird, und kombiniert diese Tatsache mit einem sparsamen Umgang mit Hardware-Resourcen und guter Robustheit zu einem Verfahren, das für zukünftige Anwendungen sehr vielversprechend ist. Die Ergebnisse, die in dieser Arbeit präsentiert werden, können als Grundlage für adaptive Anwendungen dienen, die Eye Tracking verwenden und auf die kognitive Belastung des Nutzers reagieren.

List of Publications and Contribution

Parts of this thesis have been published elsewhere. This chapter lists all publications that are part of this thesis and specifies my contribution to each of them. The publications themselves can be found in the Appendix.

1. **Appel, T.** [1], Santini, T., Kasneci, E. (2016). Brightness-and motion-based blink detection for head-mounted eye trackers. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct* (pp. 1726–1735). doi: 10.1145/2968219.2968341

   > The raw data was provided by E.K. and labeled by myself. The methodology and algorithm were designed by myself. I programmed the algorithm and performed the analysis with input by E.K and T.S. The manuscript was mainly written by myself with contributions from T.S. and supervision from E.K. I consider my contribution to this work to be 85% of total work.

2. **Appel, T.**, Scharinger, C., Gerjets, P, Kasneci, E. (2018). Cross-subject workload classification using pupil-related measures. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications* (pp. 1–8). doi: 10.1145/3204493.3204531

   > The raw data was provided by C.S. The methodology and algorithm were designed by myself with input from E.K and P.G. I programmed the algorithm and performed the analysis with input by E.K and P.G. The manuscript was mainly written by myself with supervision from E.K. I consider my contribution to this work to be 90% of total work.

[1] This publication was not authored during my PhD, but during the pursuit of my Master's degree in computer science

3. **Appel, T.**, Sevcenko, N., Wortha, F., Tsarava, K., Moeller, K., Ninaus, M., Kasneci, E., Gerjets, P. (2019). Predicting Cognitive Load in an Emergency Simulation Based on Behavioral and Physiological Measures. In *Proceedings of the 2019 International Conference on Multimodal Interaction* (pp. 154–163). doi: 10.1145/3340555.3353735

   The raw data was provided by N.S, K.T, K.M, and M.N. The experiment was carried out by N.S. and F.W. under supervision of K.M., M.N. and P.G. The methodology and algorithm were designed by myself with input from E.K, P.G., K.M. and M.N. I programmed the algorithm and performed the analysis with input by E.K, P.G., K.M. and M.N. The manuscript was mainly written by myself with contributions from K.M., E.K. and P.G. I consider my contribution to this work to be 80% of total work.

4. **Appel, T.**, Gerjets, P., Hoffmann, S., Moeller, K., Ninaus, M., Scharinger, C., Sevcenko, N., Wortha, F., Kasneci, E. (submitted). Cross-task and Cross-participant Classification of Cognitive Load in an Emergency Simulation Game. In *IEEE Transactions on Affective Computing*

   The raw data for *Emergency* was provided by S.H., N.S, K.T, K.M, and M.N. The raw data of the N-back task was provided by C.S. The *Emergency* experiment was carried out by N.S. and F.W. under supervision of K.M., M.N. and P.G. The methodology and algorithm were designed by myself with input from E.K, P.G., K.M. and M.N. I programmed the algorithm and performed the analysis with input by E.K, P.G., K.M. and M.N. The manuscript was mainly written by myself with contributions from K.M., E.K. and P.G. I consider my contribution to this work to be 80% of total work.

Contents

1 Introduction and Theoretical Framework

Our brain's resources - e.g. working memory, attention, decision making, or task-related knowledge - are limited. The degree to which these resources are currently used is commonly referred to as cognitive load [1]. It is apparent that measuring and managing cognitive load in human-computer interaction is beneficial in a plethora of situations and even essential in others. If a system would be able to adapt to the user's current state, frustration and stress caused by cognitive overload could be avoided as well as boredom that may originate from low levels of cognitive load and too little challenge [2].

This may be especially relevant for digital learning environments. In learning contexts, the zone of proximal development [3] describes the difficulty level of a learning task that is best for growth by challenging the learner, but not overwhelming them. This difficulty level can be overcome with help, differentiating it from material that is too easy (which offers no opportunity to learn) or too difficult to even solve with help. Provided that estimation of cognitive load works successfully, the difficulty of learning material could be adjusted to suit the learner's current needs or scaffolding could be employed to point the learner in the right direction. Utilizing such adaptations the learner can be kept in the zone of proximal development, thereby optimizing learning.

This concept, however, is not limited to learning scenarios. Virtually any digital environment can incorporate cognitive load assessments and many could adapt to it. Gaming is an obvious example, as difficulty can usually be adjusted in various ways depending on the specific game. By lowering the difficulty when cognitive load indicates that a user is struggling with a challenge, frustration can be prevented and the overall experience can be improved. On the other hand, certain obstacles may be to easy and can be made more enjoyable by increasing their difficulty. Adapting to cognitive load can help tailoring the experience to the user's specific skill-set, expertise, and current mental and cognitive states. Achieving a flow-state [4] by balancing the presented challenge with the user's available resources is a prime example for useful adaptation in gaming creating a state of

optimal experience [5].

Conventional ways to measure cognitive load such as questionnaires or secondary tasks are not viable for real-time adaptation as they either do not provide a continuous estimation or interfere with the main task. Predictive machine learning models that use physiological data on the other hand can provide exactly that: a continuous estimation of cognitive load without imposing additional load through a secondary task.

What limits cognitive load estimation using physiological sensors in its current form - and consequently its applications - is the lack of generalization. Machine learning classifiers that distinguish between different levels of cognitive load are usually person and task specific and do not work outside of their specific context [6], [7]. This challenge is not limited to physiological approaches, but cognitive load estimation in general, as Heard et al. conclude in their meta-review of algorithmic assessments of cognitive load [8]. In the current doctoral thesis I want to work towards a person and task independent classifier for cognitive load that relies on eye tracking. Compared to most other physiological measures eye tracking is not intrusive while still offering good classification accuracy (e.g. up to 70% as reported by Hogervost and colleagues [9]), making it a prime candidate for indirect cognitive load assessment. The goal of this thesis is first to find a way to generalize across participants, then outfit this approach with enough robustness to handle real-world situations, and finally show that it also works across different tasks. This level of generalization has not been demonstrated by researchers before and therefore can be seen as a major leap in cognitive load research that possibly enables many applications in the future.

1.1 Theoretical Framework

There are many theories concerned with cognitive load. On the one hand there is the broader term of “workload” referring to an amount of labor or quantified effort. It is often used in context of human factors and described by Wickens’ Multiple Resource Theory (MRT) [10], [11]. Wickens proposes that there are multiple different resource pools that can be tapped simultaneously and describes them in terms of modality (e.g. visual, auditory, or tactile), stage (perception, processing, and action), and reasoning (subconscious, symbolic, or linguistic). Each resource pool is finite and can be overloaded, forming a potential bottleneck.

On the other hand there are theories like Cognitive Load Theory (CLT) [1], [12], [13] that focus on instructional design and describe cognitive load as the amount of working memory resources being used. Working memory describes the small amount of information that can be held and manipulated in one's mind simultaneously for the execution of a current cognitive task [14]]. CLT mainly focuses on instructional design in educational contexts, but can also be applied to concepts of human-computer interaction [15]. The cognitive load imposed on the working memory's storage components is seen as the main bottleneck for learning, assuming that these components constrain the amount of new information that can be processed simultaneously in order to be integrated into long-term memory. If this storage is overloaded, learning is hindered and performance decreases.

The amount of cognitive load a person experiences depends on the presentation of a task as well as on the relationship between the learner's skill and the task complexity. A task can cause varying amounts of cognitive load depending on the knowledge prerequisites, making a fixed task easy for an expert and difficult for a novice. This kind of load is referred to as intrinsic cognitive load, as it is inherent to a task itself. The second type of cognitive load is called extraneous cognitive load. Extraneous cognitive load is caused by the manner in which information is presented to the learner, so it can be eased by beneficial instructional design and modalities. The final type of load defined in the framework of CLT is germane cognitive load. It describes the load caused by processing, construction and automation of schemas.

1.2 Measuring Cognitive Load

The foundation for any kind of adaptation based on cognitive load is to assess it [16], [17]. Three different avenues to assess cognitive load exist, each with its own advantages and disadvantages: self-reports, task performance, and physiological measures.

1.2.1 Self-Reports

Subjective rating scales like the NASA-TLX [18], see widespread use as a tool for cognitive load assessment and have been successfully used in various areas of research. They however, offer little flexibility and are very coarse in their granu-

larity. Measurements are only available at few points in time and not continuously throughout an experiment. Moreover, there may be other factors influencing the results of such a questionnaire with the current level of cognitive load being only one of them [19]. As a person needs to stop a task in order to fill out a questionnaire, its nature is rather disruptive and makes it unsuitable for any real-time application.

1.2.2 Task Performance

The second method of measurement is task performance. Quantifiable performance metrics like reaction times obtained from a secondary task are used to evaluate how much cognitive load is caused by the primary task. The worse the secondary task performance, the higher the cognitive load induced by the primary task. These dual-task paradigms (e.g. [20]) have the distinct advantage that they are objective, but the induced secondary load interferes with the primary task. Additionally, while the assessment is available at more points in time, it is still not continuous.

Antonenko and colleagues conclude that neither performance metrics nor questionnaire data can provide continuous and unobtrusive cognitive-load monitoring which is a prerequisite for real-time applications [21].

1.2.3 Physiological Measures

The human nervous system is divided into two sub-subsystems: the central nervous system (CNS) and the peripheral nervous system. The CNS is comprised of all cells of the brain, brain stem, and spinal cord, whereas the peripheral nervous system encompasses cells outside the skull and spinal column. Cognitive load is related to both systems and can either be measured directly or through physiological changes that accompany the changes in the nervous system.

Common ways to evaluate CNS activity are neuro-imaging procedures like electroencephalography (EEG), near-infrared spectroscopy (NIRS), or functional magnetic resonance imaging (fMRI) that measure brain activity. One successful approach for participant independent classification of cognitive load was published by Popovic et al. [22]. They used a combination of EEG and electrocardiography (ECG) achieving 72.5% classification accuracy in a leave-one-participant-out cross-validation. Wang and colleagues [23] were able to use hierarchical Bayes models

that were trained on all of their 8 participants to classify cognitive load for individual participants with roughly 80% accuracy. They used EEG measurements in their approach and concluded that their results were on par with within-participant results for similar data. Cross-task capability was demonstrated by Ke et al. for EEG data [24]. By applying feature selection to participant-specific regression models they could generalize from a working memory task to a complex simulated multi-attribute task (see [25] for details of the task). Furthermore, Krol and colleagues developed a paradigm to calibrate a task-independent EEG classifier for participant-specific use [26] that was successfully used in further studies [27], [28] and achieved results above chance-level.

The peripheral nervous system is made up of two parts: the somatic nervous system that is concerned with voluntary activation of muscles and the autonomic nervous system (ANS) governing subconscious bodily functions like breathing, heart beats, or digestion. The ANS is influenced by cognitive load through two systems that counteract each other: the parasympathetic nervous system and the sympathetic nervous system. The fundamental function of the sympathetic nervous system is to ready the human body for emergencies and mobilize the required resources. This includes increasing heart rate and blood flow to internal organs, sweating, and pupil dilation. Opposed to the sympathetic nervous system the parasympathetic nervous system regulates the conservation and maintenance of bodily resources. High cognitive load impacts the ANS in a similar way as a fight-or-flight situation by reducing activity of the parasympathetic nervous system and increased activity of the sympathetic nervous system. This is reflected in physiological changes that can be picked up by sensors and used to estimate the cognitive load that caused them.

Sarkar and colleagues successfully implemented a cross-participant method to detect cognitive load based on ECG recordings of 9 participants during a surgical simulation task [29]. After baseline correction and normalization they achieved 88.74% classification accuracy for distinguishing low and high cognitive load. Even though the number of participants is small, there seems to be potential for cross-participant cognitive load detection as long as physical contact is not an issue for the experiment.

1.3 Eye Tracking and Cognitive Load

The majority of physiological sensors are intrusive in some way (e.g. by being in physical contact with a user) and therefor may break immersion. A notable exception is eye tracking. Eye tracking - more specifically remote eye tracking - offers a great way to obtain physiological information indicating cognitive load in an non-intrusive way. It is non-invasive by nature and does not require physical contact. This helps with keeping users immersed as their awareness of the sensor is very low and the application does not need to be interrupted to get an estimation of cognitive load.

1.3.1 Overview

Eye tracking is the process of measuring either the point of gaze (where one is looking) or the motion of the eye relative to the head. A device used for this procedure is called an eye tracker. Eye movement has been studied since the 1800s but only on the basis of visual observations or using very intrusive methods, such as giant contact lenses connected to an aluminium pointer [30]. In 1937, Guy Thomas Buswell was the first to use light beams directed at the eye which were reflected and captured on film [31]. He used his apparatus to study reading behavior and noticed differences in oral and silent reading. In the 1950s and 1960s, Alfred Lukyanovich Yarbus conducted research concerning task-dependent exploration strategies. In 1967, he published a book called “Eye Movements and Vision” A. L. Yarbus, *Eye Movements and Vision*. Plenum Press, 1967 that would be quoted by eye-tracking enthusiasts for the next decades and is still one of the most influential books in the research field. Since then, eye-tracking technology has advanced and eye trackers have become increasingly versatile, accurate and popular.

In general, there are two different types of eye trackers in use today: remote eye trackers and head-mounted eye trackers. The former utilize several infrared, near-infrared or regular cameras (usually mounted as an array below a display) to measure the pupil and calculate the gaze position based on that data. The latter are similar in appearance to glasses but feature several cameras. One is facing away from the wearer and (usually) two infrared or near-infrared cameras are fixed below the wearer’s eyes and capture eye movements. From the recorded pupil and eye data, gaze vectors for both eyes are calculated with regard to the

current head position of the wearer. Since in both cases the pupil has to be detected in an image, this means that apart from the pupil center used for the gaze vector, the pupil diameter is an additional measure that can be used as a source of data.

Meanwhile, eye tracking is employed in various application fields. It can be employed in many psychological studies - especially reading studies [33]–[35]. Marketing and economics also utilize eye tracking to analyze the efficiency of advertisement and optimize their placement [36]–[38]. In the area of human-computer interaction eye-tracking has been employed as a modality to improve the interaction with a user [39]–[41]. There are also many applications for eye tracking in driving tasks and the simulations thereof. It can for example be used to detect whether an obstacle, traffic sign, or hazard has been noticed by the driver [42] or eye related features can be used to detect fatigue, drowsiness, or defects in the field of vision [43]–[52].

Advances in technology render eye trackers affordable for a broad audience [53] and many VR-headsets come already pre-equipped with eye-tracking technology. Recent studies even show that low-cost eye trackers can be used for estimating cognitive load [54], [55]. Furthermore, many manufacturers add cameras and eye trackers to their cars which enhances the opportunity for pervasive cognitive load estimation even further. As no other physiological sensor achieves the balance of cost, accuracy, and availability that eye tracking offers, it is a prime candidate for applications that aim for a large target audience.

Technological advancement does not only happen with regards to hardware, but eye-tracking software is getting more potent, too. Pupil detection has greatly improved over the last years enabling not only more accurate pupil diameter measurements, but also more precise gaze estimation [56]–[61]. These improvements make high quality eye tracking available even on low-cost devices. Further improvements in calibration procedures [62] and slippage compensation [63] dramatically increase data quality and robustness to real-world circumstances of head-mounted eye trackers.

Eye movements are very sensitive and private in nature. Even based on anonymous data inference about personal attributes like age and gender is possible [64], [65], which poses a threat to a user's privacy. Therefore, privacy-preserving methods have to be employed in order to protect users. Differential privacy approaches seem to be a good solution to these concerns on a database level [66], while ran-

domized encoding can be used to calculate a user's gaze without accessing their facial landmarks [67].

1.3.2 Features

To get closer to the goal of a general eye-tracking based cognitive load classifier, I rely on features that are in large parts independent of the presented stimulus and its structure. Assuring this criterion allows us to generalize across tasks and situations. Saccade characteristics are highly dependent on task [68]–[70] and stimulus as they are guided by visual attention and therefore are ill-suited as generalizing features. The same holds true for most features that are derived from fixations. Examples that illustrates this fact well are tasks that do not employ visual stimuli. Pupil diameter features are still reported to change as expected (see [71] for a good overview), but fixation and saccade characteristics lose their meaning. Summarizing, any eye-tracking feature that relies on gaze positions is unlikely to generalize across different tasks, consequently, the primary focus of this thesis is on measurements of pupil diameter. Still, I consider certain basic fixation and saccade metrics, microsaccades, as well as blinks.

Fixations

Fixations describe a voluntary, stable gaze on the same location typically lasting between 200 ms and 350 ms [72], but they may last up to several seconds. The frequency of fixations is influenced by a lot of factors. Time pressure tends to increase the number of fixations while reducing their duration [73]. Chen et al. related fixation duration to the level of cognitive processing whereas a high fixation duration and decreased fixation rate indicate higher working memory usage [74]. They interpret these observations as indicators of increased attention caused by tasks with higher complexity. The idea that longer fixation are associated with higher processing load and more effort is supported by further eye-tracking research [75]–[78].

Other research leans in the opposite direction by associating higher fixation duration with less invested cognitive effort [79], [80]. It is possible that these results are influenced by processing difficulty or visual demands of the stimuli [72].

Saccades

Saccades are rapid eye movements that usually occur between fixations. The effect that cognitive load has on specific saccade characteristics is highly dependent on the presented stimulus, which makes them unlikely candidates for features that generalize well. This work therefore focuses on microsaccades instead of regular saccades.

Microsaccades are small involuntary eye movements that may occur during a fixation and are associated with cognitive load and visual load. Studies reported an increase in microsaccade frequency in visually demanding tasks [81], whereas non-visual tasks (e.g, auditory tasks or mental arithmetic) seemed to reduce their frequency [82]–[84].

Pupil

The pupil is the circular black area in the center of an eye that allows light to strike the retina [85]. It is controlled by two muscles: the sphincter muscle *sphincter pupillae* responsible for contracting the pupil and the radial muscle *dilator pupillae* that dilates it. The ANS influences the pupil diameter through the parasympathetic and sympathetic nervous system respectively. High task demand decreases PNS activity and increases SNS activity, both causing the pupil to dilate [86], [87], which is often referred to as task-evoked pupillary response. Kahneman showed that this effect persists within task, between tasks, and between individuals, concluding that there is a consistent influence of cognitive load in the pupil diameter [88].

Many factors, however, play a role in pupil dilation. The most common is the light reflex - the pupil's involuntary adaptation to lighting. It regulates the amount of light that strikes the retina and assists in adaptation of vision to various levels of lightness/darkness. A phenomenon unrelated to this function is known as hippus or pupillary unrest, which is a rhythmical and regular contraction and dilation of the pupil [89]. It is independent of eye movement and lighting conditions and usually normal, however pathological hippus can occur with increased frequency and/or amplitude.

The connection between pupil dilation and cognitive load has been investigated since the early 1960s [90], [91], but is receiving more and more attention in recent years. Advances in technology make eye tracking in general and pupil diameter

estimation in particular more accurate and easier to apply. The accompanying reduction in cost helps making this avenue of CL estimation more accessibly, further increasing its use.

Pupil dilation has been used extensively as an indicator of cognitive effort throughout the last decades. Amongst the most prominent areas of research are driving and simulations thereof. Palinko and colleagues showed in multiple studies that pupil based CL estimation is feasible in driving scenarios [92], [93] - to a certain extent even under different lighting condition [94]. They administered an auditory version of the n-back task to induce different levels of cognitive load while driving and concluded that remote eye tracking might provide reliable driver cognitive load estimation. This evaluation is supported by several other researchers [95], [96].

In accordance with common preprocessing guidelines [97], [98], all studies of this thesis first remove blink artifacts and implausible outliers. Furthermore, small gaps are interpolated and the pupil signal is smoothed to reduce noise.

A key point in cross-participant CL estimation is the use of relative measures. A pupil diameter in millimeters carries little information outside a person-specific context. In order to meaningfully compare participants, pupil diameters relative to a baseline are imperative [97], [98]. Subtracting a baseline pupil diameter from a participant's pupil signal allows for analysis with reduces participant-specific influence and therefor helps generalization. In all studies presented in this thesis, I either use a tutorial or an instruction phase as a baseline in order to have greater real-world application compared to a conventional baseline measurement with a fixation cross.

Cognitive activity was observed to cause fluctuations in the pupil signal mainly consisting of sharp and pronounced spikes. Unlike the rhythmical and regular hippus, these events are more rapid and do not follow any rhythm. Marshall patented her approach to measuring this phenomenon as the Index of Cognitive Activity (ICA) in 2000 [99]. An openly accessible adaptation was developed by Duchowski and colleagues to counteract the intransparency of the ICA: the Index of Pupillary Activity (IPA) [100]. It is meant to open up this method of pupil analysis to a broader audience and help prove the validity of measuring rapid pupil dilation as an index of cognitive processing load. Both utilize wavelet decomposition in order to detect unusual increases in pupil diameter. Fairclough and colleagues found the ICA to not be significantly sensitive to isolated working memory tasks like

the N-back task [101], however Marshall reports positive results for other tasks including surgery and driving [102].

Blinks

A blink not only describes the time period when the eyelid is occluding the pupil, but also the semi-automatic process of closing and re-opening the eyelid . The main function of blinks is to keep the eye lubricated and protect it from irritants, but there is also a connection to cognitive functions [103] and visual processing [104].

Chen and Epps found blinks to indicate visual load well. Though the results for cognitive load were not significant there was a slight trend indicating that blink rate may increase with cognitive load under certain conditions [105]. Hogervorst and colleagues achieved good results using blink rate to separate different levels of cognitive load during an n-back task [9]. Since the n-back task does not increase visual load with difficulty, blinks may carry meaningful information regarding cognitive load.

1.4 Methodological Background

This thesis aims to provide a method that can be applied to other situations and users and therefore needs to provide a model that can make predictions. As a consequence, classic statistical analyses that tests group effects are not the method of choice for this task. Machine learning on the other hand provides exactly what is needed; a model can be trained based on recorded data and the same model can later be employed to make predictions for new participants or situations.

Since the ability to generalize is the focus of my work, the value of different types of eye tracking data varies. Fixation and saccade information emphasizes the gaze location and temporal sequence of gaze points, which are highly dependent on the task at hand and the presented stimulus. Consequently, features related to pupil diameter and microsaccades form the core of information that I rely on. These are also other types of eye related features that haven been shown to be discriminating features for cognitive load in a wide range of tasks, as illustrated in Section 1.3.2.

The datasets that are used in this doctoral thesis are based on tasks that have a clear separation into difficulty levels, so as far as the selection of a machine learn-

ing method is concerned, it seems reasonable to formulate cognitive load estimation as a classification problem. Task difficulty is oftentimes used as a proxy for cognitive load, even though not all participants will experience the same amount of cognitive load for each of the difficulty levels.

For the choice of classification algorithm, generalization is still the main goal. It is therefore necessary to minimize overfitting so the resulting models maintain their ability to generalize. For this reason, I decided to use ensemble methods which are good at avoiding overfitting [106]; more specifically forests of extremely randomized trees (Extra-Trees) [107]. Their tendency to not overfit [108], [109] as fast as other approaches additionally allows the use of more features with the same amount of samples per participant. This is especially important considering that the amount of available data per participant is limited. Moreover, Saez et al. found extremely randomized forest to yield the best results for cross-participant classification of physical activities [110] suggesting that they may be useful for cross-participant approaches. They also compute much faster than regular randomized forests [107], which helps greatly with developing an approach that can be executed in real-time with limited hardware resources. This is essential for a method that may be used in actual applications. If necessary, the number of trees per classifier can also be lowered to save computation time. A further advantage of ensemble methods is that they do not only output a decision for one class, but by virtue of the distribution of individual classifier within the ensemble, a probability for each class can be estimated. In the case of a distinction between low and high cognitive load, class probabilities translate to a value between 0 and 1 that in itself contains additional information beyond a dichotomous distinction. Low values could be interpreted as cognitive load being to low, while values close to 1 may indicate high cognitive overload and values around 0.5 may be optimal for some applications. Additionally, Extra-Trees can provide estimations of feature importance which greatly helps to interpret the results and thereby makes this approach very transparent.

Other methods do not explicitly rely on eye-tracking features, but rather use video material directly. One example of such an end-to-end approach was presented by Fridman et al. [111]. They had drivers perform a n-back as a secondary task while driving and captured their eye movements via a driver facing camera. Facial landmarks were detected and eye regions were extracted based on these landmarks. After transforming eye patches to be frontal facing, sequences of 6

seconds were used as input for a convolutional neural network (CNN). This approach yielded a cross-validated (with participants only being in the training or validation set) accuracy of 86.1% for a three-class problem. These results seem very promising, since they represent cross-participant classification. However, the main feature that is encoded by the CNN is pupil position within the eye - more specifically, horizontal eye movement. Horizontal dispersion has been shown to be a good indicator of cognitive load while driving [112], so it may even be possible to generalize from Fridman's study to other driving scenarios. It is, however, unlikely that the same holds true for tasks that do not involve driving. Therefore, while the accuracy is very good, this end-to-end approach is only suited for a confined set of situations.

End-to-end approaches in general run the risk of overfitting to specific circumstances since there is no theory driven component that can deliberately counteract that. I therefore think, that for a cross-participant and cross-task classifier for cognitive load to work, this may not be the right path as the degree of generalization is questionable. Additionally, a more explicit feature extraction procedure could make use of domain-specific knowledge to reduce the number of features and thereby reduce the amount of data that is needed to train a classifier.

Cross-validation with a strict separation of participants will be key to estimate how well a method works in general. Cross-validation refers to the practice of withholding a portion of the data from the training phase and using it to validate the results by applying the newly trained model to the previously withheld data. Withholding one participant for validation and only relying on data from all other participants ensures that one participant's samples are either in the training set or in the validation set, but never in both. Consequently, leave-one-participant-out cross-validation will be a recurring theme throughout all studies that compose this thesis.

2 Research Questions and Goals

The connection between eye-tracking features and cognitive load is undeniable, but reliably deriving an estimation of cognitive load from eye-based features independent of task and participant is still an open issue. Even though intra-participant results are usually good, researchers repeatedly report failure when it comes to estimating cognitive load across participants or tasks based on physiological signals (e.g. [113]–[116]). This lack of generalization is the obstacles that stands between potential practical applications and their realization.

In this thesis, I will present a partial solution in four studies that build on each other. The overarching goal of this thesis is to present a eye-based classifier for cognitive load that fulfills the following research goals (RG):

RG1 it is accurate

RG2 it is robust

RG3 it is able to generalize

RG4 it can be executed in real-time

RG1 is self-explanatory: If estimations cast by my approach are not accurate then they are useless, so accuracy is the overall measure of success regardless of any additional requirements. Real applications that I aim to facilitate with this thesis will not be used under laboratory conditions, but in real-world scenarios that include noise, unfavorable conditions, and without the luxury of extended calibration phases. This necessitates robustness in order to deal with these problems and motivates RG2. The biggest contribution of this thesis, however, lies in RG3. Although there are methods that perform cross-participant classification to a certain degree and to an even more limited extent cross-task classification, there is no method that does both at the same time, even less so based on pure eye tracking. The core method in this thesis provides a remedy to this problem by introducing a weighted scheme to combine participant-specific models into a

general one, but not without limitations. Finally, to be usable in a adaptive application, a system has to access a user's cognitive load in real-time and therefore, the processing demands for cognitive load estimation need to be low, constituting RG4. This ensure that estimations can be performed frequently, but also that the main application can use the majority of hardware resources.

2.1 Study 0: Detecting Blinks

Study 0 does not investigate the overarching research goal itself, but represents a useful tool for that exact goal. This study presents a reliable way to detect blinks with minimal hardware requirements enabling it to be employed even in low-end devices. As blinks can be an indicator for cognitive load this low-resource solution that shows high accuracy even under difficult circumstances can help to build a fast way to detect cognitive load in real-time. Due to its nature, this algorithm partially pursues RG2 (robustness) and RG4 (real-time capability).

Though the algorithm was developed for head-mounted eye trackers, it is reasonable to assume that it would perform similarly in remote eye trackers. After extraction of the eye region, the presented approach could be applied in the same manner - likely with similar results.

2.2 Study 1: Predicting Cognitive Load during a Working Memory Task across Participants

Study 1 investigates how classification of cognitive load can be performed across participants. It uses a laboratory setting with a controlled environment and a common working memory updating task: the n-back task. This minimizes interference and excludes most sources of noise like changing lighting conditions, visually complex stimuli, or inconsistent cognitive load, thereby increasing data quality.

Aiming for a classification method that works across participants, we first generate reliable classifiers for each participant and combine them in a weighted voting scheme. In this study, we focus on pupil-related measures including median and maximum pupil diameter, blink duration and frequency, as well as the ICA.

Study 1 tackles research goals 1 through 4 - at least in part. Accuracy (RG1) and real-time capability (RG4) are the main objectives in this study as we try to build

a foundation for further research. The controlled circumstances of the experiment limit the robustness (RG2) that we can test and generalization (RG3) can only be tested across participants.

In an exploratory take on RG4, real-time capability is not only evaluated in execution time, but also by simulating data recording and assessing cognitive load in real-time. This controlled and task specific evaluation represents the first small step that is necessary for real-time adaptation based on cognitive load. Additionally, different window sizes for feature extraction are examined.

The approach from study 1 serves as a base for the subsequent studies and represents a prerequisite and first but important step towards the overarching research goal.

2.3 Study 2: Expanding Cross-Participant Cognitive Load Prediction to a Real-World Application

Study 2 takes an approach similar to study 1, but applies it to a scenario with more real-world applicability: an emergency simulation. *Emergency* is a commercial software that can serve as a training simulator but may also be used for entertainment like a video game. This highlights two possible areas of application for adaptive interfaces that react to a user's cognitive load.

Compared to study 1, this study highlights the robustness (RG2) of our approach. Participants were not restrained with a headrest and the baseline that was available is far from ideal. Furthermore, the simulation does not evoke a consistent level of cognitive load and the rapidly changing stimulus adds to the obstacles that need to be overcome. RG3, too, is evaluated beyond its cross-participant aspect by testing classifiers trained on a specific scenario of *Emergency* across different scenarios. While this is not strictly cross-task application, it varies stimuli significantly and can be interpreted as an easier variation of cross-task classification. Still, RG1 and RG4 are addressed in order to get closer to the main goal of this thesis.

Study 2 not only tries to provide solutions to RG1-4, but additionally evaluates heart rate and user activity as exploratory features. A heart rate monitor may be used as an additional sensor due to its inexpensive nature and user activity can be evaluated on a situational basis.

2.4 Study 3: Applying Cognitive Load Classifiers across Participants and Tasks

Finally, study 3 expands the previous work by applying cognitive load classifiers across tasks. Classifiers are trained on n-back-data from study 1 and performance is evaluated for data from study 2 - *Emergency*. The focus in this study is RG3, since classification is attempted across tasks and participants which is the missing piece of generalization. To the author's knowledge this is the only successfully attempt at classifying cognitive load with this level of generalization.

This represents the final step towards the main research goal and offers a solution to RG1-4: an accurate and robust classifier for cognitive load that uses eye tracking data, works across tasks and participants, and is capable of real-time execution.

3 Results and Discussion

The main research goal for this doctoral thesis was to create a classifier for cognitive load that uses eye-tracking data and works independently of participant and task. Additionally, it should be executable in real-time to make it usable for applications and it was required to be robust to withstand the complications and noisy nature of real-world applications.

In Study 0 a robust, low-cost solution to detect blinks was presented that can be used in real-time detection of cognitive load. It provided a method for future application involving cognitive load. Study 1 then served as a proof-of-concept that sets the foundation for the overarching research goal by demonstrating the effectiveness of a novel approach to cross-participant classification of cognitive load in a laboratory setting using a standard working memory task. Building upon the promising results obtained in this controlled environment, Study 2 applied a similar concept to a real-world application with challenging conditions. Finally, Study 3 combined the findings of Study 1 and 2 by applying classifiers that were trained with data from Study 1 to Study 2. The success of this cross-task application represented the last missing piece in pursuit of this thesis' research goal. In the following, I present the main outcomes of the individual studies conducted for this thesis and show how they correspond to the four main research goals.

3.1 Main Outcomes

3.1.1 Study 0

Study 0 did not serve the purpose to detect cognitive load directly, but provided a basic method for blink detection, thereby providing a tool to detect cognitive load. The approach of this study achieved a leave-one-participant-out cross-validated accuracy of 96.35% for a challenging dataset and can be incorporated into future attempts at cognitive load detection.

While this study did not classify cognitive load, blinks play an important role in both cognitive and perceptual load [117]–[119] and hence it contributed to the research goals of this doctoral thesis. With regards to RG1-4, this algorithm can only be evaluated to a limited degree. It had a high accuracy with leave-one-participant-out cross-validation, but since it detects blinks and not cognitive load directly, it can only contribute to RG1 and RG3 by providing features for the subsequent studies.

The same holds true for RG2 (robustness). The approach displayed a high degree of robustness and thereby promotes RG2, if it is employed in a larger framework. The dataset that is used to train the presented classifier was extracted from an on-road driving experiment [43] and therefore was very challenging - containing changing lighting, bad angles, partial blinks, blurring, and reflections. With a recording frequency of 25Hz and a resolution of 384 x 288, the quality of the recordings themselves did also not hold up to the standards of modern Eye Trackers, making blink detection especially difficult.

The approach that was presented in Study 0 used only very basic operations and was therefore fast in its execution. On an i7-4790 at 3.60GHz using 12GB of RAM feature extraction and classification of a single sample took 0.0264ms on average. This would allow real-time calculation even on low-end devices. Incorporating this method into a eye-tracking pipeline could contribute to the whole system being real-time capable and help attain RG4.

Finally, the method in Study 0 was designed for head-mounted eye tracking, while the other three studies used remote eye tracking. When blinks are concerned, remote and head-mounted eye tracking are not too different. An eye region extracted from a remote eye tracker is very similar in structure and appearance to a frame of a head-mounted eye tracker. It therefore is reasonable to assume that an approach that yielded good results in head-mounted eye tracking would also fare well in remote eye tracking and its properties would carry over to this new setting.

3.1.2 Study 1

The main goal in Study 1 was to demonstrate the feasibility of cross-participant classification of cognitive load. Good intra-participant results are common across studies that use eye-tracking features (e.g. [120] with 84% using pupil dilation and blinks), but cross-participant results are rarely above chance level [113]–

[115]. While there are promising results in studies involving EEG or other neuroimaging methods, to the author's knowledge there is no approach that works cross-participant and relies purely on eye tracking.

With Study 1, we achieved a participant-specific accuracy of 82.4% for the distinction of low and high cognitive load (as represented by level 0 and level 2 of the n-back). Levels 0 and 1 accomplished 69.8% and levels 1 and 2 79.4%. As this exceeds the results that other researchers reached purely based on eye tracking (e.g. Haapalainen et al. with 57.4% [116], Shojaeizadeh et al. with 79% [121], or Chen et al. with 53.9% for a three class problem [122]), we can conclude, that RG1 - the classifier being accurate - was accomplished for the participant-specific scenario.

Robustness (RG2) could only be evaluated to a certain degree in Study 1, as it was recorded in a laboratory setting. The task used to induce different levels of cognitive load did not involve any emotionally charged stimuli and - by nature of the task - the amount of cognitive load should be more or less constant throughout a level. These conditions were ideal and therefore did not require any robustness to noise. The period that was used to derive a baseline from was the instruction phase, which should be slightly different for each participant. This added a certain amount of uncontrolled variability and thereby testing the algorithms robustness to a limited degree. The common practice to use a period with no cognitive load at all as a baseline would likely have resulted in slightly higher accuracy.

Study 1 introduced a novel approach to cross-participant classification by combining participant-specific classifiers into a composite classifier that generalizes. Tackling this sub-goal of RG3 was the focus of this study. Accuracy for the classification of level 0 and 2 dropped from 82.4% to 76.8%, whereas levels 0 and 1 and levels 1 and 2 went from 69.8% to 54.0% and from 79.4% to 71.5%, respectively. The decrease in classification accuracy was most pronounced for the distinction between levels 0 and 1, which can likely be attributed to the very small-scale difference between the levels of cognitive load. While results for the intra-participant setting were good, the error introduced by mismatching classifiers had the greatest effect here. Classification for levels 1 and 2 and levels 0 and 2 only suffered minor losses in accuracy when applied cross-participant. The differences in cognitive load were more pronounced in these comparisons meaning that the error for cross-participant application did not impact the results to the same degree. Considering that cross-participant accuracy for the approach presented in this study

trumps intra-participant accuracy for most other eye-tracking based studies, we can conclude that RG1 was still accomplished and we are one step closer to RG3.

Computing features of a segment with length of 5 seconds took on average 0.134 seconds on a machine with 16GB RAM and an i7-7700HQ. Our method was not optimised and did not utilize parallel processing indicating that runtime can be improved further. Even in its current state, 7Hz are attained which should be sufficient to realize real-time adaptations since users rarely experience a significant change in cognitive load that happens this rapidly, hence several cognitive load estimations per second should be sufficient. We can conclude that RG4 was fulfilled in Study 1.

Potential for real-time adaptation was further examined in real-time online classification. To this end, we used participants' data to simulate the recording of a new participant and simultaneously classified cognitive load with our cross-participant approach. The issue that scaling was involved was circumnavigated by joining the simulated data with existing data and scaling them jointly. Using this technique, not all data needed to be available to get accurate scaling and the quality of scaled data improved the more it became available. Accuracy of 70.4%, 53.8%, and 66.8% was achieved for distinctions of levels 0 and 2, levels 0 and 1, and levels 1 and 2, respectively. These results were still only slightly worse than the offline cross-participant classification results highlighting that - under the right circumstances - real-time adaptation may be performed based on eye-tracking recordings. Especially the approach of jointly scaling features may hold promise for real-time processing of eye-tracking data streams.

As expected, longer windows for feature extraction yielded better results for intra-participant classification in all combination of levels. Longer sequences meant that the features are less susceptible to noise making them more robust and consequently improving results. On the other hand, accuracy seemed to be almost unaffected in cross-participant classification, which came as a surprise. One can speculate that the error that is introduced by cross-participant application was the main source of error and it overshadows the accuracy gain from longer sequences.

3.1.3 Study 2

Study 2 focused on demonstrating robustness (RG2) by applying an approach similar to the one used in Study 1 to a challenging real-world dataset. Participants first completed the tutorial for *Emergency* - a commercially available simulation

software - followed by three scenarios each presented in three difficulty levels. Our method had to deal with subpar baselines, varying cognitive load with missing ground truth, and illumination changes caused by the dynamic stimuli. This study also included first results for classification not only across participants but also scenarios - which can be interpreted as a precursor to actual cross-task classification.

In an actual application, obtaining baseline measures would require the user to look at a fixation cross for several minutes which could already discourage them from using this application in the first place. Therefor, we used the tutorial that each participant had to complete as a baseline period circumventing an explicit, conventional baseline. It goes without saying that participants experienced different levels of cognitive load during the tutorial depending on their experience with game-like simulations, their cognitive abilities, their familiarity with input modalities, and many other factors. As a consequence, the baseline level of cognitive load varied greatly between participants which in turn added noise to the whole dataset, because we normalized all features with this baseline.

Further variance was added by the nature of the task itself. The stimuli were dynamic and colors and light changed constantly. This influenced the pupil and possibly other features even though it was not necessarily correlated with cognitive load. On the other hand, cognitive load changed with the state of the simulation and the corresponding time pressure, so cognitive load was not constant throughout one iteration of a scenario. This represented a source of error that can not be compensated for as - in the absence of ground truth - we used difficulty levels as a proxy for cognitive load. For instance, the whole period of scenario 1c was labeled as "high cognitive load" by virtue of it being the difficult version of scenario 1 although cognitive load was likely not constantly high during this period.

With additional noise caused by the deliberate decision not to employ chin rests, we still achieved cross-participant accuracy scores of 80.56%, 70.37%, and 69.81%, for classification of cognitive load in "high" and "low" for the three *Emergency* scenarios. As with Study 1, accuracy indicated that RG1 was fulfilled. Considering complications that had to be overcome and circumstances that were far from ideal, we can further conclude that our approach is robust and satisfies RG2.

It has to be noted though, that not all features were eye related. Participants' activity in the simulation was used as an exploratory feature and feature weights revealed that it carried almost as much information as pupil dilation. It is apparent

that participants had to perform more actions within the simulation if there were more tasks to perform simultaneous and more sub-goals to fulfill. The inclusion of this feature, however, did not trivialize the classification problem, since the results of Study 3 show, that almost the same accuracy can be achieved with eye-related features only (see 3.1.4 and Figure 3.1 for more details). Heart rate was the second exploratory feature, but its feature weights were negligibly, possibly because the window used for feature extraction was only 4 seconds resulting in very few beats per segment. A bigger window might yield better results and allow for more features including heart rate variability.

A feature that was not used in Study 1 but was examined in this study is microsaccade frequency. The sampling rate recommended for reliable detection and analysis of microsaccades is 300Hz [84], however, the eye tracker that we used sampled at 250Hz. This unfortunately means that more complex microsaccade features like peak velocity or magnitudes were not usable, so we restricted microsaccade features to microsaccade frequency. Detecting the occurrence and direction of a microsaccade can be performed at 250Hz as demonstrated by Engbert and Kliegl [123]. Feature weights show that microsaccade frequency is a useful feature for detecting cognitive load, but it was still outperformed by pupil diameter. We can conclude that microsaccades are a meaningful indicator for cognitive load that can be used to complement pupil diameter and blinks.

The transition from participant specific classifiers to general cross-participant classifiers did not lead to significant losses in classification accuracy. Our approach still yielded accuracy scores of 80.56%, 70.37%, and 69.81%, respectively, compared to within-participant scores of 79.03%, 70.14%, and 72.13%. We can conclude that RG3 concerning cross-participant classification was fulfilled.

Progressing towards a cross-task method, our algorithm was tested across the three scenarios of *Emergency*. We applied classifiers trained on easy and difficult samples from one scenario to different scenarios while still operating strictly cross-participant. As expected, the results were worse, but in the majority of cases the accuracy loss is less than three percentage points. This gave hope that RG3 can be satisfied with little loss of accuracy.

Since we dropped the ICA in favor of game activity and heart rate, computation time for feature extraction was a lot faster. On average feature calculation took 0.699ms and classification took another 6.723ms resulting in 7.422ms in total. Again, these runtimes were measured in an i7-770HQ with 16GB RAM and

non-optimized code. The resulting 134.7 classifications per second were almost at the frequency of the eye tracker and should be more than sufficient for real-time applications. This means that RG4 was satisfied in Study 2.

3.1.4 Study 3

The third study aimed at satisfying all four research goals. The missing piece - true cross-task application - was introduced by applying classifiers trained on data from Study 1 to data from Study 2. Furthermore, the feature set was refined to be as independent of the task at hand as possible to enable cross-task application. Finally, we used pupil diameter features, microsaccade frequency, standard deviation of pupil diameter, blink frequency, fixation frequency, and ICA.

Accuracy of this approach was at 69.25% for scenario 1 of *Emergency*, at 63.78% for scenario 2, and at 64.02% for the third scenario. In line with the results of Study 1 and Study 2, one can argue that these results are accurate and thereby satisfy RG1. Since the same conditions as in Study 2 applied to the dataset that classifiers were evaluated on, RG2 (robustness) can also be checked as satisfied. Even more so, since cross-task application introduced further noise that needed to be compensated.

At the core of Study 3 was RG3 - generalization. Applying only cross-participant, but within-task classification using methods and features presented in Study 3, N-back accuracy droped from 79.55% for intra-participant classification to 75.81% for cross-participant application. Results for *Emergency* data almost stayed the same, only decreasing from 71.91% to 71.41%, from 71.98% to 69.34%, and from 68.91% to 67.16% respectively. When going one step further and applying N-back trained classifiers to *Emergency* data, accuracy, again, decreased to 69.25%, 63.78%, and 64.02%, respectively. Even though accuracy decreased, the scores that were achieved are significantly above chance level and show that RG3 - a cognitive load classifier that can generalize - is by no means impossible to attain.

Mean runtime for our algorithm to extract features from a single sample was 59.168ms and classification added another 5.473ms resulting in a total of 66.641ms. Specification of the executing machine were the same as in Study 1 and Study 2: i7-770HQ with 16GB RAM and code that was not optimized for runtime. If no other strain is imposed on the computer, our approach could process 15 samples per second, which should be enough to fulfill the criterion for RG4.

To further prove the validity of our method, we calculated Pearson correlations

between cross-task predictions made by our algorithm and cognitive load scales of the NASA-TLX. Our predictions correlated at 0.484 with subjectively perceived effort, at 0.399 with self-reported cognitive demand, at 0.404 with stress, and at 0.459 with reported temporal demand. All correlations were significant with $p \leq 0.001$ and highlight that our method produces predictions that may be useful in real-wold applications since they correlate with participants' questionnaire data.

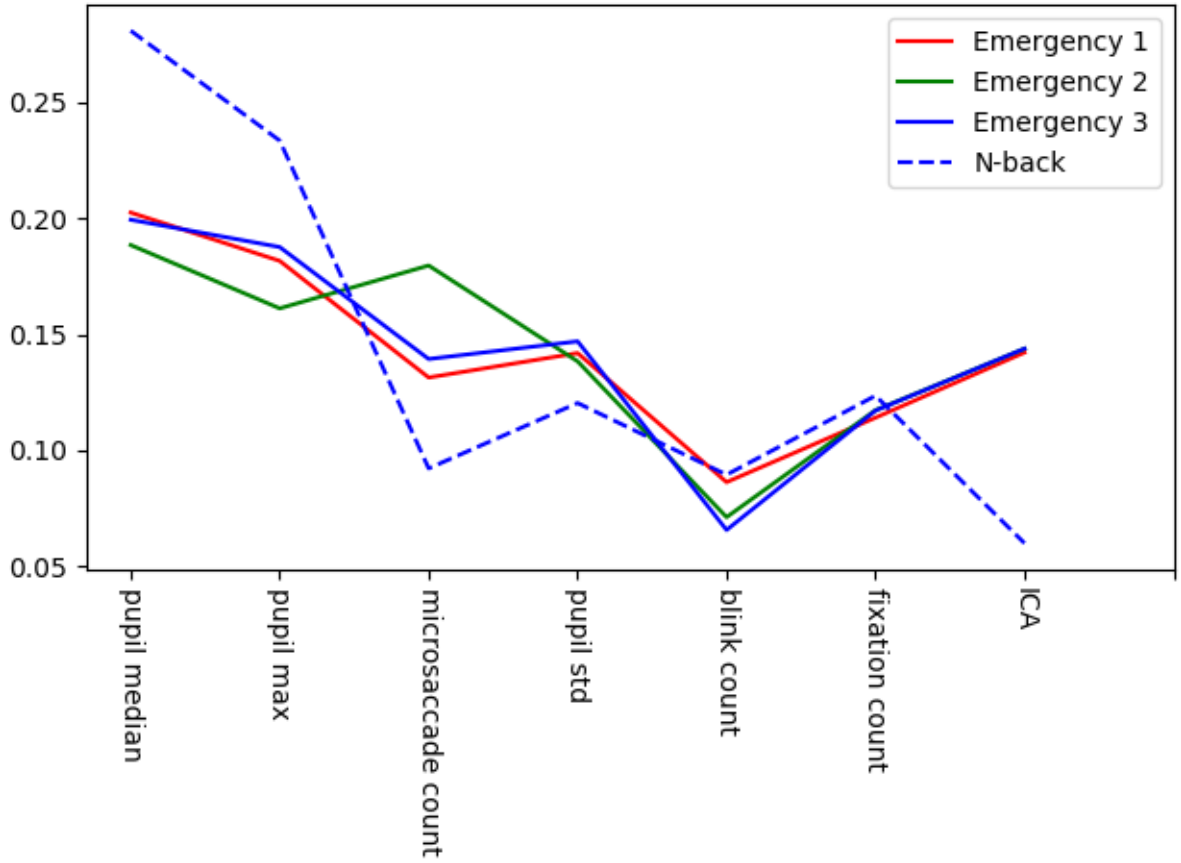


Figure 3.1: Feature weights for different scenarios of *Emergency* and the N-back

A detailed analysis of feature weights revealed that pupil diameter features work in both datasets. This holds true for median pupil diameter, pupil maximum, and standard deviation of pupil diameter. Fixation count also carried information during performance of the N-back task as well es while executing *Emergency*. Both categories of features were supposed to work rather independent of task and stimulus (e.g. summarized in [124]) and as such these findings were not surprising.

Microsaccades had a higher weight in *Emergency* compared to the N-back and therefor may not be universally useful for cognitive load. *Emergency* had a greater

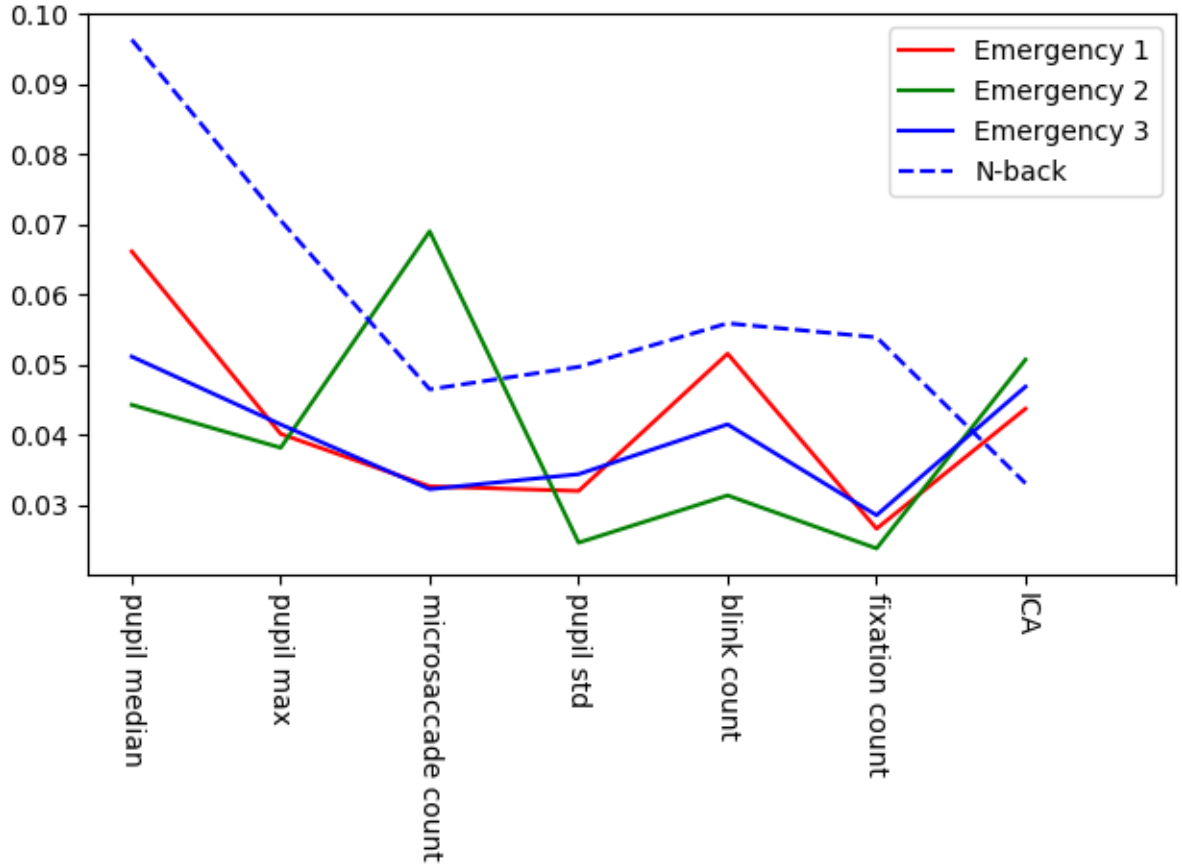


Figure 3.2: Standard deviations of all feature weights for different scenarios of *Emergency* and the N-back

visual demand on participants and required a wider distribution of attention. This is similar to ambient visual search which according to findings by Duchowski and colleagues [125] increases the number of microsaccades. Benedetto and colleagues [81] conducted a driving simulation study and found microsaccades to increase when a secondary task was added compared to a control group with only the main task. They also concluded that microsaccades are linked to visual attention. Krueger et al. drew similar conclusions in their study regarding mental arithmatic and visual load [126].

ICA is the second feature that had diverging feature weights in the two datasets. While performing the N-back task, ICA was not sensitive at all to changes in cognitive demand, but during *Emergency* it had a high feature weight. Demberg reports the ICA to be sensitive to driving and laguage tasks [127]. More difficult driving task increasd the ICA, whereas it decreased with the onset of the secondary lan-

guage task. It may be the case that the ICA only shows an effect in the presence of visual stimuli that impose a higher demand on participants or that a dual-task is needed in order to see an effect. In a multi-media learning study, Korbach and colleagues found no correlation between ICA and learning success or subjective task difficulty ratings [128]. They speculate that the ICA may not be sensitive to higher order cognitive processing. Similarly, Rerhaye et al. found no effect of a Stroop task on the ICA, but could confirm its effectiveness for a spacial processing task [129]. It seems likely that the ICA only reacts to cognitive load under certain circumstances, but more research is needed to shed light on this issue.

3.2 Integrated Discussion

The four studies that comprise my dissertation form a coherent line of research that starts at basic research with a proof of concept and consequently advances towards real-world settings and scenarios. Along the road to this thesis' goals we tested different eye-tracking features and variations of the fundamental approach that is presented in Study 1 and elaborated upon in the two subsequent studies. This section highlights the strengths and limitations of this work and embeds results and findings into the existing body of literature.

3.2.1 Strengths

Over the course of four studies it has become apparent that the approach of participant matching and combining participant-specific classifiers by weighted votes greatly helps with generalization from intra-participant classification to cross-participant classification. Reported accuracy loss was minimal showing how promising this method is compared to naive cross-participant classification that rarely exceeds chance level. Finding participants that are similar regarding their baseline measurements indirectly combats inter-individual differences that may be caused by factors such as physiology or age.

The combination of good cross-participant and cross-task results represents the biggest strength of this work. This level of generalization has not been displayed before for eye-tracking data. As the lack of generalization is what prevents the development of adaptive systems that react to a user's cognitive load, the present thesis may remove this obstacle and enable such systems.

An added benefit of this thesis is the analysis of feature weights that can help to identify what eye-tracking features indicate cognitive load and seem to be independent of stimulus and task. We confirmed the consent in the literature that pupil diameter works for different types of tasks. In addition, fixation frequency also seems to be a feature that consistently is sensitive to cognitive load. This indicates that fixation duration may not be related to visual load exclusively, but could actually be an indicator for cognitive load for a variety of tasks. Feature weights and distributions may also help discern cognitive load from perceptual load. Especially microsaccades seem a promising candidate for visual attention.

A strength of the method presented in this thesis is how short the sequences used for feature extraction can be while still offering good accuracy. A common approach is to average the pupil diameter change throughout one experimental condition or use very long periods of time (e.g. Hogervost et al. with 30, 60, and 120 seconds [9]). In Study 1 we showed that as short as 1s can be enough to collect meaningful data. This means that we can assess cognitive load with a very fine granularity and are able to capture load changes that may be missed with a larger window. It also implies that we can quickly implement adaptations that are chosen based on the last second making our method highly reactive. Additionally, the core method of this thesis only uses features that can be expressed in units per minute, thus being independent of the actual window size. One benefit of this fact is that different window sizes can be employed simultaneously to capture cognitive load within a the last x seconds in order to adapt to micro and macro changes in cognitive load. A second benefit can be exploited during training of the classifiers as training can be performed with longer intervals to enhance robustness and reduce noise.

Relying on eye tracking as a tool to estimate cognitive load offers advantages that other methods lack. On the one hand, it does not have the subjective and disruptive nature of self-reports. On the other hand, it does not impair performance like a dual-task paradigm would. In this regard eye tracking has the same advantages as any other physiological sensor that could be used for cognitive load estimation. As a video based system, (remote) eye tracking does not require direct contact with a participant, which gives it an edge over sensors like EEG or heart rate monitors. It is a sensor that is easy to use and very subtle in its application and therefore suited for a wider range of applications. The fact that eye tracking does not rely on physical contact with a participant helps to keep them in an immersed

state and increases validity of the resulting measurement. Combining this with its ease of use and cheap device options on the market, there is great potential for applications tailored towards users at home.

The fact that the proposed approach works by matching participants to find classifiers that are likely to perform well is one of the biggest strengths of this dissertation. This allows a greater degree of adaptation by virtue of adding more classifiers to the database. Classifiers that are trained with different tasks could be added and comparison between baseline conditions could then include the type of task. This way cross-task errors could be reduced and the algorithm would be more flexible overall. Similarly, different environmental conditions like time of day, ambient light, or screen brightness could be added.

3.2.2 Limitations

Missing ground truth measurement for cognitive load is a problem for all researchers trying to estimate cognitive load. Task performance or task difficulty are commonly used as proxies for degrees or levels of cognitive load, but both do not reflect the interplay between effort and skill. Self-reports on the other hand take into account ones individual skills and cognitive capacity, but are subjective and participants may rate their load on very different scales. Physiological sensors could provide ground truth as they are objective, but to interpret physiological signals validly and reliably and establish them as representing ground truth, an existing ground truth would have to be available to justify that interpretation. To the author's knowledge, there is no way of avoiding proxies in favor of any ground truth.

A major limitation that needs to be tackled in future research is the reliance on scaling. This limitation is twofold in nature. Firstly, we need a certain amount of data from one participant until scaling can be performed in a meaningful manner. If little data is available, scaling may skew the participant-specific classifier and render its predictions more harmful than useful as they could be completely off. Secondly, for scaling to be meaningful, we need similar distributions for all participants. This also entails the need for roughly equal amounts of time with low and high cognitive load. At the very least there is the need for any amount of low and high cognitive load to be present within the participant's data or else we run into problems of scaling. Imagine a participant only looking at a fixation cross and then Z-normalizing their data. Such data is meaningless once scaled.

A consequence of the limitation of scaling is that online assessment - as it would be needed for real-time adaptation - is only possible if we already have a database of classifiers that were trained under the same circumstances. Study 1 illustrated that cross-participant online classification can be performed, but only under very specific circumstances. The online estimation in Study 1 relied on joint scaling features from two participants and slowly transitioning to intra-participant scaling as more data becomes available. Before this transition, joint scaling only works reliably if the conditions during baseline recording were very similar or - ideally - identical. After the transition, scaling only works if the specific participant experienced both high and low cognitive load limiting the use of this method for real applications where this may not be guaranteed.

Another limitation is the small range of participants that were used in this thesis. All participants were students of the University of Tübingen with a limited spectrum of age and cognitive abilities. If the method was to be used in an application, one would need to ensure that it also works for users from a general population. This limitation, however, could be overcome with a larger dataset of N-back participants. If age is added to the similarity rating, appropriate classifiers can be used and a accuracy comparable to the one reported in this thesis can be expected.

A limitation similar to the small spectrum of participants is the small range of tasks. We only used two datasets from different tasks so one can argue that this transfer is not general but possibly restricted to these two tasks. Furthermore, both tasks involved visual stimuli so the approach may not work for other modalities of presentation. However, the features that receive high weights are mainly pupil diameter features, so they should be independent of the task used to invoke cognitive load as task-evoked pupillary responses were found to fulfil Kahneman's three criteria for indicating processing load [88], [130]. So even if other features fail, pupil diameter features should still yield good results, preserving at least a part of the accuracy presented in this thesis. Zekveld and colleagues also found pupil diameter to be a reliably indicator of higher cognitive load in dozens of studies that used auditory stimuli, so the argument that an increase in pupil diameter indicates cognitive load universal may apply to modalities, too [71]. Other features such as microsaccades and ICA need more investigation before they can be included in general eye-tracking features for cognitive load.

A drawback that any pupil related approach has is its high number of confound-

ing factors, especially light and emotion. This usually limits the application of pupil based algorithms to laboratory settings with controlled lighting and stimuli that do not have a strong emotional connotation. This is the only way to ensure that the pupil diameter change is caused by cognitive load by means of excluding - or at least minimizing - any other possible cause. Light plays an especially important role when it comes to confounding factors. If there is no way to compensate for the influence of light, the whole experiment has to be performed with the same lighting conditions which even limits dynamic stimuli as a dark or bright screen can cause pupil dilation that far exceeds those typically seen for cognitive load [131]–[133]. Emotions certainly influence the pupil [134], [135], but researchers found that emotional arousal only shows via pupil dilation under low cognitive load [136] and that the effect of cognitive load is dominating over the effect of emotional arousal [137]. This indicates that for most situations the effect of emotion only plays a secondary role and can be disregarded in certain applications and circumstances.

3.2.3 Theoretical Embedding

Cognitive Load Theory

From a Cognitive Load Theory perspective, varying task difficulty in order to evoke different levels of cognitive load is equivalent to manipulating intrinsic cognitive load while extraneous load is kept constant. The amount of germane load, however, can still change depending on the task. Melby-Lervåg and colleagues meta review on working memory training concludes that there is no significant transfer to other skills and only a short-term, specific training effect [138], which in turn suggests that germane load is not applicable to working memory tasks. Therefore, Study 1 should not involve germane load. This indicates that the physiological changes that were measured in Study 1 only reflect variations in intrinsic load.

Intrinsic load was also the major type of load that was manipulated in Study 2, but extraneous and germane load need a lot more attention in this study. Even though all participants completed the same tutorial, not all participants had the same foundation to work with during the simulation. Their previous gaming experience and corresponding skills influenced their intrinsic load as well as their extraneous load. Familiarity with input modalities and interfaces commonly used in video games makes *Emergency* easier to use and lowers extraneous load as the

presentation is familiar and controls are more intuitive. Likewise, prior knowledge and experience with games in general and strategy simulations in particular lower a participant's perceived difficulty while simultaneously enhancing their task performance - they lower intrinsic load. As none of the participants had performed the task before and since there was a significant transfer between scenarios and difficulty levels, it is apparent that Study 2 involved germane load. The presence of all three types of cognitive load and their different distribution for each participant and scenario makes the analysis of *Emergency* inherently more complex than Study 1.

Since general familiarity with games does no change over time, task difficulty manipulation in Study 2 is primarily limited to intrinsic load and only to a minor degree to extraneous and germane load. As participants complete more scenarios of *Emergency*, they are better acquainted with the controls and interface, thereby reducing extraneous load to a small degree. The cognitive capacity not occupied by intrinsic or extraneous load can be used for germane load - learning and developing schemata to deal with future variations of the *Emergency* task.

One can notice that cross-task application of N-back classifiers to scenario 1 of *Emergency* barely reduced accuracy compared to within-task, within-participant classifiers for *Emergency* and within-task, cross-participant classifiers. This indicates that the way cognitive load is expressed by the selected eye-tracking features was very similar between these tasks and between participants and consequently hints at a similar profile of cognitive load. We, however, expected the distribution of cognitive load to be different since Study 1 was likely only varying intrinsic load and Study 2 evoked different levels of all three types of load. This suggests that classifiers trained on N-back data are not specific to intrinsic load but react to the total amount of cognitive load. This seems to be in line with findings by Zu et al. that indicate that pupil diameter changes are affected by extraneous load and germane load [139].

To shed more light on the transfer ability of N-back classifiers, the higher accuracy loss for cross-task application for scenarios 2 and 3 of *Emergency* have to be examined. Considering that average completion rates for scenario 1 were 83.33% for both, easy and hard versions, it seems unlikely that many participants suffered from cognitive overload in this specific scenario, since completion rate for the hard variant is still high. On the other hand, scenarios 2 and 3 had a completion rate for their easy version of 97.22% and a rates of 36.11% and 33.33% respectively for

their hard versions. This makes it highly likely that many participants experienced intrinsic cognitive overload caused by the high difficulty of this part of the task. Chen and Epps report a decrease in pupil diameter for a visual search task that induced intrinsic overload with its visual demand [105], suggesting that this could also be the case in scenarios 2 and 3. Similar findings for a short term memory task by Johnson and colleagues further support this hypothesis [140]. As a consequence, N-back trained classifiers that rely primarily on pupil diameter would fare worse in scenario 2 and 3 and seem to be best suited for situations that can cause high cognitive load, but no overload.

Multiple Resource Theory

Through the lens of Wicken's multiple resource theory, the N-back task used a single perceptual modality (the focal visual channel), focused on spacial cognition, and required a manual response. The strain on perception should be negligible and most resources should be used in the cognition stage, especially since the mode and difficulty of perception were not altered during presentation of this task. Manipulation of task difficulty focused in cognition and important features in Study 1 consequently represented that.

As with Study 1, Study 2's *Emergency* also used the visual perception channel, but it made use of focal as well as ambient visual perception. Similarly, Study 2 mainly used spacial cognition and required a manual output. These two studies seem to be very similar in the structure of their resource demand, but Study 2 was more complex. Varying task difficulty was performed by adding more available emergency personnel and more sub-tasks that needed to be completed. It did, therefore, not only change the strain on the cognition stage, but more difficult scenarios and version featured a lot more visual perception load. Furthermore, a lot more and faster responses were required in order to successfully complete difficult versions of scenarios.

The manipulation in cognitive demand during Study 1 was well reflected in the pupil dilation features. In Study 2, microsaccades and ICA had a significantly higher importance while pupil dilation features still played an important role. This discrepancy in feature importance may be caused by the manipulation in perceptual load that happened in Study 2. Microsaccades are linked to visual attention [125], [141] which played a crucial role in Study 2, but not in Study 1. The difference regarding ICA was not so clear since it did not seem to be tied to visual

demand, but task demand. However, pupil fluctuations were not indicative of task difficulty in Study 1, which again gives rise to the question what exactly ICA and IPA are sensitive to.

Cross-task application of N-back trained classifiers to the first scenario of Study 2 maintained a high accuracy, while it decreased for the second and third scenario. Since cross-scenario results in Study 2 barely reduced accuracy, this result is puzzling. All three scenarios seem to be similar enough for classifiers to work for all of them, but classifiers trained on N-back data can only be applied to scenario 1 without a significant loss.

83.33% of participants successfully completed the hard version of scenario 1 indicating that most of them were not overloaded and allowing the difficulty manipulation to manifest itself equally for perception and cognition. Since N-back classifiers are mainly sensitive to strains in the cognition stage, their perform well for scenario 1.

The high importance of microsaccades in scenario 2 combined with the low rate of success for the hard scenario could mean that participants' overload can be attributed to perception instead of cognition. Fire as an additional hazard (that can spread if not controlled) was first introduced in scenario 2 and could be the source of perceptive overload providing a possible explanation for this observation.

Scenario 3 showed a feature importance distribution very similar to scenario 1. The emphasis on microsaccades that can be observed for scenario 2 is no longer apparent, meaning that perception overload does no longer seem to be the discriminating factor. Meanwhile, only 33.33% of all participants completed the hard version of scenario 3 successfully, suggesting that many were overloaded with the combination of perceptual and cognition load. Also, according to participant's self-reports, cognitive demand and invested effort were already high for the easy version. Since features of all scenarios were normalized with the same baseline, this means that the difference in eye-tracking features between the easy and hard version were less pronounced in scenario 3 and therefore the separation into the two classes worked less accurate. This may be amplified further by the fact that the step from maximum load to overload reduces the features that typically show high cognitive load (see [105], [140]).

For future work it may be advantageous to estimate the load on perception and cognition separately and combine them in accordance with Wicken's multiple resource theory to get a more complete picture about a participant's load situation.

3.3 Contribution of this Thesis

At its core, this thesis contributes to the field of computer science, but the applications that can be derived from this contribution extend into psychology, educational science, and potentially commercial applications.

Cognitive load estimation based on eye tracking has been investigated for decades, but generalization has always been an issue. Here lies the main contribution of my doctoral thesis. It offers methodological remedies to combat the inadequacy of conventional approaches. Cross-participant performance of the presented core method is only marginally worse compared to within-participant accuracy and seems to work well across different tasks. This degree of generalization has not been achieved by other researchers making this thesis a valuable methodological contribution to the field of computer science.

Furthermore, the presented method is robust to noise and complications that real-world application entail. It is also computationally inexpensive allowing it to be executed concurrently with an application that has higher demands. It's low hardware demand additionally allows it to be used in real-time, enabling interactive systems to react to a user's current cognitive state. Combining all these strengths, the core method of this doctoral thesis could serve as a basis for a large number of applications in various areas.

The impact of this work, however, could reach far beyond computer science. It represents a tool that can be applied to different fields provided that eye tracking is involved. Cognitive load is an important factor in learning and instructional design and correctly assessing cognitive load in real-time is a prerequisite for many adaptive human-computer interfaces. Especially in the field of e-learning the applications are manifold. As it is desirable to keep learners in the *zone of proximal development* [3], we first have to detect when they leave it. This happens when a task is too difficult for a learner to be completed with assistance of the e-learning system or when they easily can do it on their own. It can also be interpreted in terms of cognitive load: If a task cognitively overloads a learner, they will not benefit from it, but if the task or learning material does not pose a challenge, their learning outcomes will be equally subpar. As a consequence, adapting a system to suit a learner's current needs may help them learn at their optimal rate.

Following this train of thought, assessing cognitive load in real-time is a first step necessary for adapting a learning interface to fit the learners' needs. If a systems detects a overwhelming amount of cognitive load while learning with texts,

the system could adjust the selection of available text to learn from accordingly. Similar methods could be applied with selection of mathematical exercises or vocabulary training.

If one was to reverse the idea, this method can be used as an evaluation tool. If the material is fixed, assessing cognitive load can serve as a proxy for difficulty as cognitive load increases with task demands. This way a learning task can be evaluated even if it does not feature a performance metric.

There are also applications outside of education. Any computer system that exposes a user to cognitive load could employ eye tracking to make use of cognitive load estimation. During driving cognitive load could be monitored and incorporated in a model that estimates fatigue or exhaustion. If a driver is experiencing high cognitive demands for a prolonged time, performance may suffer and the risk of error increases. Here a prompt to either take a break or switch drivers can be made to help ensure the safety of all involved parties.

3.4 Implications and Future Research

The current thesis can serve as a point of origin for many improvements and applications. The biggest limiting factor of this thesis is that online cognitive load estimation is only viable under very specific circumstances using the presented methods. At the core of this limitation is the need to scale features to make participants easier to compare. Study 1 counteracted this issue to a certain degree by jointly scaling data of multiple participants, but only to a certain degree. In order to not rely on scaling anymore, we need to be able to computationally eliminate inter-individual differences. In part this task is accomplished by deducting the baseline, but this only helps for within-task classification. It is difficult to compare participants' baselines across tasks since the recording conditions are different (especially lighting) and as Pfleging et al. have demonstrated, the magnitude of pupil dilation is dependent on the light and baseline condition [133], which is what made scaling pupil diameter features necessary in the first place. The self-evident solution to this problem is to incorporate an illumination model into our algorithm that for a given screen illumination and ambient lighting outputs the expected pupil diameter. There has been research conducted on this topic and a promising candidate for a pupil model seems to be the unified model for light-adapted pupil sizes by Watson and colleagues [142]. If for any given moment we

have an accurate estimate of the expected pupil size, we can use it as a baseline eliminating the need for a dedicated baseline recording and improve comparability between participants. Once this is achieved, we can also start jointly scaling provided that participants are similar enough.

More research has to be conducted regarding baselines in general - not limited to pupil dilation. It is unclear what baseline operation works best for eye-tracking features that are not directly related to average pupil diameter. As Mathôt et al. suggest, pupil dilation is best corrected by subtraction and not division [97]. It is, however, unclear how other features like blinks, ICA/IPA, or microsaccades are best corrected. Furthermore, compared to pupil dilation, it is not well researched, how environmental factors influence these other features and if we can devise a model that helps us to reduce these effects.

A major factor that influences cognitive performance and eye related measurements is fatigue [143]–[145]. For a true out-of-the-box solution to work, the influence of fatigue has to be taken into account and integrated into the cognitive load detection model. A possible way would be to monitor cognitive load over time and estimate fatigue based on these observations. It has been suggested by Mizuno and colleagues that prolonged cognitive load induces mental fatigue and is associated with sympathetic hyperactivity based on decreased parasympathetic activity [146]. Grounded in this observation, it seems plausible to devise a cognitive load model that uses an estimation of the level of cognitive load over time to - in turn - estimate fatigue. More research has to be conducted, however, in order to confirm this hypothesis and develop such a model.

As the feature weight differences and cross-task results indicate, it many be useful to investigate cognitive load and perceptual load separately. Both can represent a bottleneck in overall load and considering them separately could help disentangle them and take appropriate actions.

One of the most apparent avenues for future work that builds on this thesis is expanding the database of classifiers. In theory it is possible to achieve the same accuracy as within-task, within-participant classifiers, if there are just enough different classifiers available to choose from. Once the database contains data from multiple task, lighting conditions, and environmental circumstances for a wide array of participants, one will always find a classifier that is very similar and thus should be accurate for the current situation. It goes without saying that sufficiently populating this enormous classifier space is an extraordinary laborious task and is

likely not the most efficient way to solve this problem. but for a limited application this brute force approach may yield acceptable results until a more efficient solution can be engineered.

Considering that this thesis shows that cross-participant cognitive load estimation works without significant loss of accuracy, it could already be deployed in adaptive applications if the application's scope is small enough and the nature of the application is suited. A small set of initial participants would be needed to train classifiers which then will be employed for the actual application. Study 1 suggests that such an approach could give decent results, but pilot testing and additional research are needed to confirm this.

3.5 Conclusion

In summary, the present doctoral thesis shows certain limitations - some are due to eye tracking and the nature of pupil signals, others are methodological. However, these limitation can be addressed in future research using this thesis as a foundation. The four studies of this thesis have demonstrated that it is possible to use eye tracking as a tool to estimate cognitive load in a way that is accurate (RG1), robust (RG2), works across tasks and participants (RG3), and is capable of execution in real-time (RG4). Pupil related features still seem to be the gold standard when it comes to cognitive load estimation, but there is promise in pupil fluctuation measures like ICA or IPA as well as microsaccades. Moreover, Studies 2 and 3 showed that eye-tracking based approaches can cope with real-world applications that are inherently more noisy than laboratory tasks.

Successfully combining cross-task and cross-participant classification of cognitive load based on eye tracking, the present doctoral thesis demonstrates a level of generalization that has not been displayed before, rendering it a valuable contribution and tool with high potential. The low hardware requirements and real-time capabilities further add to this fact. It represents a step towards out-of-the-box solutions for the detection of cognitive load that could be the foundation for a multitude of applications.

References

[1] J. Sweller, J. J. Van Merrienboer, and F. G. Paas, "Cognitive architecture and instructional design," *Educational psychology review*, vol. 10, no. 3, pp. 251–296, 1998.

[2] P. Gerjets, C. Walter, W. Rosenstiel, and M. Bogdan, "Cognitive state monitoring and the design of adaptive instruction in digital environments: Lessons learned from cognitive workload assessment using a passive brain-computer interface approach," Jan. 2014. DOI: 10.15496/publikation-821.

[3] S. Chaiklin, "The zone of proximal development in vygotsky's analysis of learning and instruction," *Vygotsky's educational theory in cultural context*, vol. 1, pp. 39–64, 2003.

[4] M. Csikszentmihalyi, S. Abuhamdeh, J. Nakamura, *et al.*, *Flow*, 1990.

[5] K. Kiili, A. Lindstedt, and M. Ninaus, "Exploring characteristics of students' emotions, flow and motivation in a math game competition," Jan. 2018.

[6] J. L. Lobo, J. D. Ser, F. De Simone, R. Presta, S. Collina, and Z. Moravek, "Cognitive workload classification using eye-tracking and eeg data," in *Proceedings of the International Conference on Human-Computer Interaction in Aerospace*, ACM, 2016, p. 16.

[7] G. F. Wilson, C. Russell, J. Monnin, J. Estepp, and J. Christensen, "How does day-to-day variability in psychophysiological data affect classifier accuracy?" In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, SAGE Publications Sage CA: Los Angeles, CA, vol. 54, 2010, pp. 264–268.

[8] J. Heard, C. E. Harriott, and J. A. Adams, "A survey of workload assessment algorithms," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 5, pp. 434–451, Oct. 2018. DOI: 10.1109/THMS.2017.2782483.

[9] M. A. Hogervorst, A.-M. Brouwer, and J. B. F. van Erp, "Combining and comparing eeg, peripheral physiology and eye-related measures for the assessment of mental workload," *Front Neurosci*, vol. 8, p. 322, Oct. 2014, 25352774[pmid], ISSN: 1662-4548. DOI: 10.3389/fnins.2014.00322. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4196537/.

[10] C. D. Wickens, "The structure of attentional resources," *Attention and performance VIII*, vol. 8, pp. 239–257, 1980.

[11] C. D. Wickens, "Processing resources and attention," *Multiple-task performance*, vol. 1991, pp. 3–34, 1991.

[12] J. Sweller, "Cognitive load theory, learning difficulty, and instructional design," *Learning and instruction*, vol. 4, no. 4, pp. 295–312, 1994.

[13] J. L. Plass, R. Moreno, and R. Brünken, *Cognitive load theory*. Cambridge university press, 2010.

[14] N. Cowan, "Working memory underpins cognitive development, learning, and education," *Educational psychology review*, vol. 26, no. 2, pp. 197–223, 2014.

[15] N. Hollender, C. Hofmann, M. Deneke, and B. Schmitz, "Integrating cognitive load theory and concepts of human–computer interaction," *Computers in human behavior*, vol. 26, no. 6, pp. 1278–1288, 2010.

[16] R. Brünken, S. Steinbacher, J. L. Plass, and D. Leutner, "Assessment of cognitive load in multimedia learning using dual-task methodology.," *Experimental psychology*, vol. 49, no. 2, p. 109, 2002.

[17] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. Van Gerven, "Cognitive load measurement as a means to advance cognitive load theory," *Educational psychologist*, vol. 38, no. 1, pp. 63–71, 2003.

[18] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Advances in Psychology*, vol. 52, Elsevier, 1988, pp. 139–183.

[19] R. D. McKendrick and E. Cherry, "A deeper look at the nasa tlx and where it falls short," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 62, no. 1, pp. 44–48, 2018. DOI: 10.1177/1541931218621010. eprint: https://doi.org/10.1177/1541931218621010. [Online]. Available: https://doi.org/10.1177/1541931218621010.

[20] R. Brunken, J. L. Plass, and D. Leutner, "Direct measurement of cognitive load in multimedia learning," *Educational psychologist*, vol. 38, no. 1, pp. 53–61, 2003.

[21] P. Antonenko, F. Paas, R. Grabner, and T. Van Gog, "Using electroencephalography to measure cognitive load," *Educational Psychology Review*, vol. 22, no. 4, pp. 425–438, 2010.

[22] D. Popovic, M. Stikic, T. Rosenthal, D. Klyde, and T. Schnell, "Sensitive, diagnostic and multifaceted mental workload classifier (physioprint)," vol. 9183, Aug. 2015. DOI: 10.1007/978-3-319-20816-9_11.

[23] Z. Wang, R. M. Hope, Z. Wang, Q. Ji, and W. D. Gray, "Cross-subject workload classification with a hierarchical bayes model," *NeuroImage*, vol. 59, no. 1, pp. 64–69, 2012.

[24] Y. Ke, H. Qi, F. He, S. Liu, X. Zhao, P. Zhou, L. Zhang, and D. Ming, "An eeg-based mental workload estimator trained on working memory task can work well under simulated multi-attribute task," *Frontiers in Human Neuroscience*, vol. 8, p. 703, 2014, ISSN: 1662-5161. DOI: 10.3389/fnhum.2014.00703. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnhum.2014.00703.

[25] Y. Santiago-Espada, R. R. Myer, K. A. Latorella, and J. R. Comstock Jr, "The multi-attribute task battery ii (matb-ii) software for human performance and workload research: A user's guide," 2011.

[26] L. R. Krol, S.-C. Freytag, M. Fleck, K. Gramann, and T. O. Zander, "A task-independent workload classifier for neuroadaptive technology: Preliminary data," in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, 2016, pp. 003 171–003 174.

[27] L. M. Andreessen, P. Gerjets, D. Meurers, and T. O. Zander, "Toward neuroadaptive support technologies for improving digital reading: A passive bci-based assessment of mental workload imposed by text difficulty and

presentation speed during reading," *User Modeling and User-Adapted Interaction*, pp. 1–30, 2020.

[28] X. Zhang, L. R. Krol, and T. O. Zander, "Towards task-independent workload classification: Shifting from binary to continuous classification," in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, 2018, pp. 556–561.

[29] P. Sarkar, K. Ross, A. J. Ruberto, D. Rodenbura, P. Hungler, and A. Etemad, "Classification of cognitive load and expertise for adaptive simulation using deep multitask learning," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 2019, pp. 1–7.

[30] E. Huey, *The Psychology and Pedagogy of Reading (Reprint)*. MIT Press, 1968 (originally published 1908).

[31] G. T. Buswell, *How adults read*, 45. University of Chicago, 1937.

[32] A. L. Yarbus, *Eye Movements and Vision*. Plenum Press, 1967.

[33] E. D. Reichle, A. E. Reineberg, and J. W. Schooler, "Eye movements during mindless reading," *Psychological science*, vol. 21, no. 9, pp. 1300–1310, 2010.

[34] P. C. Gordon, R. Hendrick, M. Johnson, and Y. Lee, "Similarity-based interference during language comprehension: Evidence from eye tracking during reading.," *Journal of experimental psychology: Learning, Memory, and Cognition*, vol. 32, no. 6, p. 1304, 2006.

[35] D. Beymer, P. Z. Orton, and D. M. Russell, "An eye tracking study of how pictures influence online reading," in *IFIP Conference on Human-Computer Interaction*, Springer, 2007, pp. 456–460.

[36] D. M. Krugman, R. J. Fox, J. E. Fletcher, and T. H. Rojas, "Do adolescents attend to warnings in cigarette advertising? an eye-tracking approach," *Journal of advertising research*, vol. 34, no. 6, pp. 39–53, 1994.

[37] G. Hervet, K. Guérard, S. Tremblay, and M. S. Chtourou, "Is banner blindness genuine? eye tracking internet text advertising," *Applied cognitive psychology*, vol. 25, no. 5, pp. 708–716, 2011.

[38] N. Scott, C. Green, and S. Fairley, "Investigation of the use of eye tracking to examine tourism advertising effectiveness," *Current Issues in Tourism*, vol. 19, no. 7, pp. 634–642, 2016.

[39] C. Katsanos, N. Tselios, and N. Avouris, "Evaluating website navigability: Validation of a tool-based approach through two eye-tracking user studies," *New review of Hypermedia and Multimedia*, vol. 16, no. 1-2, pp. 195–214, 2010.

[40] P. Majaranta and A. Bulling, "Eye tracking and eye-based human–computer interaction," in *Advances in physiological computing*, Springer, 2014, pp. 39–65.

[41] A. Papoutsaki, P. Sangkloy, J. Laskey, N. Daskalova, J. Huang, and J. Hays, "Webgazer: Scalable webcam eye tracking using user interactions," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence-IJCAI 2016*, 2016.

[42] E. Kasneci, G. Kasneci, T. C Kübler, and W. Rosenstiel, "Online recognition of fixations, saccades, and smooth pursuits for automated analysis of traffic hazard perception," in *Artificial neural networks*, Springer, 2015, pp. 411–434.

[43] E. Kasneci, K. Sippel, K. Aehling, M. Heister, W. Rosenstiel, U. Schiefer, and E. Papageorgiou, "Driving with binocular visual field loss? a study on a supervised on-road parcours with simultaneous eye and head tracking," *PloS one*, vol. 9, no. 2, e87470, 2014.

[44] E. Kasneci, K. Sippel, K. Aehling, M. Heister, W. Rosenstiel, U. Schiefer, and E. Papageorgiou, "Driving with binocular visual field loss? a study on a supervised on-road parcours with simultaneous eye and head tracking," *PloS one*, vol. 9, no. 2, e87470, 2014.

[45] C. Braunagel, E. Kasneci, W. Stolzmann, and W. Rosenstiel, "Driver-activity recognition in the context of conditionally autonomous driving," in *Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on*, IEEE, 2015, pp. 1652–1657.

[46] C. Braunagel, W. Rosenstiel, and E. Kasneci, "Ready for take-over? a new driver assistance system for an automated classification of driver take-over readiness," *IEEE Intelligent Transportation Systems Magazine*, vol. 9, no. 4, pp. 10–22, 2017.

[47] C. Braunagel, D. Geisler, W. Rosenstiel, and E. Kasneci, "Online recognition of driver-activity based on visual scanpath classification," *IEEE Intelligent Transportation Systems Magazine*, vol. 9, no. 4, pp. 23–36, 2017.

[48] E. Bozkir, D. Geisler, and E. Kasneci, "Assessment of driver attention during a safety critical situation in vr to generate vr-based training," in *ACM Symposium on Applied Perception 2019*, 2019, pp. 1–5.

[49] T. C. Kübler, E. Kasneci, W. Rosenstiel, U. Schiefer, K. Nagel, and E. Papageorgiou, "Stress-indicators and exploratory gaze for the analysis of hazard perception in patients with visual field loss," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 24, pp. 231–243, 2014.

[50] E. Kasneci, T. Kübler, K. Broelemann, and G. Kasneci, "Aggregating physiological and eye tracking signals to predict perception in the absence of ground truth," *Computers in Human Behavior*, vol. 68, pp. 450–455, 2017.

[51] M. H. Baccour, F. Driewer, E. Kasneci, and W. Rosenstiel, "Camera-based eye blink detection algorithm for assessing driver drowsiness," in *2019 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2019, pp. 987–993.

[52] E. Bozkir, D. Geisler, and E. Kasneci, "Person independent, privacy preserving, and real time assessment of cognitive load using eye tracking in a virtual reality setup," in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, IEEE, 2019, pp. 1834–1837.

[53] B. Hosp, S. Eivazi, M. Maurer, W. Fuhl, D. Geisler, and E. Kasneci, "Remoteeye: An open-source high-speed remote eye tracker," *Behavior Research Methods*, pp. 1–15, 2020.

[54] T. Čegovnik, K. Stojmenova, G. Jakus, and J. Sodnik, "An analysis of the suitability of a low-cost eye tracker for assessing the cognitive load of drivers," *Applied ergonomics*, vol. 68, pp. 1–11, 2018.

[55] J. Coyne and C. Sibley, "Investigating the use of two low cost eye tracking systems for detecting pupillary response to changes in mental workload," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 60, no. 1, pp. 37–41, 2016. DOI: 10.1177/1541931213601009. eprint: https://doi.org/10.1177/1541931213601009. [Online]. Available: https://doi.org/10.1177/1541931213601009.

[56] T. Santini, W. Fuhl, and E. Kasneci, "Pure: Robust pupil detection for real-time pervasive eye tracking," *Computer Vision and Image Understanding*, vol. 170, pp. 40–50, 2018.

[57] T. Santini, W. Fuhl, and E. Kasneci, "Purest: Robust pupil tracking for real-time pervasive eye tracking," in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, 2018, pp. 1–5.

[58] W. Fuhl, T. Santini, G. Kasneci, and E. Kasneci, "Pupilnet: Convolutional neural networks for robust pupil detection," *arXiv preprint arXiv:1601.04902*, 2016.

[59] W. Fuhl, T. Santini, G. Kasneci, W. Rosenstiel, and E. Kasneci, "Pupilnet v2. 0: Convolutional neural networks for cpu based real time robust pupil detection," *arXiv preprint arXiv:1711.00112*, 2017.

[60] W. Fuhl, D. Geisler, T. Santini, T. Appel, W. Rosenstiel, and E. Kasneci, "Cbf: Circular binary features for robust and real-time pupil center detection," in *Proceedings of the 2018 ACM symposium on eye tracking research & applications*, 2018, pp. 1–6.

[61] F. Manuri, A. Sanna, and C. P. Petrucci, "Pdif: Pupil detection after isolation and fitting," *IEEE Access*, vol. 8, pp. 30 826–30 837, 2020.

[62] T. Santini, W. Fuhl, and E. Kasneci, "Calibme: Fast and unsupervised eye tracker calibration for gaze-based pervasive human-computer interaction," in *Proceedings of the 2017 chi conference on human factors in computing systems*, 2017, pp. 2594–2605.

[63] T. Santini, D. C. Niehorster, and E. Kasneci, "Get a grip: Slippage-robust and glint-free gaze estimation for real-time pervasive head-mounted eye tracking," in *Proceedings of the 11th ACM symposium on eye tracking research & applications*, 2019, pp. 1–10.

[64] V. Cantoni, C. Galdi, M. Nappi, M. Porta, and D. Riccio, "Gant: Gaze analysis technique for human identification," *Pattern Recognition*, vol. 48, no. 4, pp. 1027–1038, 2015.

[65] N. Sammaknejad, H. Pouretemad, C. Eslahchi, A. Salahirad, and A. Alinejad, "Gender classification based on eye movements: A processing effect during passive face viewing," *Advances in cognitive psychology*, vol. 13, no. 3, p. 232, 2017.

[66] J. Steil, I. Hagestedt, M. X. Huang, and A. Bulling, "Privacy-aware eye tracking using differential privacy," in *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, 2019, pp. 1–9.

[67] E. Bozkir, A. B. Ünal, M. Akgün, E. Kasneci, and N. Pfeifer, "Privacy preserving gaze estimation using synthetic images via a randomized encoding based framework," in *Symposium on Eye Tracking Research and Applications*, 2020, pp. 1–5.

[68] E. Kasneci, "Towards the automated recognition of assistance need for drivers with impaired visual field," *Wilhelmstr*, vol. 32, p. 72 074, 2013.

[69] D. A. Cronin, E. H. Hall, J. E. Goold, T. R. Hayes, and J. M. Henderson, "Eye movements in real-world scene photographs: General characteristics and effects of viewing task," *Frontiers in Psychology*, vol. 10, p. 2915, 2020.

[70] J. Henderson, J. Brockmole, M. Castelhano, and M. Mack, "Eye movements: A window on mind and brain," *Visual saliency does not account for eye movements during visual search in real-world scenes. Elsevier, Oxford (in prep)*, 2007.

[71] A. A. Zekveld, T. Koelewijn, and S. E. Kramer, "The pupil dilation response to auditory stimuli: Current state of knowledge," *Trends in hearing*, vol. 22, p. 2 331 216 518 777 174, 2018.

[72] K. Rayner, "Eye movements in reading and information processing: 20 years of research.," *Psychological Bulletin*, vol. 124, no. 3, p. 372, 1998.

[73] M. De Rivecourt, M. Kuperus, W. J. Post, and L. J. Mulder, "Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight," *Ergonomics*, vol. 51, no. 9, pp. 1295–1319, 2008.

[74] S. Chen, J. Epps, N. Ruiz, and F. Chen, "Eye activity as a measure of human mental effort in hci," in *Proceedings of the 16th international conference on Intelligent user interfaces*, 2011, pp. 315–318.

[75] S. Inamdar and M. Pomplun, "Comparative search reveals the tradeoff between eye movements and working memory use in visual tasks," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 25, 2003.

[76] J. He and J. S. McCarley, "Executive working memory load does not compromise perceptual processing during visual search: Evidence from additive factors analysis," *Attention, Perception, & Psychophysics*, vol. 72, no. 2, pp. 308–316, 2010.

[77] Q. Wang, S. Yang, M. Liu, Z. Cao, and Q. Ma, "An eye-tracking study of website complexity from cognitive load perspective," *Decision Support Systems*, vol. 62, pp. 1 –10, 2014, ISSN: 0167-9236. DOI: https://doi.org/10.1016/j.dss.2014.02.007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167923614000402.

[78] K. F. Van Orden, W. Limbert, S. Makeig, and T.-P. Jung, "Eye activity correlates of workload during a visuospatial memory task," *Human factors*, vol. 43, no. 1, pp. 111–121, 2001.

[79] T. Van Gog, F. Paas, and J. J. Van Merriënboer, "Uncovering expertise-related differences in troubleshooting performance: Combining eye movement and concurrent verbal protocol data," *Applied Cognitive Psychology*, vol. 19, no. 2, pp. 205–221, 2005.

[80] F. Amadieu, T. Van Gog, F. Paas, A. Tricot, and C. Mariné, "Effects of prior knowledge and concept-map structure on disorientation, cognitive load, and learning," *Learning and Instruction*, vol. 19, no. 5, pp. 376–386, 2009.

[81] S. Benedetto, M. Pedrotti, and B. Bridgeman, "Microsaccades and exploratory saccades in a naturalistic environment," *Journal of Eye Movement Research*, vol. 4, no. 2, pp. 1–10, 2011.

[82] X. Gao, H. Yan, and H.-j. Sun, "Modulation of microsaccade rate by task difficulty revealed through between-and within-trial comparisons," *Journal of Vision*, vol. 15, no. 3, pp. 3–3, 2015.

[83] E. Siegenthaler, F. M. Costela, M. B. McCamy, L. L. Di Stasi, J. Otero-Millan, A. Sonderegger, R. Groner, S. Macknik, and S. Martinez-Conde, "Task difficulty in mental arithmetic affects microsaccadic rates and magnitudes," *European Journal of Neuroscience*, vol. 39, no. 2, pp. 287–294, 2014.

[84] K. Krejtz, A. T. Duchowski, A. Niedzielska, C. Biele, and I. Krejtz, "Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze," *PLOS ONE*, vol. 13, no. 9, e0203629, 2018.

[85] B. Cassin, M. L. Rubin, and S. Solomon, *Dictionary of eye terminology*. Triad Publishing Company Gainsville, 1984, vol. 10.

[86] J. Beatty, B. Lucero-Wagoner, *et al.*, "The pupillary system," *Handbook of psychophysiology*, vol. 2, no. 142-162, 2000.

[87] A. F. Kramer, "Physiological metrics of mental workload: A review of recent progress," *Multiple-task Performance*, pp. 279–328, 1991.

[88] D. Kahneman, *Attention and effort*. Citeseer, 1973, vol. 1063.

[89] J. McLaren, J. Erie, and R. Brubaker, "Computerized analysis of pupillograms in studies of alertness," *Investigative ophthalmology & visual science*, vol. 33, pp. 671–6, Apr. 1992.

[90] E. H. Hess and J. M. Polt, "Pupil size in relation to mental activity during simple problem-solving," *Science*, vol. 143, no. 3611, pp. 1190–1192, 1964.

[91] D. Kahneman and J. Beatty, "Pupil diameter and load on memory," *Science*, vol. 154, no. 3756, pp. 1583–1585, 1966.

[92] O. Palinko, A. L. Kun, A. Shyrokov, and P. Heeman, "Estimating cognitive load using remote eye tracking in a driving simulator," in *Proceedings of the 2010 Symposium on Eye-Tracking Research and Applications*, ser. ETRA '10, Austin, Texas: Association for Computing Machinery, 2010, 141–144, ISBN: 9781605589947. DOI: 10.1145/1743666.1743701. [Online]. Available: https://doi.org/10.1145/1743666.1743701.

[93] A. L. Kun, O. Palinko, Z. Medenica, and P. A. Heeman, "On the feasibility of using pupil diameter to estimate cognitive load changes for in-vehicle spoken dialogues," in *INTERSPEECH*, 2013.

[94] O. Palinko and A. L. Kun, "Exploring the effects of visual cognitive load and illumination on pupil diameter in driving simulators," in *Proceedings of the S2012 ymposium on Eye Tracking Research and Applications*, 2012, pp. 413–416.

[95] T. M. Gable, A. L. Kun, B. N. Walker, and R. J. Winton, "Comparing heart rate and pupil size as objective measures of workload in the driving context: Initial look," in *Adjunct Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, ser. AutomotiveUI '15, Nottingham, United Kingdom: Association for Computing Machinery, 2015, 20–25, ISBN: 9781450338585. DOI: 10.1145/2809730.2809745. [Online]. Available: https://doi.org/10.1145/2809730.2809745.

[96] V. Faure, R. Lobjois, and N. Benguigui, "The effects of driving environment complexity and dual tasking on drivers' mental workload and eye blink behavior," *Transportation research part F: traffic psychology and behaviour*, vol. 40, pp. 78–90, 2016.

[97] S. Mathôt, J. Fabius, E. Van Heusden, and S. Van der Stigchel, "Safe and sensible preprocessing and baseline correction of pupil-size data," *Behavior research methods*, vol. 50, no. 1, pp. 94–106, 2018.

[98] M. E. Kret and E. E. Sjak-Shie, "Preprocessing pupil size data: Guidelines and code," *Behavior research methods*, vol. 51, no. 3, pp. 1336–1342, 2019.

[99] S. P. Marshall, "Method and apparatus for eye tracking and monitoring pipil dilation to evaluate cognitive activity," Patent US 006090051A, Jul. 2000. [Online]. Available: `https://patentimages.storage.googleapis.com/pdfs/9171d27ab488a900c7db/US6090051.pdf`.

[100] A. Duchowski, K. Krejtz, I. Krejtz, C. Biele, A. Niedzielska, P. Kiefer, R. Martin, and I. Giannopoulos, "The index of pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation," Apr. 2018, pp. 1–13. DOI: `10.1145/3173574.3173856`.

[101] S. H. Fairclough, L. J. Moores, K. C. Ewing, and J. Roberts, "Measuring task engagement as an input to physiological computing," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, Sep. 2009, pp. 1–9. DOI: `10.1109/ACII.2009.5349483`.

[102] S. P. Marshall, "Identifying cognitive state from eye metrics," *Aviation, space, and environmental medicine*, vol. 78, no. 5, B165–B175, 2007.

[103] K. Fukuda, J. A. Stern, T. B. Brown, and M. B. Russo, "Cognition, blinks, eye-movements, and pupillary movements during performance of a running memory task," *Aviation, space, and environmental medicine*, vol. 76, no. 7, pp. C75–C85, 2005.

[104] D. Bristow, C. Frith, and G. Rees, "Two distinct neural effects of blinking on human visual processing," *Neuroimage*, vol. 27, no. 1, pp. 136–145, 2005.

[105] S. Chen and J. Epps, "Using task-induced pupil diameter and blink rate to infer cognitive load," *Human–Computer Interaction*, vol. 29, no. 4, pp. 390–413, 2014. DOI: 10.1080/07370024.2014.892428. eprint: https://doi.org/10.1080/07370024.2014.892428. [Online]. Available: https://doi.org/10.1080/07370024.2014.892428.

[106] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and systems magazine*, vol. 6, no. 3, pp. 21–45, 2006.

[107] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006.

[108] A. Narassiguin, M. Bibimoune, H. Elghazel, and A. Aussem, "An extensive empirical comparison of ensemble learning methods for binary classification," *Pattern Analysis and Applications*, vol. 19, no. 4, pp. 1093–1128, 2016.

[109] M. Bibimoune, H. Elghazel, and A. Aussem, "An empirical comparison of supervised ensemble learning approaches," in *International Workshop on Complex Machine Learning Problems with Ensemble Methods COPEM@ ECML/PKDD*, vol. 13, 2013, pp. 123–138.

[110] Y. Saez, A. Baldominos, and P. Isasi, "A comparison study of classifier algorithms for cross-person physical activity recognition," *Sensors*, vol. 17, no. 1, p. 66, 2017.

[111] L. Fridman, B. Reimer, B. Mehler, and W. T. Freeman, "Cognitive load estimation in the wild," in *Proceedings of the 2018 chi conference on human factors in computing systems*, 2018, pp. 1–9.

[112] B. Reimer, B. Mehler, Y. Wang, and J. F. Coughlin, "A field study on the impact of variations in short-term memory demands on drivers' visual attention and driving performance across three age groups," *Human factors*, vol. 54, no. 3, pp. 454–468, 2012.

[113] C. L. Baldwin and B. Penaranda, "Adaptive training using an artificial neural network and eeg metrics for within- and cross-task workload classification," *NeuroImage*, vol. 59, no. 1, pp. 48 –56, 2012, Neuroergonomics: The human brain in action and at work, ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2011.07.047. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S105381191100824X.

[114] C. Walter, S. Schmidt, W. Rosenstiel, P. Gerjets, and M. Bogdan, "Using cross-task classification for classifying workload levels in complex learning tasks," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, Sep. 2013, pp. 876–881. DOI: 10.1109/ACII.2013.164.

[115] M. Spüler, T. Krumpe, C. Walter, C. Scharinger, W. Rosenstiel, and P. Gerjets, "Brain-computer interfaces for educational applications," in *Informational Environments*, Springer, 2017, pp. 177–201.

[116] E. Haapalainen, S. Kim, J. F. Forlizzi, and A. K. Dey, "Psycho-physiological measures for assessing cognitive load," in *Proceedings of the 12th ACM international conference on Ubiquitous computing*, 2010, pp. 301–310.

[117] D. E. Irwin and L. E. Thomas, "Eyeblinks and cognition.," 2010.

[118] K. F. Van Orden, W. Limbert, S. Makeig, and T.-P. Jung, "Eye activity correlates of workload during a visuospatial memory task," *Human factors*, vol. 43, no. 1, pp. 111–121, 2001.

[119] J. A. Stern, L. C. Walrath, and R. Goldstein, "The endogenous eyeblink," *Psychophysiology*, vol. 21, no. 1, pp. 22–33, 1984.

[120] S. Chen, J. Epps, and F. Chen, "A comparison of four methods for cognitive load measurement," in *Proceedings of the 23rd Australian Computer-Human Interaction Conference*, 2011, pp. 76–79.

[121] M. Shojaeizadeh, S. Djamasbi, R. C. Paffenroth, and A. C. Trapp, "Detecting task demand via an eye tracking machine learning system," *Decision Support Systems*, vol. 116, pp. 91–101, 2019.

[122] S. Chen, J. Epps, and F. Chen, "Automatic and continuous user task analysis via eye activity," in *Proceedings of the 2013 international conference on Intelligent user interfaces*, 2013, pp. 57–66.

[123] R. Engbert and R. Kliegl, "Microsaccades uncover the orientation of covert attention," *Vision Research*, vol. 43, no. 9, pp. 1035 –1045, 2003, ISSN: 0042-6989. DOI: https://doi.org/10.1016/S0042-6989(03)00084-1. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0042698903000841.

[124] M. P. Coral, "Analyzing cognitive workload through eye-related measurements: A meta-analysis," 2016.

[125] A. Duchowski, K. Krejtz, J. Zurawska, and D. House, "Using microsaccades to estimate task difficulty during visual search of layered surfaces," *IEEE transactions on visualization and computer graphics*, 2019.

[126] E. Krueger, A. Schneider, A. Chavaillaz, R. Groner, B. D. Sawyer, A. Sonderegger, and P. Hancock, "Microsaccades distinguish looking from seeing," *Journal of Eye Movement Research*, vol. 12, no. 6, p. 2, 2019.

[127] V. Demberg, "Pupillometry: The index of cognitive activity in a dual-task study," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 35, 2013.

[128] A. Korbach, R. Brünken, and B. Park, "Measurement of cognitive load in multimedia learning: A comparison of different objective measures," *Instructional science*, vol. 45, no. 4, pp. 515–536, 2017.

[129] L. Rerhaye, T. Blaser, and T. Alexander, "Evaluation of the index of cognitive activity (ica) as an instrument to measure cognitive workload under differing light conditions," in *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)*, S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, and Y. Fujita, Eds., Cham: Springer International Publishing, 2019, pp. 350–359, ISBN: 978-3-319-96059-3.

[130] J. Beatty, "Task-evoked pupillary responses, processing load, and the structure of processing resources.," *Psychological bulletin*, vol. 91, no. 2, p. 276, 1982.

[131] H. H. Telek, H. Erdol, and A. Turk, "The effects of age on pupil diameter at different light amplitudes," *Beyoglu Eye J*, 2018.

[132] Y. Zhang, S. Li, J. Wang, P. Wang, Y. Tu, X. Li, and B. Wang, "Pupil size estimation based on spatially weighted corneal flux density," *IEEE Photonics Journal*, vol. 11, no. 6, pp. 1–9, 2019.

[133] B. Pfleging, D. K. Fekety, A. Schmidt, and A. L. Kun, "A model relating pupil diameter to mental workload and lighting conditions," in *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2016, pp. 5776–5788.

[134] R. R. Henderson, M. M. Bradley, and P. J. Lang, "Emotional imagery and pupil diameter," *Psychophysiology*, vol. 55, no. 6, e13050, 2018.

[135] M. M. Bradley, L. Miccoli, M. A. Escrig, and P. J. Lang, "The pupil as a measure of emotional arousal and autonomic activation," *Psychophysiology*, vol. 45, no. 4, pp. 602–607, 2008.

[136] R. F. Stanners, M. Coulter, A. W. Sweet, and P. Murphy, "The pupillary response as an indicator of arousal and cognition," *Motivation and Emotion*, vol. 3, no. 4, pp. 319–340, 1979.

[137] S. R. Steinhauer, G. J. Siegle, R. Condray, and M. Pless, "Sympathetic and parasympathetic innervation of pupillary dilation during sustained processing," *International journal of psychophysiology*, vol. 52, no. 1, pp. 77–86, 2004.

[138] M. Melby-Lervåg and C. Hulme, "Is working memory training effective? a meta-analytic review.," *Developmental psychology*, vol. 49, no. 2, p. 270, 2013.

[139] T. Zu, J. Hutson, L. C. Loschky, and N. S. Rebello, "Use of eye-tracking technology to investigate cognitive load theory," *arXiv preprint arXiv:1803.02499*, 2018.

[140] E. L. Johnson, A. T. Miller Singley, A. D. Peckham, S. L. Johnson, and S. A. Bunge, "Task-evoked pupillometry provides a window into the development of short-term memory capacity," *Frontiers in psychology*, vol. 5, p. 218, 2014.

[141] Z. M. Hafed, C.-Y. Chen, and X. Tian, "Vision, perception, and attention through the lens of microsaccades: Mechanisms and implications," *Frontiers in systems neuroscience*, vol. 9, p. 167, 2015.

[142] A. B. Watson and J. I. Yellott, "A unified formula for light-adapted pupil size," *Journal of vision*, vol. 12, no. 10, pp. 12–12, 2012.

[143] J. F. Hopstaken, D. van der Linden, A. B. Bakker, and M. A. Kompier, "The window of my eyes: Task disengagement and mental fatigue covary with pupil dynamics," *Biological psychology*, vol. 110, pp. 100–106, 2015.

[144] R. Schleicher, N. Galley, S. Briest, and L. Galley, "Blinks and saccades as indicators of fatigue in sleepiness warnings: Looking tired?" *Ergonomics*, vol. 51, no. 7, pp. 982–1010, 2008, PMID: 18568959. DOI: 10.1080/00140130701817062. eprint: https://doi.org/10.1080/00140130701817062. [Online]. Available: https://doi.org/10.1080/00140130701817062.

[145] L. L. Di Stasi, M. B. McCamy, A. Catena, S. L. Macknik, J. J. Cañas, and S. Martinez-Conde, "Microsaccade and drift dynamics reflect mental fatigue," *European Journal of Neuroscience*, vol. 38, no. 3, pp. 2389–2398, 2013. DOI: 10.1111/ejn.12248. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ejn.12248. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/ejn.12248.

[146] K. Mizuno, M. Tanaka, K. Yamaguti, O. Kajimoto, H. Kuratsune, and Y. Watanabe, "Mental fatigue caused by prolonged cognitive load associated with sympathetic hyperactivity," *Behavioral and brain functions*, vol. 7, no. 1, pp. 1–7, 2011.

Appendix

Manuscript 0

Title:
Brightness-and motion-based blink detection for head-mounted eye trackers

Authors:
Tobias Appel, Thiago Santini, and Enkelejda Kasneci

Published in:
Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct.

Time of publication:
September 2016

Included in this dissertation with permission from the PETMEI.

Brightness- and Motion-Based Blink Detection for Head-Mounted Eye Trackers

Tobias Appel
Perception Engineering
University of Tübingen, Germany
tobias.appel@student.uni-tuebingen.de

Thiago Santini
Perception Engineering
University of Tübingen, Germany
thiago.santini@uni-tuebingen.de

Enkelejda Kasneci
Perception Engineering
University of Tübingen, Germany
enkelejda.kasneci@uni-tuebingen.de

Ubicomp/ISWC'16 Adjunct, September 12-16, 2016, Heidelberg, Germany.
ACM 978-1-4503-4462-3/16/09
http://dx.doi.org/10.1145/2968219.2968341

Abstract
Blinks are an indicator for fatigue or drowsiness and can assist in the diagnose of mental disorders, such as schizophrenia. Additionally, a blink that obstructs the pupil impairs the performance of other eye-tracking algorithms, such as pupil detection, and often results in noise to the gaze estimation. In this paper, we present a blink detection algorithm that is tailored towards head-mounted eye trackers and is robust to calibration-based variations like translation or rotation of the eye. The proposed approach reached 96,35% accuracy for a realistic and challenging data set and in real-time even on low-end devices, rendering the proposed method suited for pervasive eye tracking.

Author Keywords
Blink detection; Pervasive eye tracking, Real time; Image processing

ACM Classification Keywords
I.5.4 [PATTERN RECOGNITION]: Applications; I.4.8 [IMAGE PROCESSING AND COMPUTER VISION]: Object recognition,Shape; I.4.9 [IMAGE PROCESSING AND COMPUTER VISION]: Applications

Introduction
A blink is a rapid closing and opening of the eyelids that falls within three classes: endogenous, reflex, and volun-

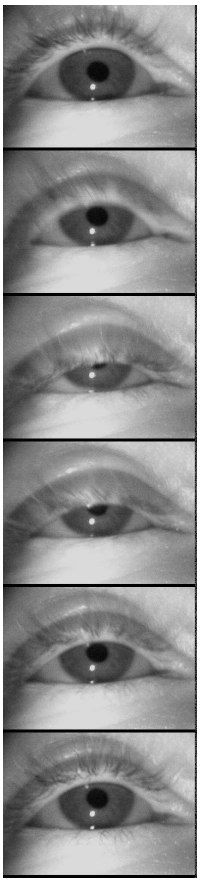


Figure 1: Eyelid movement during an endogenous blink, which typically lasts for 75 to 400 ms [4]. For a low-end eye tracker (≈ 25 frames per second), this results in approximatelly 4 to 16 frames.

tary movements [21]. Endogenous (or spontaneous) blinks serve to spread the tear film over the cornea and remove irritants.

Reflex blinks originate from the startle reflex to protect the eye from external stimuli. For these two classes, blinks usually range from 75 to 400 ms [4]. In contrast, voluntary blinks are performed consciously, can be used for interaction – e.g., in human-computer interfaces (HCI) [10] – and have no determined duration patterns. Apart from these biological functions, unusual blink patterns are also indicative of a person's state of vigilance, fatigue, and drowsiness [15, 25, 22]. Such states are specially important in situations that require quick reactions, e.g., during driving [3]; in this context, real-time blink detection combined with pervasive eye tracking has the potential to prevent dangerous and life-threatening circumstances. Furthermore, blinks are a significant source of noise for eye-tracking algorithms. For instance, there is a trade-off between detecting pupils in realistic and challenging scenarios and false positives during blink (when no pupil is visible). Moreover, mid-blink the pupil becomes partially occluded causing pupil detection algorithms to bias the pupil center towards the still visible part; as a result, blinks must be taken into account during the automatic classification of eye movements [19]. Thus, a robust and accurate blink detection algorithm enables not only the employment of blink-related data (e.g., frequency) but also circumvents the noise introduced by blinks in other eye-tracking algorithms.

Due to the many advantages that head-mounted eye trackers offer – e.g., mobility and unintrusivnes – they are promising candidates for pervasive eye tracking, and, thus, this work focuses on these eye trackers. An algorithm for use in head-mounted eye trackers has different requirements than one for remote eye tracking. There is no need for head or eye localization, but the exact location, alignment, and angle of the eye in the video depends on the eye camera position, which varies significantly from subject to subject. This makes it uncertain, where to expect the pupil or eyelids and mostly prohibits the use of any priors in an algorithm. Motion bluring and frame skips also pose problems. The former renders blink detection based on edges almost impossible, and the latter can significantly distort a blink sequence. Further challenges are added by reflections, uncommon angles, and illumination changes.

In this work, we propose a brightness and motion based algorithm that runs in real-time in systems equivalent to those used in state-of-the-art eye tracking systems and does not rely on prior information. The fact that a sequence is analyzed in contrast to a frame-by-frame approach makes our algorithm robust to frame skips and the brightness-based detection does not rely on edges of any kind. Furthermore, we introduce a new labeled data set of realistic and challenging images from on-road driving experiments. To foster further research in the field and effortless replication of our result, we contribute the algorithm implementation and data sets openly at:
www.ti.uni-tuebingen.de/perception

Related work

The great majority of video-based blink identification concerns remote eye trackers or regular cameras. In a first stage, a plethora of methods, such as Viola-Jones [24] and KLT trackers [23], are used to identify and track the subject's face and eyes region. As previously mentioned, in this work we focus on head-mounted eye trackers. On one hand, these devices impose extra constraints on the blink identification task. For instance, since near-infrared images are employed, no color information is available. Furthermore, the orientation of the eye image and eye corner

positions are not known a priori. On the other hand, head-mounted devices do not require the aforementioned stage; thus, henceforth we discuss related work assuming this initial stage is performed appropriately and eye boxes have been identified correctly.

Smith et al. [20] first identify the *eye-white color* as the brightest pixel in the eye region on an initial frame; further frames are classified on whether *eye-white pixels* exist in the eye region (non-blinks) or not (blinks). Grauman et al. [7] employ an open eye template and correlation scores to determine whether a frame contains an open or closed eye (based on a predetermined threshold). Ito et al. [9] divide the input image into vertical sections and, for each section, find the pair of maximal and minimal intensity derivatives most distant from the darkest point in the section. The candidates from five sections are grouped, and two groups are estimated to represent the upper and lower eyelid; the average distance between the upper and lower eyelid points is used to measure the degree of closure. A threshold then discriminates between blinks and open eyes. Moriyama et al. [16] rely on the average illumination intensity for the upper and lower halves of the eye region. Crossings between these values are employed to determine when the eyelid crosses the line separating these regions and, thus, blinks. Morris et al. [17] use a mean image and variance map to detect blinks; these are updated with each new frame, and the resulting variance maps is thresholded. If the number of pixels remaining is larger in relation to the eye box, a blink is assumed. Lalonde et al. [13] employ scale-invariant feature transform to identify tracking points. The optical-flow of the points inside the eye region is then employed to identify when the eyelid descends and ascends. Bavivarov et al. [1] models the eyes through an active appearance model and define a blink criteria based on the ratio between the resulting eye width and aperture. Lee et al. [14] first normalize the eye region illumination to account for illumination variations. The cumulative difference of black pixels in a binarized version of the normalized eye region and the ratio between the eye height/width are used as features for a support vector machine classifier that discriminates between open and closed eyes. Drutarovsky and Fogelton [5] employ a flock of KLT trackers within the eye region, which are used to evaluate the average motion of nine equally sized cells in the region. The variance of the superior six cells drive a finite state machine that identifies downward and upward eyelid movements. Jiang et al. [11] consider images provided by head-mounted eye trackers. Their approach consists with thresholding the difference between two subsequent frames. The resulting image is then morphologically opened, and pupil and eyelids identified. Blink onset is determined based on eyelid position changes between the two consecutive frames, whereas blink offset is identified based on pupil size changes during an ongoing blink.

Method

The nature of our data leads to an approach that does not rely on edge detection or prior knowledge about the location of the pupil or shape of the eye. It is based on two simple assumptions: the pupil is dark and gets at least partly obstructed by the eyelid during a blink. Those two facts can be exploited if we look at the brightness of consecutive frames. During the blink onset, the frame brightness steadily increases, reaching its maximum at the blink apex. Afterwards, frame brightness decreases during the blink offset until it approaches a level similar to the one prior the blink (see Figure 2).

Percentile values serve as a measure of brightness in our algorithm. In addition, differences between consecutive frames are used to determine if there is enough change to

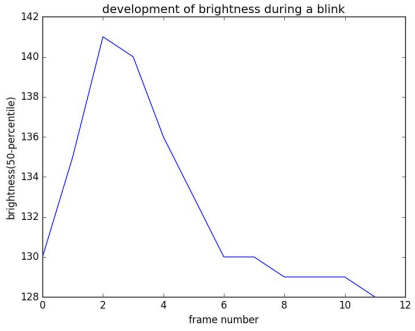


Figure 2: The brightness development for a typical blink.

classify a sequence as a blink or not. For this, the frames f_i and f_{i-1} were blurred to reduce noise, and the absolute difference between them are calculated and summed up. Both together form the features of a single frame. In the following formula, c and r denote column number and row number respectively, and $b(f_i)$ is the blurred version of frame f_i.

$$\text{diff}_{i,i-1} = \sum_{r,c} |b(f_i)(r,c) - b(f_{i-1})(r,c)| \quad (1)$$

$$\text{feature}_i = (P_{percentile}(f_i), \text{diff}_{i,i-1}) \quad (2)$$

These features are calculated for k consecutive frames, which together make up the feature vector for a window of size k. Figure 3 illustrates the feature extraction procedure.

$$\text{feature}_{i-k,i} = (\text{feature}_{i-k}, \ldots, \text{feature}_i) \quad (3)$$

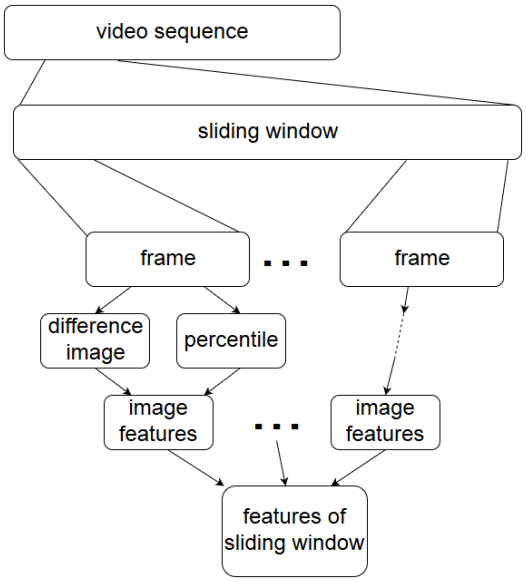


Figure 3: Feature extraction process

Even though blinks may vary in length, their basic structure remains similar. For a small amount of frames the eyelid descends, whereas ascension takes a larger amount of frames since eyelid velocity is higher during blink onset. Thus, it is reasonable to assume a fixed window size that appropriately models this behavior exists. Different choices for the chosen percentile and the window size k are discussed in the evaluation section.

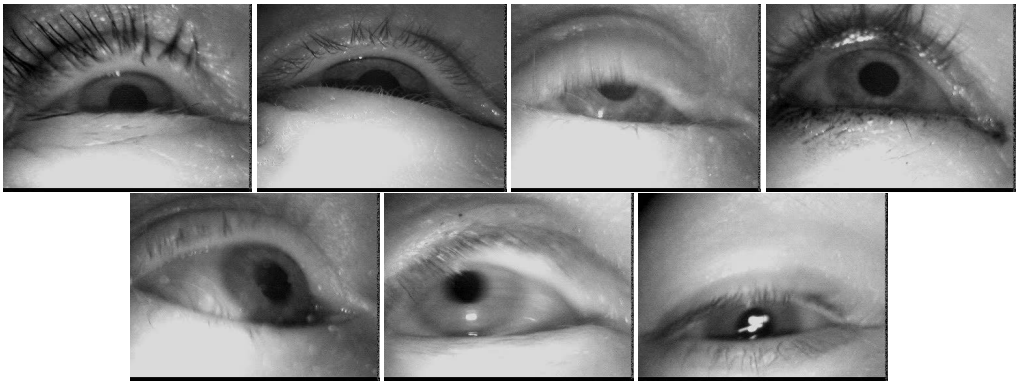


Figure 4: Challenging examples due to bad angle, incomplete blink, make-up, motion blurring and reflections

Duration	Blinks
<5	5
5	69
6	316
7	699
8	683
9	349
10	156
11	65
12	33
13	19
>13	16

Table 1: Duration distribution in terms of frames for all recorded blinks. Each frame encompasses $\approx$ 40 ms.

Based on these features, a Random Forest Classifier [8] with 100 trees is trained. If more computational power is available, the amount of trees can be increased to increase accuracy. In contrast, the number of trees can be scaled down to allow a faster evaluation. Random Forests are the method of choice, because they are quick to train and very fast to evaluate in addition to being able to handle non-linearity. The possibility to parallelize both processes increases the speed further. In addition to the already mentioned advantages, Random Forests do not need scaling for their input and, thus, effectively handle the combined feature vector of summed differences and brightness changes. Furthermore, these classifiers are resilient to outliers, so eccentric blinks do not influence its ability to generalize.

The algorithm was implemented in *Python* using the *OpenCV* [2] and *Scikit-learn* [18] libraries. Testing was done with an Intel® Core™ i7-4790 at 3.60GHz and with 12GB of RAM. This system is consistent with those used by state-of-the-art eye trackers (e.g., Dikablis [6]). The mean runtime of the feature extraction process was 0.6630ms and predicting the class of all 4820 training samples took 0.0264ms on average. This amounts to a processing rate of 1450.54 frames per second once the images are loaded.

Evaluation

The proposed algorithm was evaluated using a data set of 20 video sequences of 5 minutes each extracted from a on-road driving experiment [12]; thus, the data contains endogenous and, possibly, reflex blinks. Each video corresponds to a different subject. The videos were recorded using a Dikablis Essential eye tracker at a sampling rate of 25Hz and a resolution of 384 x 288 pixels. All blink sequences were annotated, amounting to a total of 2410 blinks; their duration distribution is shown in 1. Every blink sequence starts with a completely open eye, is followed by

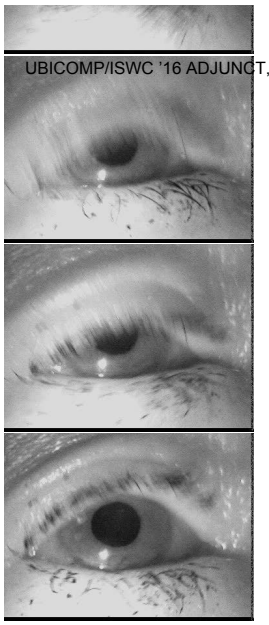


Figure 5: Subject 17 showed only little occlusion of the pupil during most blinks. This leads to a relativly high false negative rate.

the blink apex, and continues until roughly the same degree of openness that the eye had before the blink is reached. It is worth noticing that this data set provides realistic and challenging eye images, including quick changing illumination, blurring, reflections, and makeup. In contrast, related work usually employs data from indoor scenarios, which differ little from laboratory settings and are not realistic in the context of pervasive eye tracking. This data set is available for download at:
www.ti.uni-tuebingen.de/perception

We investigate the algorithm behaviour for window sizes that encompass expected durations of endogenous and reflexive blinks, ranging from $\approx$ 80 ms (i.e., $k = 2$) to $\approx$ 600 ms (i.e., $k = 15$). Blinks durations vary inside this range. However, in order to train a classifier, we need a fixed amount of features; thus, it is necessary to clip or extend blink sequences in order to fit the selected window size. In the case that our blink sequence was too short to fit the window, subsequent frames were added to the blink sequence until it was the same length as the window. Since mostly the start of a blink is needed for classification, extending the sequence at the end only influences the classification negatively if there are blinks in extremely rapid succession – i.e., the sequence covers two or more blinks. If a blink sequence was to long, we trimmed it at the end. The second parameter under investigation is the percentile used as brightness indicator. For this, we investigate a wide range, from 5 up to 90-percentile. The method was evaluated through leave-one-out cross-validation. In other words, results were obtained by training on the other 19 subjects and evaluating on the remaining one; this procedure was performed for each subject. Negative training samples were chosen at random from periods not encompassing blinks, and the amount of negative sequences was chosen as to equal the amount of positive (blink) sequences for each

Subject	**Accuracy (%)**			
	5 Percentile		50 Percentile	
	Blink	Non-Blink	Blink	Non-Blink
1	96.0	97.0	97.0	95.0
2	100.0	53.1	96.9	89.0
3	100.0	98.5	98.5	100.0
4	95.0	95.0	95.6	95.6
5	97.0	99.0	96.0	99.0
6	96.6	97.8	97.8	97.8
7	100.0	92.3	100.0	89.3
8	97.8	98.9	96.6	97.8
9	92.0	96.6	92.5	94.3
10	100.0	100.0	100.0	100.0
11	92.1	99.5	91.6	98.0
12	96.7	97.5	96.7	95.0
13	91.1	97.8	91.7	97.8
14	100.0	91.3	100.0	95.7
15	96.0	98.0	97.0	93.9
16	97.6	99.2	96.8	100.0
17	90.0	100.0	83.8	100.0
18	93.3	98.7	88.7	100.0
19	91.0	99.6	91.9	97.9
20	96.5	100.0	94.8	99.1

Table 2: Individual results for the subjects obtained with a window size of 12

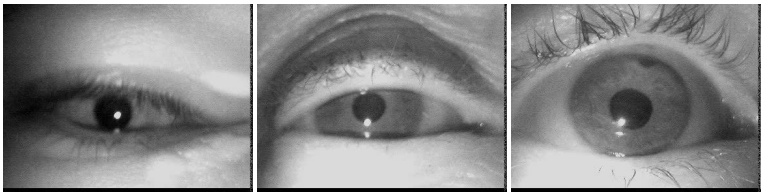


Figure 6: Difficult cases in terms of false positives: subjects 2, 7 and 14.

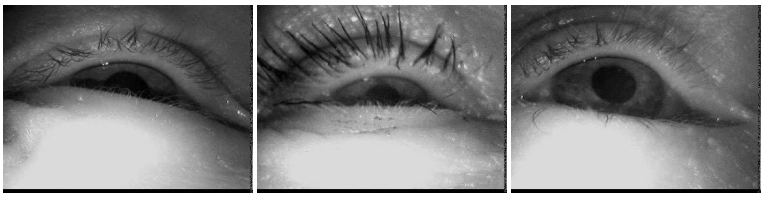


Figure 7: Difficult cases in terms of false negatives: subjects 9, 11 and 13

subject.

Figure 8 reports the average positive and negative predictive value as the window size and brightness percentile change. This figure clearly shows that the detection rate of a blink increases with the window size. This is to be expected because by clipping blink sequence information is lost. Nonetheless, since blinks usually do not last longer than 400ms, there is a point where increasing the window size further does not further improve detection. Moreover, larger windows enables a more clear distinction between blinks and non-blinks, because the features of a non-blink sequence are less likely to match those of blink sequences by chance. However, problems arise when there are several blinks in rapid succession that are to short to fill the window. Oftentimes, this leads to misclassification as non-blink. The brightness percentile affects both accuracies of blinks and non-blinks. The smaller the percentile, the better in terms of blink accuracy. With the choice of the 5-percentile only changes from very dark pixels to brighter ones are measured, and, thus, mainly changes in pupil pixels occur. This increases the accuracy of blink detection. However, if a subject has distinctly dark eyelashes, these have roughly the same pixel intensity as the pupil; as a result, the algorithm responds to every sequence, classifying it as a blink. This is especially true for small narrow eyes. Per subject results are reported in Table 2 for a window size of 12 frames. Averaged over all evaluated percentile, a window size of 12 yielded the best F1-score. As can be seem in this table, only one subject (subject 2) presented such

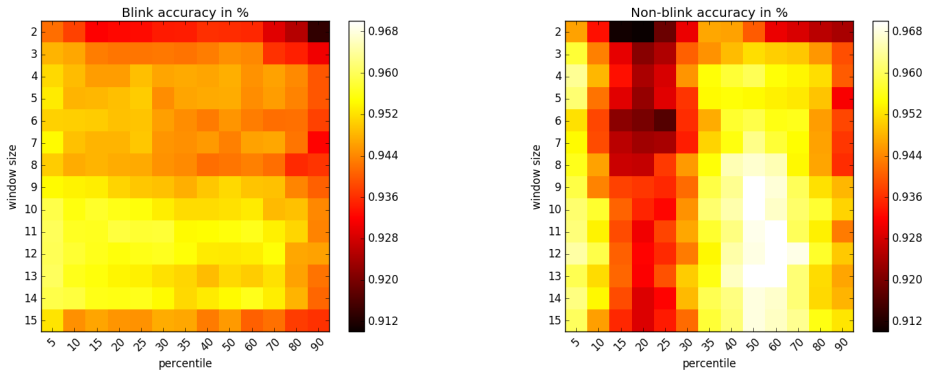


Figure 8: Average predictive value across subjects for blink (left) and non-blink (right) sequences as the window size k and brightness percentile change.

distinctly dark eyelashes. Remaining subjects were classified properly with the 5-percentile or equally well with both percentiles. With 96,3795% the overall best F1-score was achieved using the 50th percentile and a window size of 11.

Figure 6 illustrates the problems with subject 2 as well as subject 7, who too had a small pupil and wore make-up. In addition to that, subject 7 had a partly occluded pupil even when not blinking. Subject 14 suffered from an iris defect that can be misinterpreted as a pupil and in the process of looking downwards it gets obscured, mimicking a blink.

Figure 7 shows three subjects that had a false negative rate above average. This can stem from a bad angle or illumination. Both can lead to a low visibility of the pupil and lessen the brightness change, which ultimately results in a lower detection rate. In the case of subject 17 the lower detection rate arises from the fact that the pupil does not get occluded significantly during a blink (see figure 5). Double blinks only occurred in the video of subject 19. The rapid sequence of blinks lead to some misclassification and a lower detection rate.

Conclusion

We presented an approach that is fast and has very high detection rate for blinks. For further elaboration of the proposed algorithm, we plan to broaden the spectrum of subjects to have a representative data base to train, which will significantly decrease false positives. A greater data base would also allow us to train different classifiers for different window sizes and would enable us to narrow down the time span of a blink. This way the blink duration can be estimated. Additionally a calibration phase can be integrated

at the start of an experiment where the subject is asked to open and close its eyes to have samples to construct blink sequences of different lengths and estimate the expected brightness change during a blink, which allows for normalization. To foster further research in the field and effortless replication of our result, we contribute the algorithm implementation and data sets openly at:
www.ti.uni-tuebingen.de/perception

REFERENCES

1. I. Bacivarov, M. Ionita, and P. Corcoran. 2008. Statistical models of appearance for eye tracking and eye-blink detection and measurement. *Consumer Electronics, IEEE Transactions on* 54, 3 (2008), 1312–1320.
2. G. Bradski and others. 2000. The opencv library. *Doctor Dobbs Journal* 25, 11 (2000), 120–126.
3. C. Braunagel, E. Kasneci, W. Stolzmann, and W. Rosenstiel. 2015. Driver-Activity Recognition in the Context of Conditionally Autonomous Driving. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*. 1652–1657.
4. P. P Caffier, U. Erdmann, and P. Ullsperger. 2003. Experimental evaluation of eye-blink parameters as a drowsiness measure. *European journal of applied physiology* 89, 3-4 (2003), 319–325.
5. T. Drutarovsky and A. Fogelton. 2014. Eye Blink Detection Using Variance of Motion Vectors. In *Computer Vision-ECCV 2014 Workshops*. Springer, 436–448.
6. Ergoneers. 2016. Dikablis Glasses. `http://www.ergoneers.com/`. (2016).
7. K. Grauman, M. Betke, J. Gips, and G. R Bradski. 2001. Communication via eye blinks-detection and duration analysis in real time. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, Vol. 1. IEEE, I–1010.
8. T. K. Ho. 1995. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, Vol. 1. IEEE, 278–282.
9. T. Ito, S. Mita, K. Kozuka, T. Nakano, and S. Yamamoto. 2002. Driver blink measurement by the motion picture processing and its application to drowsiness detection. In *Intelligent Transportation Systems, 2002. Proceedings. The IEEE 5th International Conference on*. IEEE, 168–173.
10. R. J. Jacob and K. S. Karn. 2003. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. *Mind* 2, 3 (2003), 4.
11. X. Jiang, G. Tien, D. Huang, B. Zheng, and M. S. Atkins. 2013. Capturing and evaluating blinks from video-based eyetrackers. *Behavior research methods* 45, 3 (2013), 656–663.
12. E. Kasneci, K. Sippel, K. Aehling, M. Heister, W. Rosenstiel, U. Schiefer, and E. Papageorgiou. 2014. Driving with Binocular Visual Field Loss? A Study on a Supervised On-road Parcours with Simultaneous Eye and Head Tracking. *Plos One* 9, 2 (2014), e87470.
13. M. Lalonde, D. Byrns, L. Gagnon, N. Teasdale, and D. Laurendeau. 2007. Real-time eye blink detection with GPU-based SIFT tracking. In *Computer and Robot Vision, 2007. CRV'07. Fourth Canadian Conference on*. IEEE, 481–487.

14. W. O. Lee, E. C. Lee, and K. R. Park. 2010. Blink detection robust to various facial poses. *Journal of neuroscience methods* 193, 2 (2010), 356–372.

15. L. K. McIntire, R. A. McKinley, C. Goodyear, and J. P. McIntire. 2014. Detection of vigilance performance using eye blinks. *Applied ergonomics* 45, 2 (2014), 354–362.

16. T. Moriyama, T. Kanade, J. F Cohn, J. Xiao, Z. Ambadar, J. Gao, and H. Imamura. 2002. Automatic recognition of eye blinking in spontaneously occurring behavior. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, Vol. 4. IEEE, 78–81.

17. T. Morris, P. Blenkhorn, and F. Zaidi. 2002. Blink detection for real-time eye tracking. *Journal of Network and Computer Applications* 25, 2 (2002), 129–143.

18. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

19. T. Santini, W. Fuhl, T. Kübler, and E. Kasneci. 2016. Bayesian Identification of Fixations, Saccades, and Smooth Pursuits. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications (ETRA '16)*. ACM, New York, NY, USA, 163–170. DOI: http://dx.doi.org/10.1145/2857491.2857512

20. P. Smith, M. Shah, and N. da Vitoria Lobo. 2000. Monitoring head/eye motion for driver alertness with one camera. In *icpr*. IEEE, 4636.

21. J. A. Stern, L. C. Walrath, and R. Goldstein. 1984. The endogenous eyeblink. *Psychophysiology* 21, 1 (1984), 22–33.

22. M. Suzuki, N. Yamamoto, O. Yamamoto, T. Nakano, and S. Yamamoto. 2006. Measurement of driver's consciousness by image processing-a method for presuming driver's drowsiness by eye-blinks coping with individual differences. In *Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference on*, Vol. 4. IEEE, 2891–2896.

23. C. Tomasi and T. Kanade. 1991. *Detection and tracking of point features*. School of Computer Science, Carnegie Mellon Univ. Pittsburgh.

24. P. Viola and M. J. Jones. 2004. Robust real-time face detection. *International journal of computer vision* 57, 2 (2004), 137–154.

25. F. Yang, X. Yu, J. Huang, P. Yang, and D. Metaxas. 2012. Robust eyelid tracking for fatigue detection. In *2012 19th IEEE International Conference on Image Processing*. 1829–1832.

Manuscript 1

Title:
Cross-subject workload classification using pupil-related measures

Authors:
Tobias Appel, Christian Scharinger, Peter Gerjets, and Enkelejda Kasneci

Published in:
Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications.

Time of publication:
June 2018

Included in this dissertation with permission from the ETRA.

Cross-subject workload classification using pupil-related measures

Tobias Appel
LEAD Graduate School and Research Network
Tübingen
tobias.appel@uni-tuebingen.de

Christian Scharinger
Knowledge Media Research Center
Tübingen

Peter Gerjets
Knowledge Media Research Center
Tübingen

Enkelejda Kasneci
Perception Engeneering, University of Tübingen
Tübingen

ABSTRACT

Real-time evaluation of a person's cognitive load can be desirable in many situations. It can be employed to automatically assess or adjust the difficulty of a task, as a safety measure, or in psychological research. Eye-related measures, such as the pupil diameter or blink rate, provide a non-intrusive way to assess the cognitive load of a subject and have therefore been used in a variety of applications. Usually, workload classifiers trained on these measures are highly subject-dependent and transfer poorly to other subjects. We present a novel method to generalize from a set of trained classifiers to new and unknown subjects. We use normalized features and a similarity function to match a new subject with similar subjects, for which classifiers have been previously trained. These classifiers are then used in a weighted voting system to detect workload for an unknown subject. For real-time workload classification, our methods performs at 70.4% accuracy. Higher accuracy of 76.8% can be achieved in an offline classification setting.

CCS CONCEPTS

• **Applied computing** → **Psychology**; • **Human-centered computing** → *Human computer interaction (HCI)*;

KEYWORDS

eye tracking, pupillometry, blinks, workload, cross-subject, classification

ACM Reference Format:
Tobias Appel, Christian Scharinger, Peter Gerjets, and Enkelejda Kasneci. 2018. Cross-subject workload classification using pupil-related measures. In *ETRA '18: 2018 Symposium on Eye Tracking Research and Applications, June 14–17, 2018, Warsaw, Poland*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3204493.3204531

ETRA '18, June 14–17, 2018, Warsaw, Poland

ACM ISBN 978-1-4503-5706-7/18/06…$15.00
https://doi.org/10.1145/3204493.3204531

1 INTRODUCTION

Workload or *cognitive load* refers to the load on ones working memory in response to a particular task. It can either be measured by subjective ratings, performance data, or physiological response. Already in 1982, Beatty showed that the pupil qualifies as a indicator for cognitive load as its diameter increases linearly with task difficulty [Beatty 1982]. Building on this work, scientists from many different fields investigated the relationship between pupillary changes and workload, e.g., in the fields of aviation [Peysakhovich et al. 2015], driving [Kun et al. 2013; Marquart et al. 2015], medicine [Szulewski et al. 2015], or psychology [Klingner et al. 2011; Laeng et al. 2011].

As a physiological measure, the pupil is objective compared to subjective ratings by the user. Additionally, eye tracking is easily applicable in a lot of different situations and less cumbersome and intrusive compared to measures like EEG, but more reliable than skin conductance and heart rate.

To detect cognitive load, usually subject-related data is analyzed in hindsight. However, various applications, such as automotive, HCI, learning, and many more would benefit from a real-time detection of cognitive load. A widely known application area, where the online detection of cognitive load might be very beneficial is driving. In fact, with the increasing use of eye-tracking and driver observation technology as input to adaptive assistance systems [Braunagel et al. 2017; Kasneci et al. 2017], workload detection could be implemented based on the physiological response of the pupil to increasing workload. Another field with great potential is virtual reality. Most VR-headsets can meanwhile be equipped with eye trackers, at the same time offering a working environment with controlled illumination. In such a setting, workload estimation could be perfectly utilized to automatically adjust the task difficulty to the user.

To date, what is preventing most of the above mentioned real-time application from making practical use of workload detection based on the pupillary response, is that most of the available algorithmic approaches that have been implemented so far, prove low capability to generalize [Lobo et al. 2016]. Classifiers are very subject-dependent and as such, each user who would want to use such an application would need to undergo a lengthy calibration procedure.

Walter et al. successfully applied regression algorithms to EEG data in a cross-subject setting [Walter et al. 2014] and applied their method to an online environment for arithmetic learning [Walter

et al. 2017]. In this environment they used the detected workload of university students to adjust the difficulty of the task and improve learning outcomes. While the results are promising, EEG measurements require more time and effort to prepare and are impractical to use on a wide basis.In another work, Lobo and colleagues combined EEG and eye-tracking features trying to achieve cross-subject classification of three different workload levels, but were unable to transfer the good subject-specific results to a cross-subject level [Lobo et al. 2016]. Another more successful attempt at cross-subject and cross-task workload estimation was done by Smith and colleagues [Smith et al. 2015]. They aggregated 66 data-channels, 4 of which were eye-related and performed a regression across different tasks and subjects. An approach only relying on eye tracking was chosen by Fritz et al. achieves 69% precision, but it incorporates not only the pupil, but the fixations and saccades during a programming task [Fritz et al. 2014]. This makes it unlikely to transfer well to different tasks.

In this work we use pure pupillometry which is not task-dependent like fixation- or saccade-related features. We evaluate how feasible it is to apply classifiers across different subjects and evaluate subjects with classifiers not trained on their data. We present a novel method to tackle this task with promising results. The loss in accuracy distinguishing low and high cognitive load is low when we move from intra-subject to cross-subject classification.

The remaining of the paper is organized as follows. Section 2 introduces the task and collection process of the dataset. In Section 3, light is shed on the features that we use for our classification and Section 4 presents the details of our workload classification method. The results are shown in Section 5 and discussed in Section 6.

2 DATA

We use data collected from a modified n-back task by Scharinger et al. [Scharinger et al. 2015] and an unpublished dataset also collected by Scharinger et al. based on the same task. We combined both datasets and treat them as one, since the task and collection process were identical for both.

2.1 Task and Stimulus

The n-back task consists of a randomly generated sequence of letter from the set $L = \{C, F, H, S\}$ that are shown one after another. Each letter is presented for 0.5 seconds and is followed by a 1.5 second period of black screen. Over the course of the experiment subjects must state whether the currently presented stimulus is the same as the one n trials before by pressing one of two buttons. For level 0, the task is to compare the presented stimulus letter with a constant one that is shown at the beginning of the task.

Scharinger and colleagues introduced flanking letters that were either congruent to the actual stimulus or incongruent, in which case they served as distractors. For the purpose of distinguishing different levels of cognitive load that correspond with the n-back levels, this minor modification will be disregarded for this work. Since these distractors were evenly distributed across all levels and subjects, their effect is negligible when it comes to the classification of the levels.

Every subject completed two blocks, where each of them consists of one 0-back task, one 1-back task and one 2-back task that were ordered randomly and had 154 trials each. Half of the trials were targets that coincided with the stimulus n trials before and the other half were non-targets that did not coincide. Participants underwent a training phase for each level before they started the trials that are analyzed here. They were presented with a block from the level they were training for and continued to receive a training block until they reached an accuracy of at least 60%.

2.2 Participants

The experiment was conducted with 28 students of the University of Tübingen (mean age 24.71, SD 4.12, 14 females), with German being their mother tongue. They received a payment of 8€ per hour and did not have any neurological disorders. All participants had normal or corrected-to-normal vision. The local ethics committee of Knowledge Media Research Center in Tübingen approved the study and participants gave their written informed consent at the beginning of the study.

Three subjects were excluded from the analysis, because of problems during the eye-tracking recording or poor tracking ratios.

2.3 Devices

The stimulus was presented on a 22-inch monitor with a resolution of 1,680 × 1,050 and a font-size of 25 utilizing the Arial font. All letters were presented in gray on a black background and eyes were tracked with a 250 Hertz SMI remote eye-tracking system. A chin rest ensured the distance of 70 cm was constant over the course of the test. Recording took place in SMI iView X 2.7.13. Calibration was done at the beginning and during each break using SMI's built in 9-point calibration.

3 FEATURES

Several different features can be derived from the pupil signal. We used the pupil diameter, blinks and the Index of Cognitive Activity as proposed by Marshall [Marshall 2000], which is described in more detail in the following.

3.1 Pupil

Across a wide variety of tasks and conditions, the pupil has been observed to increase in diameter with increasing cognitive load [Alnaes et al. 2014; Chen and Epps 2014; G. Brown et al. 1999; Klingner et al. 2011]. The task causes cognitive load, and as a result, decreases the parasympathetic activity in the peripheral nervous system, leading to an increase in pupil diameter [Kramer 1990]. This behavior of the pupil shows in many of tasks, including short-term memory, language processing, reasoning, perception, sustained attention and selective attention [Beatty 1982]. Additionally, it was found to fulfill Kahneman's three criteria for indicating processing load: the ability to reflect differences in processing load within a task, between different tasks, and finally between different individuals [Kahneman 1973].

We observed blink artifacts in our dataset, so we removed those parts from the pupil signal that had an unreasonably large slope right before or after a sequence of missing data. During this sequences, the eyelid had either already covered part of the pupil during the onset of a blink or was not completely visible yet after

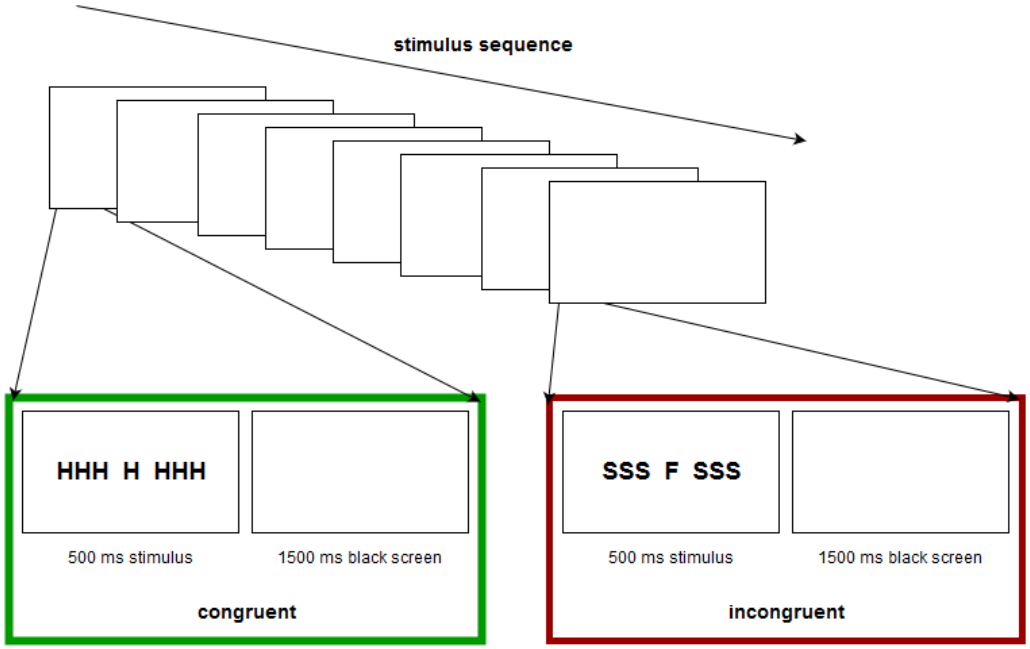


Figure 1: The *n*-back flanker task.

a blink. The pupil being tracked during that time results in pupil sizes changes that are not related to workload.

In this work we will employ the median and maximum of the pupil diameter as a feature to derive cognitive load. The median is used in order to be robust to outliers, remaining blink artifacts and other noise, while the maximum captures the peaks that occur while performing the *n*-back task.

3.2 Blinks

Studies have shown, that workload influences the rate and duration at which humans blink [Benedetto et al. 2011; Orden et al. 2001]. Increased task difficulty for a visual task also tends to increase the delay between two blinks, while the blink duration decreases [Veltman and Gaillard 1998]. Since the *n*-back task only varies the load on the memory capability, we expect a higher number of blinks per minute and little difference in blink duration [Fukuda et al. 2005].

Two different characteristics of blinks were used: average blink duration and number of blinks per minute.

3.3 ICA

The Index of Cognitive Activity as described by Marshall [Marshall 2000] uses a wavelet decomposition to detect sudden changes in pupil diameter that may be indicative of cognitive activity. These specific kind of changes can be caused by two different sources; either by cognitive activity or by the light reflex that causes the pupil to pulsate irregularly and continually. The first are short and abrupt while the latter are slower and larger. This difference is caused by the different activation and inhibition pattern of the radial muscles responsible for dilating and contracting the pupil respectively. Wavelet decomposition is suited to differentiate between the two sorts of events.

Since interpolation may have a great influence on the decomposition when large gaps occur, but a continuous signal is need for the decomposition, we followed the suggestion given by Marshall and leave out those parts. As suggested for the recording rate of 250 Hz, we use the Daubechies 16 wavelet. What we finally use as a feature for classification is the number of ICA-events per minute.

4 METHODS

To get a representative set to train models with, 50 sequences of set length were randomly sampled from every session, subject and level without overlap. Every level was treated as one long sequence from which sub-sequences were taken at random. The random sampling ensures that the regular structure of the dataset with 154 2-second-trials does not influence the training set and it is indeed representative. Collecting 50 samples from each level and session results in 100 samples per level for each subject which ensures that enough data per class is available to handle 5-10 features.

Samples with less than 70% usable data points were dismissed. With one level consisting of 154 trials, we can work with 308 seconds which - in theory - allows these sequences to be of length 6 seconds

at most. We will examine sequences of 1 to 5 seconds, however. With longer sequence we would not be able to reach 50 random samples, considering that some are unsuitable due to too much missing data and the combination of the exact position of a sample being random and sequences not overlapping.

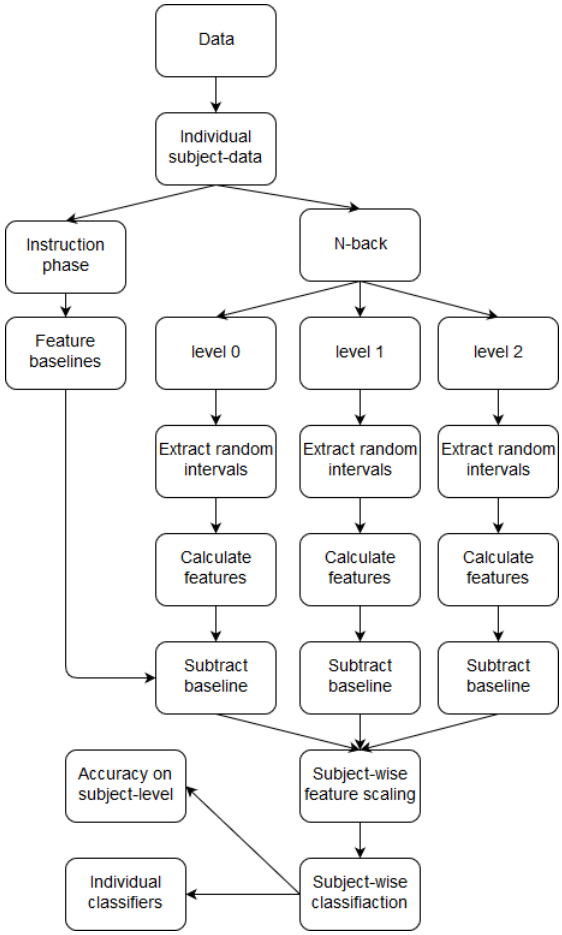


Figure 2: Summary of the construction of individual classifiers. This procedure is applied to the data of each subject in order to train classifiers explicitly suited for them.

We first use our data to train subject-wise classifiers to distinguish between different workload levels and then apply these classifiers to data of subjects that they were not trained on. For the cross-subject classification, two cases were examined: offline and online. Offline refers to the situation that one wants to analyze data that is already collected, whereas online means workload classification while the subject is still completing the given task.

4.1 Intra-subject classification

For better inter-subject comparability, the same features that will serve for classification are also calculated for the instruction phase to have a baseline. This baseline is then subtracted. This way, only relative changes are regarded and the great variance in baselines becomes less relevant. In addition, we normalize the resulting features to have a mean of 0 and a standard deviation of 1. Figure 2 illustrates the classification procedure. The per-subject classifiers described in this section will later be used for cross-subject classification.

We then use forests of 100 extremely randomized trees (Extra-Trees)[Geurts et al. 2006] to classify different levels of the *n*-back task, either pairwise or all three levels jointly.

One criterion for the choice of classifier is its good accuracy compared to simpler approaches and its tendency not to overfit to the data. In addition, Extra-Trees are far less computationally expensive compared to other tree ensemble classifiers and allow real-time evaluation and quick training. The implementation was done in python using the scikit-learn toolbox [Pedregosa et al. 2011]. The number of trees can be adjusted at will. A larger forest mainly increases the training time, whereas the time needed to evaluate a new sample only marginally increases.

4.2 Offline cross-subject classification

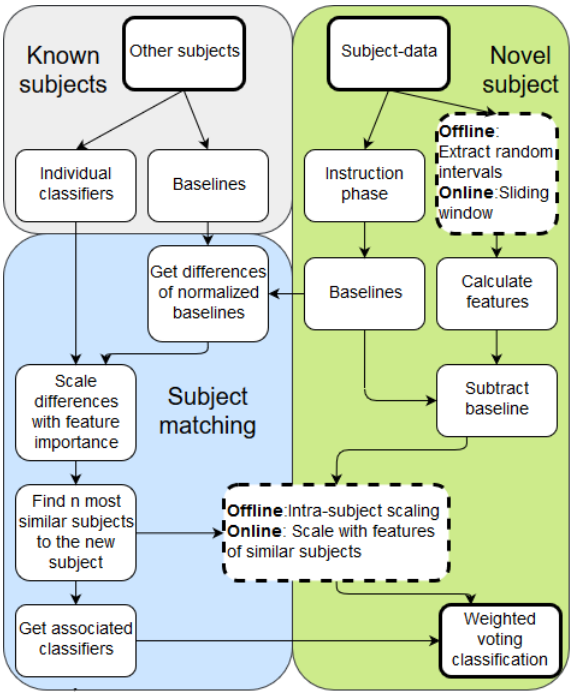


Figure 3: Voting scheme for cross-subject classification. The dashed line shows the difference between online and offline classification.

Once the per-subject workload classifiers are trained (4.1), they will be employed to estimate the workload of other subjects in a weighted voting scheme. For this purpose, features from the instructional phase are gathered and used as a baseline to determine the

pairwise similarity between two subjects. The similarity between subject x_{new} that we want to classify and subject x_{old} that already has a classifier is calculated as follows:

$$s(x_{new}, x_{old}) = \frac{1}{\sum w_{x_{old}} |f_{x_{new}} - f_{x_{old}}|}, \quad (1)$$

whereas f_{x_i} denotes the baseline feature vector of subject i and w_{x_i} the feature importance for the classifier of i. This way important features get higher weights and contribute more to the similarity measure. In random trees, good splits are chosen for nodes near the root, so the discriminating power of a feature can be seen as the expected fraction of samples that a feature is contributing to. Averaging over all trees in the forest helps in reducing the variance.

Based on this similarity, a nearest neighbor search is conducted and the closest n subjects are considered. Their classifiers are applied to the normalized data from the new subject x_{new} and their votes are weighted according with their normalized similarity. Figure 3 illustrates the process of cross-subject classification.

4.3 Online cross-subject classification

To simulate the use-case of actual real-time cross-subject application with out-of-the-box classifiers, we parse the data of a subject and every data point that is parsed is treated as a newly made measurement. This simulates the data recording of a new subject x_{new} performing the n-back task.

Over the course of this simulation, every newly available measurement is added to the existing ones of subject x_{new} to gradually build up a dataset for x_{new}. All currently available data from subject x_{new} is joined with the data of x_{old} whose classifier we want to use. Candidates for x_{old} are chosen in the same manner as in the offline scenario: by similarity of baselines weighted with the feature importance of the classifier. Then they are jointly scaled to have a mean of 0 and a standard deviation of 1, in order to reasonably apply the pre-trained classifier to the new data from subject x_{new}. This is a critical point, as the first measurements are without reference points for low and high workload and so scaling them does not lead to the outcome we want. Here scaling the features jointly with those of x_{old} for which we have those reference points makes sense. A certain error is introduced because the pupil features for low and high workload of x_{new} and x_{old} will not be exactly the same, but this error decreases as more data from x_{new} becomes available.

The same voting scheme as in the offline classification is then applied to classify the new sample. For this procedure we did not dismiss any sample due to quality issues or missing data, as such a measure of quality may not yet be available. This procedure is also presented in Figure 3.

5 RESULTS

To better evaluated which levels of workload - represented by the n-back levels - can be differentiated best, we trained classifiers to distinguish between specific pairs of n-back levels. To judge what amount of workload is caused by each n-back level, we first took a look at the descriptive statistics of EEG and performance measures presented in Table 1. It is apparent that level 0 and level 1 do not differ in difficulty as much as level 1 and level 2 do. This shows in all measurements, from EEG data to pupil diameter and performance measures. Therefore, we expect a better distinction between level 1 and level 2 than between level 0 and level 1. Consequently, the accuracy of the classifiers should be better.

Table 1: Descriptive statistics about the difficulty of each level illustrated by pupil performance and EEG measures.

These values are taken from Scharinger et al. [Scharinger et al. 2015]

	level 0	level 1	level 2
Pupil diameter [mm]	5.59	5.72	6.14
Reaction time [ms]	462	506	632
Accuracy [%]	88	86	79
Pz, P300 mean amplitude [μV]	4.05	3.76	2.75
Pz, upper aplha [μV^2/Hz]	6.80	6.25	5.43
Fz, theta [μV^2/Hz]	11.34	11.87	12.12

Table 2: Results overview

	intra-subject	cross-subject offline	cross-subject online
0 vs 1	69.8%	54.0%	53.8%
0 vs 2	82.4%	76.8%	70.4%
1 vs 2	79.4%	71.5%	66.8%
0 vs 1 vs 2	63.7%	46.7%	37.8%

The actual level of cognitive load that a subject experiences is not necessarily identical with the n-back level, because many factors can influence the workload. Since we cannot determine the cognitive load exactly, we use the task difficulty represented by the level of the n-back task as an indicator for workload.

For our study, the number of neighbors for the cross-subject classification was set to 9, but it can easily be adjusted to the computational power available. More neighbors means greater runtime, but possibly better accuracy. Table 2 shows the exact scores for 5 second sequences.

5.1 Intra-subject results

We used 10-fold cross-validation for the intra-subject classification to have a point of reference for the cross-subject accuracy. These are the scores we may be able to achieve, if the database of available classifiers is large enough to find good fits for every new subject. The levels 0 and 1 can be distinguished with an accuracy of 69.8%, while the levels 0 and 2 score 82.4% and the levels 1 and 2 79.4%. All three levels jointly, the classification is 63.7% accurate. Figure 4 shows the detailed results averaged for all subjects.

To evaluate the contribution of individual features to this result, the feature importance for the classifiers was calculated and averaged for every subject and specific level distinction. The details can be seen in Figure 7. Especially for short sequences, the pupil diameter is the most important feature, but with longer intervals,

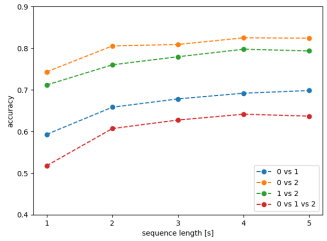


Figure 4: Accuracy of the intra-subject workload classification

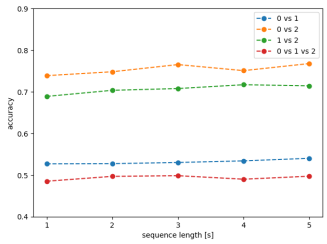


Figure 5: Accuracy of the cross-subject offline workload classification

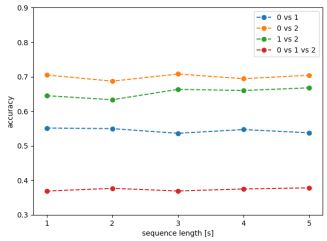


Figure 6: Accuracy of the cross-subject online workload classification

blink-related features increase in weight. The weight of the ICA is almost constant independent of the sequence length.

5.2 Cross-subject results

To measure the accuracy in a cross-subject setting where workload estimation is performed after all data is collected, we used the same samples for the intra-subject classification, but classified them with our cross-subject method as described in 4.2. Classification was performed for every sample individually and results were averaged across subjects. In this offline scenario, the levels 0 and 1 could be correctly distinguished with 54.0% accuracy, while in the case of level 0 and 2 76.8% were achieved and for level 1 and 2 71.5%. Figure 5 depicts the outcome of this classification.

For online classification, we used a sliding window to simulate newly available data and applied the procedure described in 4.3. The width of the sliding window was set to the sequence length that we also used for the training data, but in theory this is not strictly necessary. Since all features are either per second or median values, The window size can be set independently from the sequence length used for training. Using levels 0 and 1, 53.8% were classified correctly, with levels 0 and 2 it were 70.4% and with levels 1 and 2 66.8%. The detailed results are presented in Figure 5. On a computer equipped with an i7-7700HQ and 16GB RAM, feature calculation and classification took on average 0.134 seconds per sample of 5 seconds length, making it more than real-time capable.

It is unexpected that the accuracy of the intra-subject classification gets better with longer sequences, while the scores of the cross-subject classification do not. Most likely this is due to the fact that the main portion of the error is caused by choosing sub-optimal classifiers. This could be remedied with a larger set of subjects or a longer instruction for more stable baselines.

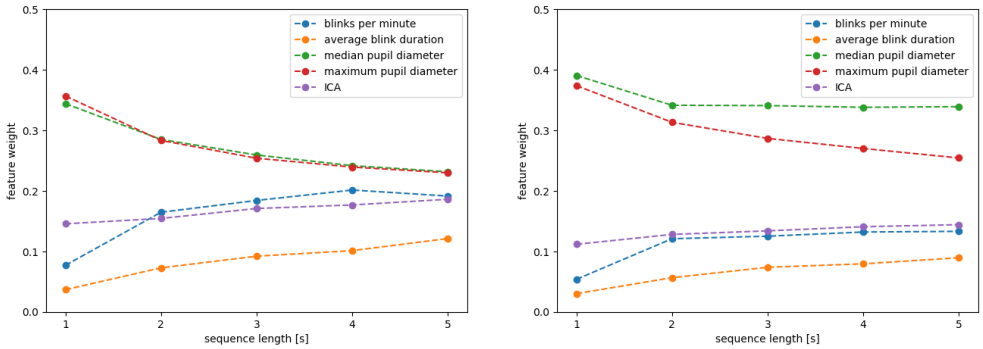
6 DISCUSSION

We have shown that cross-subject classification of cognitive load can be done with reasonable accuracy. With 82.4% our intra-subject scores for the distinction of low and high workload are slightly better than those reported in other *n*-back studies that employ eye tacking (e.g. [Hogervorst et al. 2014] with about 75%). One explanation may be the fact that we use the ICA as an additional feature, even though its feature importance is low. Another reason may be the removal of blink artefacts that increases the usefulness of the average pupil diameter as a feature or that we use the median instead of the mean. Since the lighting conditions also influence the pupil diameter change, maybe that is part of the reason, too. The pupil diameter change that occurs with increasing workload is more pronounced if the illumination is low, but the standard deviation also increases [Pfleging et al. 2016].

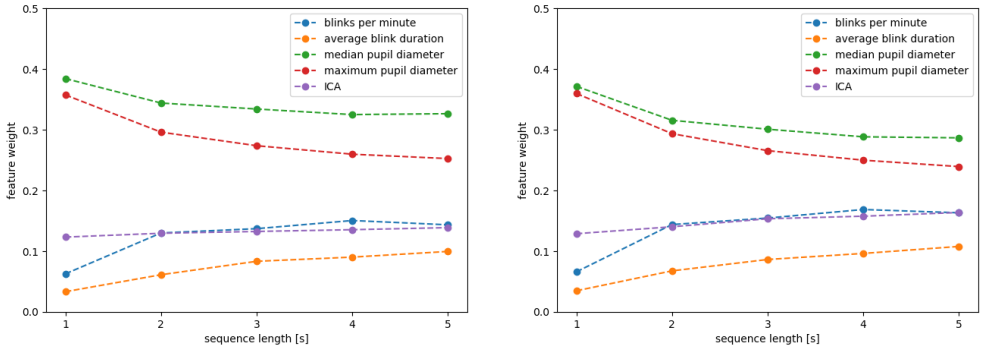
Looking at the feature weights, we can conclude that the pupil diameter is the best pupil-related indicator for workload. The median is the more stable indicator, whereas the importance of the maximum decreases over time which is expected. It is worth noting, that the importance of blink count and blink duration increases with the sequence length. If blinks reflect workload, this is to be expected. Sequences of short length most likely do not contain blinks and therefore they are not a feature to distinguish with. The longer the sequences get, the more likely it is for them to contain more than 0 blinks and the discriminating power of the feature increases. The ICA suffers from the same problem as blinks, but longer sequences only slightly increase its importance for classification.

It is curious that the performance of the cross-subject classifiers stagnates or slightly increases at best with the sequence length. It is possible that the quality of the classifiers as well as the individual data sample increases with longer sequences, but the main portion of the error is caused by choosing an unfitting classifier in the first place. Further investigation with a larger set of subjects is needed to test this hypothesis.

The biggest drop in accuracy can be observed for the level distinction of levels 0 and 1. Here the class differences were the smallest and so the error introduced by using an inappropriate classifier is punished the hardest.



(a) 0 vs 1 **(b) 0 vs 2**



(c) 1 vs 2 **(d) 0 vs 1 vs 2**

Figure 7
Average feature weights of the classifiers for specific workload level distinction.

What sets our method apart from others is the short time window that we evaluate. With 1-5 seconds, we only use a fraction of the time-span usually used which is 30-120 seconds [Hogervorst et al. 2014]. This gives us the opportunity to detect changes in workload that only last for a very short amount of time. Combined with the low computational demand of our method, real-time workload estimation can be performed at a very fine-grained level.

One of the benefits of our method is how easily the database can be expanded. Since we use per-subject classifiers with a voting scheme, there is no need to retrain a large model with all the data. If new data is available, we just need to train a model for a single subject and integrate its classifier into the database. Additionally, our approach is not restrictive when it comes to classifiers. The same neighbor matching and weighted voting can be performed with classifiers of different kinds. This way we can use Extra-Trees alongside SVM-classifiers or any classifier that can output its feature weights.

One factor that influences the accuracy of our workload classification is fatigue. Fatigue and exhaustion increase a subject's cognitive load while the pupil shrinks in size [Lowenstein and Loewenfeld 1962]. This makes it hard to compare samples from the same task that have long time intervals between them. In our study this does not have a systematic effect, that our classifiers may have learned, because we randomized the order of *n*-back levels for every subject. This however means that a classifier from a subject that first completed level 2 and one that did level 2 last are quite different. If we are able to compensate for this effect, classification accuracy will dramatically improve, because the difference in pupil diameter caused by fatigue can be substantial [Morad et al. 2000]. In our dataset the pupil decrease by 0.53mm averaged over all subjects, levels and sessions. This difference is about as large as the average pupil diameter difference between level 0 and level 2 and is likely to be responsible for a large portion of wrong classifications.

The same procedure is applicable to wearable eye trackers and the small but constant distance to the subject's eyes would likely yield even better results. This hypothesis needs testing, but we see great potential in head-mounted eye tracking. In addition, more advanced methods for pupil detection such as ElSe [Fuhl et al. 2016b] or PupilNet [Fuhl et al. 2016a] could be employed to detect eye features more robustly.

A next step would be to test how well our cross-subject approach performs across different tasks and with a database of different

lighting conditions. Also, a bigger database of subjects is needed to minimize the error if unsuited classifiers. This would be a big step towards an out-of-the-box algorithm for online workload detection that is invariant to circumstances and is universally applicable.

The generalization ability across different tasks and lighting conditions still needs to be examined, but enough data may make accurate online estimation of a subjects workload independent of task and circumstances possible.

ACKNOWLEDGMENTS

This research was funded by the LEAD Graduate School and Research Network [GSC1028], a project of the Excellence Initiative of the German federal and state governments. Tobias Appel was a doctoral student of the LEAD Graduate School and Research Network.

REFERENCES

D. Alnaes, M. H. Sneve, T. Espeseth, T. Endestad, S. H. P. Van de Pavert, and B. Laeng. 2014. Pupil size signals mental effort deployed during multiple object tracking and predicts brain activity in the dorsal attention network and the locus coeruleus. *JOURNAL OF VISION* 14, 4 (2014). https://doi.org/{10.1167/14.4.1}

J. Beatty. 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin* 91, 2 (March 1982), 276–292.

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

C. Braunagel, W. Rosenstiel, and E. Kasneci. 2017. Ready for Take-Over? A New Driver Assistance System for an Automated Classification of Driver Take-Over Readiness. *IEEE Intelligent Transportation Systems Magazine* 9, 4 (2017), 10–22.

S. Chen and J. Epps. 2014. Using Task-Induced Pupil Diameter and Blink Rate to Infer Cognitive Load. *HUMAN-COMPUTER INTERACTION* 29, 4 (JUL 4 2014), 390–413. https://doi.org/{10.1080/07370024.2014.892428}

T. Fritz, A. Begel, S. C. Müller, S. Yigit-Elliott, and M. Züger. 2014. Using Psycho-physiological Measures to Assess Task Difficulty in Software Development. In *Proceedings of the 36th International Conference on Software Engineering (ICSE 2014)*. ACM, New York, NY, USA, 402–413. https://doi.org/10.1145/2568225.2568266

W. Fuhl, T. Santini, G. Kasneci, and E. Kasneci. 2016a. PupilNet: Convolutional Neural Networks for Robust Pupil Detection. *arXiv preprint arXiv:1601.04902* (2016).

W. Fuhl, T. C Santini, T. Kübler, and E. Kasneci. 2016b. Else: Ellipse selection for robust pupil detection in real-world environments. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*. ACM, 123–130.

K. Fukuda, J. A. Stern, T. B. Brown, and M. B. Russo. 2005. Cognition, blinks, eye-movements, and pupillary movements during performance of a running memory Task. 76 (07 2005), C75–85.

G. G. Brown, S. S. Kindermann, G. Siegle, E. Granholm, E. Wong, and R. Buxton. 1999. Brain activation and pupil response during covert performance of the Stroop Color Word task. 5 (06 1999), 308–19.

P. Geurts, D. Ernst, and L. Wehenkel. 2006. Extremely randomized trees. *Machine Learning* 63, 1 (01 Apr 2006), 3–42. https://doi.org/10.1007/s10994-006-6226-1

M. A. Hogervorst, A-M. Brouwer, and J. B. F. van Erp. 2014. Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. *Front Neurosci* 8 (14 Oct 2014), 322. https://doi.org/10.3389/fnins.2014.00322 25352774[pmid].

D. Kahneman. 1973. *Attention and Effort*. Prentice Hall.

E. Kasneci, T. Kübler, K. Broelemann, and G. Kasneci. 2017. Aggregating physiological and eye tracking signals to predict perception in the absence of ground truth. *Computers in Human Behavior* 68 (2017), 450–455.

J. Klingner, B. Tversky, and P. Hanrahan. 2011. Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *PSYCHOPHYSIOLOGY* 48, 3 (MAR 2011), 323–332. https://doi.org/{10.1111/j.1469-8986.2010.01069.x}

A. E. Kramer. 1990. Physiological Metrics of Mental Workload: A Review of Recent Progress. (1990).

A. L. Kun, O. Palinko, Z. Medenica, and P. A. Heeman. 2013. On the Feasibility of Using Pupil Diameter to Estimate Cognitive Load Changes for In-Vehicle Spoken Dialogues. In *14TH ANNUAL CONFERENCE OF THE INTERNATIONAL SPEECH COMMUNICATION ASSOCIATION (INTERSPEECH 2013), VOLS 1-5 (Interspeech)*, Bimbot, F and Cerisara, C and Fougeron, C and Gravier, G and Lamel, L and Pellegrino, F and Perrier, P (Ed.). Int Speech Commun Assoc; europa org; amazon; Microsoft; Google; TcL SYTRAL; European Language Resources Assoc; ouaero; imaginove; VOCAPIA res; acapela; speech ocean; ALDEBARAN; orange; vecsys; IBM Res; Raytheon BBN Technol; voxygen, 3733–3737. 14th Annual Conference of the International-Speech-Communication-Association (INTERSPEECH 2013), Lyon, FRANCE, AUG 25-29, 2013.

B. Laeng, M. Ørbo, t. Holmlund, and M. Miozzo. 2011. Pupillary Stroop effects. *Cognitive Processing* 12, 1 (01 Feb 2011), 13–21. https://doi.org/10.1007/s10339-010-0370-z

J. L. Lobo, J. De Ser, F. De Simone, R Presta, S. Collina, and Z. Moravek. 2016. Cognitive Workload Classification Using Eye-tracking and EEG Data. In *Proceedings of the International Conference on Human-Computer Interaction in Aerospace (HCI-Aero '16)*. ACM, New York, NY, USA, Article 16, 8 pages. https://doi.org/10.1145/2950112.2964585

O. Lowenstein and I. E. Loewenfeld. 1962. The pupil. 2 (01 1962).

G. Marquart, C. Cabrall, and J. de Winter. 2015. Review of Eye-related Measures of Drivers' Mental Workload. *Procedia Manufacturing* 3 (2015), 2854 – 2861. https://doi.org/10.1016/j.promfg.2015.07.783 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015.

S. P. Marshall. 2000. Method and apparatus for eye tracking and monitoring pipil dilation to evaluate cognitive activity. (18 07 2000). https://patentimages.storage.googleapis.com/pdfs/9171d27ab488a900c7db/US6090051.pdf

Y. Morad, H. Lemberg, N. Yofe, and Y. Dagan. 2000. Pupillography as an objective indicator of fatigue. *Curr. Eye Res.* 21, 1 (Jul 2000), 535–542.

K. F. Van Orden, W. L., Scott Makeig, and Jung T-P. 2001. Eye Activity Correlates of Workload during a Visuospatial Memory Task. *Human Factors* 43, 1 (2001), 111–121. https://doi.org/10.1518/001872001775992570 arXiv:https://doi.org/10.1518/001872001775992570 PMID: 11474756.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

V. Peysakhovich, F. Dehais, and M. Causse. 2015. Pupil diameter as a measure of cognitive load during auditory-visual interference in a simple piloting task. In *6TH INTERNATIONAL CONFERENCE ON APPLIED HUMAN FACTORS AND ERGONOMICS (AHFE 2015) AND THE AFFILIATED CONFERENCES, AHFE 2015 (Procedia Manufacturing)*, Ahram, T and Karwowski, W and Schmorrow, D (Ed.), Vol. 3. 5199–5205. https://doi.org/{10.1016/j.promfg.2015.07.583} 6th International Conference on Applied Human Factors and Ergonomics (AHFE), Las Vegas, NV, JUL 26-30, 2015.

B. Pfleging, D. K. Fekety, A. Schmidt, and A. L. Kun. 2016. A Model Relating Pupil Diameter to Mental Workload and Lighting Conditions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 5776–5788. https://doi.org/10.1145/2858036.2858117

C. Scharinger, A. Soutschek, T. Schubert, and P. Gerjets. 2015. When flanker meets the n-back: What EEG and pupil dilation data reveal about the interplay between the two central-executive working memory functions inhibition and updating. *Psychophysiology* 52, 10 (2015), 1293–1304. https://doi.org/10.1111/psyp.12500

A. M. Smith, B. J. Borghetti, and C. F. Rusnock. 2015. Improving Model Cross-Applicability for Operator Workload Estimation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 59, 1 (2015), 681–685. https://doi.org/10.1177/1541931215591148 arXiv:https://doi.org/10.1177/1541931215591148

A. Szulewski, N. Roth, and D. Howes. 2015. The Use of Task-Evoked Pupillary Response as an Objective Measure of Cognitive Load in Novices and Trained Physicians: A New Tool for the Assessment of Expertise. *ACADEMIC MEDICINE* 90, 7 (JUL 2015), 981–987. https://doi.org/{10.1097/ACM.0000000000000677}

J. A. Veltman and A. W. K. Gaillard. 1998. Physiological workload reactions to increasing levels of task difficulty. *Ergonomics* 41, 5 (1998), 656–669. https://doi.org/10.1080/001401398186829 arXiv:https://doi.org/10.1080/001401398186829 PMID: 9613226.

C. Walter, W. Rosenstiel, M. Bogdan, P. Gerjets, and M. Spüler. 2017. Online EEG-Based Workload Adaptation of an Arithmetic Learning Environment. *Frontiers in Human Neuroscience* 11 (2017), 286. https://doi.org/10.3389/fnhum.2017.00286

C. Walter, P. Wolter, W. Rosenstiel, M. Bogdan, and M. Spüler. 2014. Towards Cross-Subject Workload Prediction.. In *Proceedings of the 6th International Brain-Computer Interface Conference*. Graz, Austria.

Manuscript 2

Title:
Predicting Cognitive Load in an Emergency Simulation Based on Behavioral and Physiological Measures

Authors:
Tobias Appel, Natalia Sevcenko,Franz Wortha, Katharina Tsarava, Korbinian Moeller, Manuel Ninaus, Enkelejda Kasneci, and Peter Gerjets

Published in:
Proceedings of the 2019 International Conference on Multimodal Interaction.

Time of publication:
October 2019

Included in this dissertation with permission from the ICMI.

Predicting Cognitive Load in an Emergency Simulation Based on Behavioral and Physiological Measures

Tobias Appel
tobias.appel@uni-tuebingen.de
LEAD Graduate School and Research Network
Tübingen, Germany

Natalia Sevcenko
Leibniz-Institut für Wissensmedien
Tübingen, Germany

Franz Wortha
LEAD Graduate School and Research Network
Tübingen, Germany

Katerina Tsarava
Leibniz-Institut für Wissensmedien
Tübingen, Germany

Korbinian Moeller
Leibniz-Institut für Wissensmedien
Tübingen, Germany

Manuel Ninaus
Leibniz-Institut für Wissensmedien
Tübingen, Germany

Enkelejda Kasneci
University of Tübingen
Tübingen, Germany

Peter Gerjets
Leibniz-Institut für Wissensmedien
Tübingen, Germany

ABSTRACT

The reliable estimation of cognitive load is an integral step towards real-time adaptivity of learning or gaming environments. We introduce a novel and robust machine learning method for cognitive load assessment based on behavioral and physiological measures in a combined within- and cross-participant approach. 47 participants completed different scenarios of a commercially available emergency personnel simulation game realizing several levels of difficulty based on cognitive load. Using interaction metrics, pupil dilation, eye-fixation behavior, and heart rate data, we trained individual, participant-specific forests of extremely randomized trees differentiating between low and high cognitive load. We achieved an average classification accuracy of 72%. We then apply these participant-specific classifiers in a novel way, using similarity between participants, normalization, and relative importance of individual features to successfully achieve the same level of classification accuracy in cross-participant classification. These results indicate that a combination of behavioral and physiological indicators allows for reliable prediction of cognitive load in an emergency simulation game, opening up new avenues for adaptivity and interaction.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → *Cross-validation.*

KEYWORDS

Cognitive Load, Eye Tracking, Heart Rate, Multimodal, Classification

ACM Reference Format:
Tobias Appel, Natalia Sevcenko, Franz Wortha, Katerina Tsarava, Korbinian Moeller, Manuel Ninaus, Enkelejda Kasneci, and Peter Gerjets. 2019. Predicting Cognitive Load in an Emergency Simulation Based on Behavioral and Physiological Measures. In *2019 International Conference on Multimodal Interaction (ICMI '19), October 14–18, 2019, Suzhou, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3340555.3353735

ICMI '19, October 14–18, 2019, Suzhou, China

ACM ISBN 978-1-4503-6860-5/19/10.
https://doi.org/10.1145/3340555.3353735

1 INTRODUCTION

Real-time user modeling in general and cognitive modeling in particular are essential for successfully implementing adaptive interfaces and environments. One important aspect of cognitive modeling is its consideration of cognitive load, which refers to the degree to which cognitive resources such as working memory are recruited while performing a task [6, 10]. Often it is beneficial to adapt a system in a way that the cognitive load experienced by a user does not exceed a critical level. An e-learning environment, for example, should

neither provide trivial learning materials, nor overstrain the user with materials and tasks they cannot cope with [30]. This also holds true for most computer environments involving different levels of difficulty, such as games, training simulations, or tutoring systems. An example of successful implementation of this kind of adaptivity is demonstrated by Yuksel and colleagues [35]. They used EEG data for adapting difficulty during piano lessons, which they observed to increase participants' learning gains. Wilson and Russel used physiological measures such as EEG, respiration, and heart rate, but also eye-fixation behavior to realize adaptivity [34] in an aviation simulation. The provided real-time adaptive feedback enhanced participants' performance.

However, assessing cognitive load can not only be used for adaptation. In many situations, its assessment can be valuable on its own, in particular in the absence of other outcome measures. In this case, reliable indicators of cognitive load may allow for evaluating which interaction modality is easiest to use, which type of interface causes the desired degree of cognitive load, or how difficult a certain activity is [3].

Yet many traditional ways of assessing cognitive load are impractical for many situations. Using questionnaires like the NASA task-load index (TLX) [15] make the user aware of the assessment, interrupting and impairing task performance and reducing immersion. As a retrospective measure, this is not an issue, but for real-time measuring and adaptation it is not the right tool.

Other methods make use of physiological changes caused by cognitive load to derive respective indicators. Direct impact of cognitive load was suggested to be measurable considering brain activity (e.g., by means of EEG) and participant-specific predictions were found to be accurate [22, 24]. Methods for measuring brain activity such as EEG have the drawback of usually being intrusive, uncomfortable to use, and requiring a high expertise to be set up. While there are mobile versions of such devices, mitigating some of the drawbacks, the intrusive nature making users aware of being monitored still remains. There are, however, physiological parameters which can be measured indirectly or with little effort. Eye-fixation behavior is a prime example of such measures because it is easy to measure, less invasive for the user, and allows for using changes in pupil diameter as an indicator of cognitive load [1, 20]. Moreover, changes in heart rate are also a common effect of variations of cognitive load that can be measured with little effort and intrusion. Combining these two streams of data increases predictive power [31].

In addition to estimating cognitive load through measure of pupil dilation, eye tracking has the benefit of allowing for evaluating participants' gaze behavior. This allows insights into cognitive processes and offers behavioral data as another means for assessing cognitive load. Gaze information can be used in conjunction with interaction log-files to better grasp how participants react under low or high cognitive load.

The advantage of jointly considering several (physiological) data streams for the assessment of cognitive load was previously demonstrated in several studies [13, 16, 17, 28]. Hussain et al. used a combination of heart rate, skin conductance, respiration data, and eye-tracking [17]. In addition, they synchronized these physiological measures with facial features and behavioral data in pursuit of classifying cognitive load under the effect of affective interference. Haapalainen and colleagues further added a skin temperature sensor and an EEG-headset to this set of sensors and achieved a mean classification rate of 81.1% for the distinction between low and high cognitive load for 6 cognitive tasks [13].

A frequent problem in practice arising from the estimation of cognitive load being participant-specific results in a lack of generalizability [23]. As a consequence, a cognitive model for each participant needs to be trained in order for adaptive systems to work. This would necessitate a lengthy calibration period involving examples of low and high cognitive load for each participant. This often seems impractical and may deter users from actually using the system in the first place.

In this article, we combine behavioral and physiological measures in a novel multimodal approach for classifying cognitive load in an emergency simulation game. We specifically aim to increase cross-participant generalizability up to the point where the accuracy of cross-participant classification is the same as within-participant classification. In particular, we want to develop a method for classifying cognitive load that satisfies the criteria of i) allowing for a robust estimation with ii) high classification accuracy and iii) generalizability across participants for potential iv) real-time application.

2 SETUP AND STIMULI

Task

Participants had to perform in different scenarios taken from an adapted version of the real-time simulation game Emergency [12]. The three scenarios were adapted specifically for the purpose of this study, Figure 1 providing an example. The goal of the game is to coordinate emergency forces. Participants take control of paramedics, ambulances, and firefighters to save people from emergency situations and fight fires in the scenario.

Participants first completed a tutorial introducing all game mechanics with instructions and in absence of time pressure, followed by three different scenarios: a car crash, burning buildings, and a train crash. Each scenario was presented in three versions, "easy", "medium", and "hard". The scenarios were always presented in the same order, starting with the easiest version of the car crash scenario going through the hardest version of the train crash scenario. The simulation's

difficulty was raised by increasing the number of tasks a participant had to complete as well as the number of units available, while maintaining the same time constraint. This necessitates planning more steps ahead, while also requiring more micro-managing and better prioritizing. As planning gets more sophisticated and time pressure increases, cognitive load should increase as a consequence. Because we lack means of measuring cognitive load directly, we instead aimed to classify task difficulty as a proxy for cognitive load. All instructions were presented in German.

Scenario 0: Tutorial. The tutorial provided instructions on how to give orders to emergency forces and which tasks needed to be carried out for successfully completing the game. It also introduced the different units participants needed to coordinate and their purpose in the simulation. There was no time limit.

Scenario 1: Car Crash. The first scenario featured a car crash at a crossroads. Some accident victims were trapped within their cars and the participant needed to send firefighters to free them. At the same time, other victims needed treatment by a paramedic before being transported to the hospital. There was a 5 minute time limit for this scenario.

Scenario 2: Burning Buildings. In this scenario, several buildings were on fire. This poses a more dynamic threat as the fire can spread to neighbouring buildings and paramedics cannot operate in close proximity to fire. Several victims needed to be saved from burning buildings with ladders, while others were in need of medical aid. To extinguish the flames, fire trucks and firefighters could be used, which differ in their effectiveness. The dynamics of the fire made this scenario inherently more difficult than the first one. The time limit was 7 minutes and 30 seconds.

Scenario 3: Train Crash. In the final scenario, participants faced a derailed train that had crashed into a building. As a consequence of the train having hit a building, the building had caught on fire. This also threatened the surrounding buildings. Furthermore, train wagons were deformed, making it necessary for them to be cut open to save trapped victims. Adding to these tasks, there were victims in need of medical aid. By combining all challenges participants faced before, the final scenario further increased the difficulty compared to the previous ones. This especially emphasizes the need for prioritizing actions because firefighters could either cut victims free or extinguish fires. Taking into consideration the complexity and volume of this scenario, the time limit was set to 10 minutes.

Apparatus

The experiment took place in individual sessions in a laboratory setting under constant light conditions. The emergency simulation and the eye tracker were installed on a notebook with a 16" screen driven at 1920 x 1080 resolution. The position of each participant was individually determined based on the calibration of the eye tracker. No chin rest was used. For the recording of eye-fixation behavior, we used a RED250 eye tracker from SensoMotoric Instruments (SMI) in combination with the SMI Experiment Center 3.7.60 software. The eye tracker was calibrated using SMI's integrated 9-point calibration procedure.

For the measurement of heart rate, a custom-made Bitalino wearable was used in combination with OpenSignals Revolution software. To increase the signal-to-noise ratio, we attached three pre-gelled electrodes to participants' chests, which were cable-connected with the Bitalino unit. For technical reasons, an additional laptop was necessary to acquire heart-rate data. The laptop was connected to the measuring computer via Bluetooth. The whole setup is depicted in Figure 2.

Participants

We gathered data from 47 right-handed participants (*mean age* = 24.6, *SD* = 6.3, 33 females). None of them reported psychological, neurological, or cardiovascular diseases. All participants provided written informed consent and were able to speak German on native speaker level. Participants were informed about the study aims and received monetary compensation after the experiment. We had to exclude 7 participant due to problems with recordings of eye-tracking, physiological or meta-data. Another 2 were excluded because they reported that they did not take the experiment seriously. Finally, 2 participants did not provide enough usable data for all scenarios, rendering their data partly unusable. The data of all remaining 36 participants was considered in the analyses. Participants with periods of low tracking ratios – time spans with invalid data caused by the pupil not being detected reliably – were deliberately included in our dataset. Including them renders the dataset more realistic and consequently the results more accurately reflect a real-world scenario.

3 FEATURES

We looked at features which have been used successfully in the past to detect cognitive load. We grouped the features into different categories of i) pupil dilation, ii) fixations, iii) saccades, iv) heart rate, and v) in-game activity. In this section, we describe them in more detail.

Pupil

Many studies have successfully used pupil diameter as an indicator for cognitive load [7, 19, 25]. Increasing cognitive load was observed to decrease parasympathetic activity in the peripheral nervous system, leading to an increase in pupil

Figure 1: Example of how the task looked like.

Figure 2: Setup that was used during data collection.

diameter [20]. This effect was found for different tasks, including short-term memory, language processing, reasoning, perception, as well as sustained and selective attention [1]. The effect was also consistently observed within a task, between tasks, and between individuals [18]. In a preprocessing step, we first removed all data points where pupil diameter was 0 or negative. Such artifacts typically occur in case of invalid data points. We also removed data points up to 100 ms directly before and after a blink. During these periods, the pupil is partly occluded by the eyelid or eyelashes and cannot be detected reliably. Finally, we linearly interpolated small gaps of up to 50 ms to increase the amount of usable data. Because the analyzed time periods are very short and thus susceptible to noise, we used the median of pupil diameter instead of the more commonly used mean. We expect this to result in more robust features and more reliable predictions. Additionally, we used the pupil diameter maximum to also consider peaks in pupil diameter. Both parameters are expected to increase with increasing task difficulty.

Fixations

Fixations describe a stable gaze on the same location usually lasting between 200 ms and 350 ms [27]. The number of fixations per second depends on many factors. Higher cognitive load was found to lead to fewer but longer fixations [8], whereas time pressure tended to decrease fixation duration while increasing the number of fixations per second [33]. We used the number of fixations per second as a feature. The difficulty levels of our scenarios were mainly driven by the number of emergency personnel to coordinate, the number of sub-tasks to perform and increases in time pressure, leading to the expectation of an increase in the number of fixations per second.

Saccades

Rapid eye movements that usually occur between fixations are called saccades. How cognitive load and task difficulty influence saccade characteristics strongly depends on the task at hand. There is evidence pointing towards an increase of average saccade amplitude in search tasks compared to free viewing [32]. With increasing task difficulty, the amount of visual exploration should decrease and participants' gaze

behavior should be dominated by specific goal-directed saccades. Hence, we expect to find higher saccade frequency during high-difficulty tasks. Because participants have to navigate the interface efficiently to perform well during the simulation, we expect a higher saccade amplitude during phases of high task difficulty.

Another form of saccades are so-called microsaccades. Microsaccades are small involuntary eye movements that can occur during a fixation. Studies have tied them to cognitive load in different situations. Non-visual tasks appear to reduce the number of microsaccades [9, 21, 29], while visually more demanding tasks appear to increase the frequency of microsaccades [2]. We used the method suggested by Krejtz and colleagues [21] to detect microsaccades, but focused on microsaccade frequency rather than amplitude or velocity. We expected to see an increase of miscrosaccade frequency with task difficulty.

Heart Rate

Just like pupil diameter increases as a consequence of reduced parasympathetic activity, heart rate was observed to increase as well. Various studies investigated the relationship between cognitive load and heart rate [4] and substantiated this association.

We used the raw ECG signal recorded by the Bitalino unit and processed it with the python package biosppy [5] which uses an approach by Hamilton [14] for QRS detection and provides us with a timestamp for every R-peak. A QRS complex is the main spike in an ECG signal, marking a heart beat, and an R-peak is the highest point of this complex. Using the exact R-peak points, we then calculated the number of heart beats per second, which we used as a feature for assessing task difficulty.

In-game Activity

Not only physiological measures can be used to assess task difficulty, but also the way participants interact with the simulation. Periods of high task difficulty should be accompanied by higher in-game activity. Successfully completing different sub-tasks – such as putting out fires or saving victims under time pressure – requires coordinating emergency forces while keeping the overall setting in mind. [1]

We used actions per second as a measure for in-game activity. Actions comprise opening the menu of an emergency respondent, selecting a specific command, and selecting a target for that command. For instance, commanding a paramedic to care for a certain victim results in 3 actions.

[1]Using the number of clicks may seem to render the classification problem trivial, but even without the in-game activity as a feature, classification accuracy is still around 70% and as our results show, it is not the dominant feature.

4 METHOD

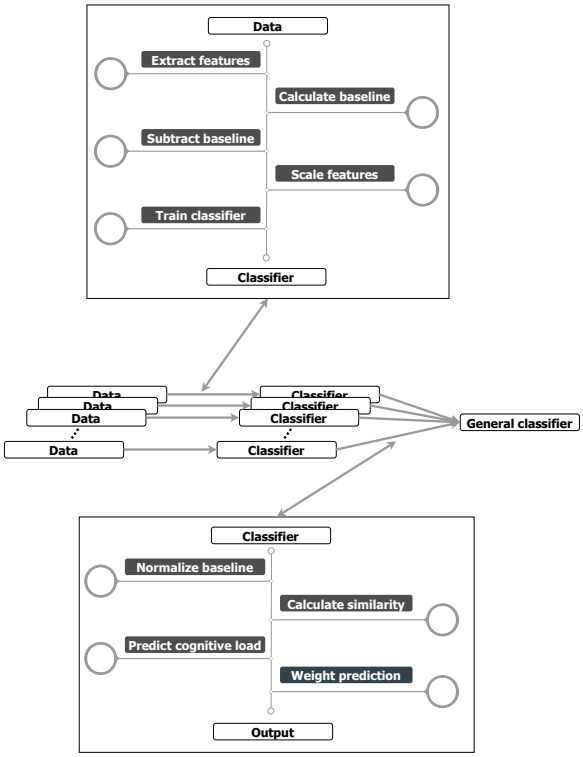


Figure 3: Schematic overview of our method showing how we train individual classifiers and combine them to a general one.

Our goal was to develop a robust classifier that works reliably as well as independently of individual participants. The main rationale underlying our approach was to train classifiers based on participants from a database and apply these classifiers to novel participants.

Participant-specific Classifiers

In the within-participant part of our approach, we aimed to train a classifier for each participant and scenario, classifying whether a participant currently plays the easy or hard version of the simulation. To train such classifiers, we needed samples from low- and high-difficulty periods to learn from. This prerequisite was achieved by drawing 60 random samples of 4 s intervals, half of which are obtained from the easy versions of each scenario and the other half from the hard version. This process was repeated for each participant and scenario. The 4 s interval enabled us to sample without overlap and still include participants that had completed the scenario well before the time limit was reached. Furthermore,

this interval length enabled us to reject samples with more than 30 % missing data, while still including all participants. This prevented any bias that may have been introduced by systematically excluding well-performing participants.

For each sample obtained, we calculated the features mentioned in Section 3. We used the tutorial as a baseline, meaning that we extracted the same features for the whole tutorial and used them for standardization, resulting in features that were a percentage change of the baseline. This did not require any further processing because all the used features were standardized per second. Choosing the tutorial as a baseline made it unnecessary to employ a dedicated calibration or baseline period, which increases real-world applicability of our approach.

After extracting the features, we z-standardized all features for each scenario. This resulted in all features having a mean of 0 and a standard deviation of 1, ensuring that all features had the same normalized scale and that any machine learning algorithm did not weight features according to their scaling. Finally, we trained a forest of 100 extremely randomized trees (Extra-Trees) [11] with a single participant's data to obtain a classifier specific to that participant. This resulted in one classifier per participant and scenario being able to distinguish a sample as either originating from a period of high or low task difficulty. Instead of just classifying into "low" or "high", we used class probabilities, which are scores between 0 and 1. All outputs above 0.5 were considered to reflect high task difficulty and all below 0.5 were considered to indicate low task difficulty. This provided the advantage of incorporating confidence into the prediction, improving generalizability in the following step. Our method was implemented in Python using the scikit-learn toolbox [26].

General Classifiers

Based on the classifiers for each participant, the next step was to generalize across participants. A naive way of approaching this issue would be to either train one classifier for all participants or to blindly apply the trained classifiers to other participants. It does, however, make little sense to apply participant A's classifier to participant B if their physiological or behavioral features are very different. Therefore, we weighted the prediction of individual classifiers according to how similar they were to the participant we wanted to apply them to.

First, we standardized the baselines for participant, guaranteeing that certain baseline features do not receive a higher weight when calculating the Euclidean distance between two baselines. Features on a larger scale tend to dominate the distance, because they result in larger numbers (e.g., there are a lot more fixations per second than in-game actions).

Let x be a novel participant whose cognitive load we want to classify, $sample_x$ a sample of x characterized with a set of features, and Y the set of participants we trained on. Every $y \in Y$ has a classifier c_y that predicts a value between 0 and 1 for $sample_x$. We combine these predictions according to the following equations:

$$sim(x, y) = \frac{1}{\sum weights_{c_y} |baseline_x - baseline_y|}$$

$$pred(sample_x) = \frac{\sum_{y \in Y} sim(x, y) pred_{c_y}(sample_x)}{\sum_{y \in Y} sim(x, y)}$$

$sim(x, y)$ refers to the baseline similarity between participants x and y, $weights_{c_y}$ to the normalized feature weights of classifier c_y, and $pred_y$ to the prediction of classifier c_y. This means we let each classifier c_y predict the cognitive load and weight these predictions according to the similarity between participants x and y. Additionally, we factor the feature weights of classifier c_y into the similarity, giving a higher weight to more important features. Dividing by the sum of all similarities normalizes these similarities and ensures that the prediction's final result is within the interval of $[0, 1]$.

5 RESULTS

In order to provide a frame of reference for the accuracy of our method, we first present descriptive statistics for the 3 scenarios in Table 1. This illustrates how difficult the individual scenarios were. The higher the increase in difficulty, the larger differences between the easy and the hard version should be and consequently the performance of our classification should be better. We focus on the distinction between easy and hard task difficulty, but our method can also be applied to the classification problems "easy vs. medium", "medium vs. hard" and for the multi-class problem "easy vs. medium vs. hard". Evaluating these problems and their results, however, is beyond the scope of this article.

Table 1: Descriptive statistics about the difficulty of the scenarios.

scenario	finished on time	average time of completion, if finished
1 easy	83.33 %	4:12
1 hard	83.33 %	4:16
2 easy	97.22 %	3:58
2 hard	36.11 %	6:31
3 easy	97.22 %	7:01
3 hard	33.33 %	8:57

Participant-specific Results

To evaluate how well our cross-participant approach works, we first ran participant-specific classifiers. This was performed using 10-fold cross-validation, to ensure that we do not artificially increase classification accuracy by overfitting. Table 2 shows the average accuracy for each scenario.

Table 2: Mean accuracy for within-participant classification of cognitive load.

scenario	accuracy
1	79.03%
2	70.14%
3	72.13%

Feature Weights

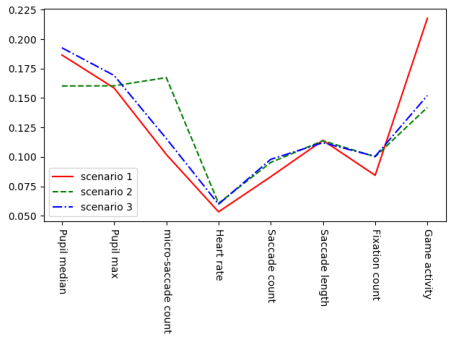


Figure 4: Average feature importance for each of the 3 scenarios.

To better understand our results, we also present average feature weights in Figure 4. Interestingly, pupil features have the highest weight, making them the most important features. We expected a large contribution from the pupil data and this result is in line with the literature. Another important feature is in-game activity. In particular, it is very important in the first scenario, because participants are not yet familiar with the controls, but by the time they have to perform the hard version they already acquired more experience, increasing the difference in in-game activity. Learning and experience reduce this effect in the subsequent scenarios. Microsaccades are the next most important category. Several publications indicated that cognitive load has an effect on microsaccade frequency [21], but usually with higher relative influence. Our recording frequency of 250 Hz may be too low to detect all microsaccades and most studies recorded microsaccades during very long fixations of several seconds, whereas our task resulted in rather short fixations.

As expected, saccade and fixation characteristics are predictive for task difficulty, but do not carry as much weight as other features. Finally, heart rate does not seem to matter very much for our algorithm. Nevertheless, looking at mean values per difficulty level, we still consistently find heart rate to increase with difficulty. A possible reason might be the short intervals analyzed. 4 s may be a very short time frame to evaluate heart rate meaningfully.

Cross-participant Results

After we trained our participant-specific classifiers, we applied them for cross-participant classification according to the schema described in Section 4. We performed leave-one-out cross-validation on the participant level, meaning that we withheld the participant to whom we applied our cross-participant algorithm. To also test how well our approach performs across different situations, we used classifiers which had been trained either on the same scenario or on one of the two others. The average results for classification accuracy are shown in Table 3.

Table 3: Mean accuracy for cross-participant classification of cognitive load when applying classifiers trained on different scenarios.

Classifier from \ applied to	1	2	3
1	80.56%	67.92%	70.10%
2	73.70%	70.37%	67.04%
3	76.13%	67.73%	69.81%

It is noticeable that our approach does not lose accuracy performing cross-subject classification compared to within-subject classification. Only in the case of scenario 3, accuracy drops slightly. This is likely to be the case, because towards the end of our experiment subjects adapt strategies that differ between them. Overall our approach generalizes well on participant level, which can be attributed to the weighting and various forms of standardization. We performed the same classification with only the 10 most similar participants instead of all participants and still got the same accuracy. This may be used to save runtime in case little processing power is available or when the number of participants gets very large.

Also note that the accuracy barely drops when we use data from another scenario, so our method does not only generalize well across participants but also across different scenarios. This is signified by the columns of Table 3. The diagonal shows the the results obtained by training and evaluating on the same scenario setting the bar for other classifiers, whereas the rest of the column depicts results from sub-optimal classifiers trained on other scenarios. As the

scenarios differ in regards to how well easy and hard version can be distinguished, meaningful row-wise comparisons can not be made.

With regard to runtime, calculation of features took 0.699 ms on average and weighted classification took 6.723 ms, thus resulting in a total runtime of 7.422 ms. The reported performance was measured on a laptop with an Intel(R) Core(TM) i7-7700HQ with 16 GB RAM running a non-optimized version of our algorithm that does not make use of parallel processing. Limiting the number of participants we consider for classification to the n most similar ones would further speed up our algorithm while still maintaining a high accuracy.

6 DISCUSSION

Our goal was a i) robust estimator offering ii) high accuracy, iii) generalizing well across participants, and is also iv) real-time capable. In conclusion, our approach satisfies all these criteria.

Firstly, we showed the robustness of our approach. We operate with a suboptimal baseline derived from the tutorial and did not exclude participants with low tracking rates. Furthermore, we use only one baseline from the very beginning of the experiment. As a consequence, fatigue influences pupil data over the course of the experiment, decreasing its diameter and thereby counteracting some effects of changes in cognitive load. Furthermore, luminance changes caused by the dynamic nature of the simulation also add noise that our methods shows resilience towards.

Secondly, a mean classification accuracy of 72 % appears sufficient for most real-world applications. Actual classification accuracy may most likely be higher because we use task difficulties for classification and not actual cognitive load. During a heterogeneous task, cognitive load is usually not at a constant level and can be low within a difficult task high within an easy task. Moreover, for participants that finished on time, we noticed a drop in predicted task difficulty towards the end of the task indicating they may not be challenged anymore. Additionally, most participants started any version of a scenario with low predicted task difficulty, likely because they were adjusting to the task. Both of these circumstances reduce nominal accuracy of our approach even though predictions may be accurate.

Thirdly, our results show that our method is able to generalize across participants. There is no drop in accuracy when we apply weighted predictions to withheld participants, indicating that we can expect the same level of accuracy when we apply our method to new participants. This even holds true when we restrict training the algorithm to the 10 most similar participants.

This is relevant for the last criterion of being executed in real-time. When processing power is limited, the number of participants whose classifiers are applied can be restricted, reducing runtime to a fraction of its original time without loss of classification accuracy. This will allow most devices to run the algorithm at an acceptable frequency.

Finally, our method maintaining its high classification accuracy even when applied across different scenarios indicates an additional degree of robustness.

Apart from the ability to generalize there are other benefits to the presented method. For instance, converting the continuous output to a binary classification is not mandatory, as the output can be used directly. The higher the score, the more likely is cognitive load to also be high. In the case of an adaptive environment, task difficulty could be adjusted in case it is found, for example, to be below 0.3 or above 0.7.

There are, however, limitations to our approach. We standardized features of each participant to a mean of 0 and a standard deviation of 1, which may cause problems when done in real time. If we assume the same order of tasks as in this work, participants start with an easy version of the first scenario. This means we only have data from periods of low cognitive load during this task and scaling does not work as expected. This may partly be mitigated by applying the scaling function from participants from our database, meaning we subtract the mean of a recorded participant and divide by their standard deviation. This will still cause minor loss in accuracy, but partly solves the problem. As soon as we have data from periods of low and high cognitive load, this is not an issue anymore.

While our approach can be used for many different tasks, the trained estimators can not. The fixation and saccade characteristics are specific to our simulation, so any classifier we trained has limited use for other tasks. If we ignore fixations and saccades to focus on the less task-dependent features like pupil diameter, microsaccades and heart rate, the range of application widens, but the accuracy for a specific task is reduced.

As a future perspective, we are planning a follow-up study using the method presented in this article to create an adaptive version of the emergency simulation. The 36 participant-specific classifiers we trained should suffice as a database to evaluate cognitive load of new participants in real time and we are confident that we can implement a system allowing for real-time adaptation of cognitive load caused by the simulation.

ACKNOWLEDGMENTS

This research was funded by the LEAD Graduate School and Research Network [GSC1028], a project of the Excellence Initiative of the German federal and state governments. Tobias Appel was a doctoral student of the LEAD Graduate School and Research Network.

REFERENCES

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

[2] Simone Benedetto, Marco Pedrotti, and Bruce Bridgeman. 2011. Microsaccades and exploratory saccades in a naturalistic environment. *Journal of Eye Movement Research* 4, 2 (2011), 1–10.

[3] Maneesh Bilalpur, Mohan Kankanhalli, Stefan Winkler, and Ramanathan Subramanian. 2018. Eeg-based evaluation of cognitive workload induced by acoustic parameters for data sonification. ACM, 315–323.

[4] Yati N Boutcher and Stephen H Boutcher. 2006. Cardiovascular response to Stroop: effect of verbal response and task difficulty. *Biological Psychology* 73, 3 (2006), 235–241.

[5] Carlos Carreiras. 2015–. BioSPPy: Biosignal Processing in Python. https://github.com/PIA-Group/BioSPPy

[6] Fang Chen, Jianlong Zhou, Yang Wang, Kun Yu, Syed Z Arshad, Ahmad Khawaji, and Dan Conway. 2016. *Robust multimodal cognitive load measurement.* Springer.

[7] Siyuan Chen and Julien Epps. 2013. Automatic classification of eye activity for cognitive load measurement with emotion interference. *Computer Methods and Programs in Biomedicine* 110, 2 (2013), 111–124.

[8] Michel De Rivecourt, Marianne Kuperus, Wendy J Post, and Lambertus JM Mulder. 2008. Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight. *Ergonomics* 51, 9 (2008), 1295–1319.

[9] Xin Gao, Hongmei Yan, and Hong-jin Sun. 2015. Modulation of microsaccade rate by task difficulty revealed through between-and within-trial comparisons. *Journal of Vision* 15, 3 (2015), 3–3.

[10] Peter Gerjets, Carina Walter, Wolfgang Rosenstiel, Martin Bogdan, and Thorsten O. Zander. 2014. Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain-computer interface approach. *Frontiers in Neuroscience* 8 (2014), 385. https://doi.org/10.3389/fnins.2014.00385

[11] Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine Learning* 63, 1 (2006), 3–42.

[12] Promotion Software GmbH. 1999. World of Emergency. [https://www.world-of-emergency.com/?lang=en

[13] Eija Haapalainen, SeungJun Kim, Jodi F Forlizzi, and Anind K Dey. 2010. Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*. ACM, 301–310.

[14] Patrick S Hamilton. 2002. Open source ECG analysis software documentation. *Computers in Cardiology* 2002 (2002), 101–104.

[15] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in Psychology*. Vol. 52. Elsevier, 139–183.

[16] SM Hadi Hosseini, Jennifer L Bruno, Joseph M Baker, Andrew Gundran, Lene K Harbott, J Christian Gerdes, and Allan L Reiss. 2017. Neural, physiological, and behavioral correlates of visuomotor cognitive load. *Scientific Reports* 7, 1 (2017), 8866.

[17] M Sazzad Hussain, Rafael A Calvo, and Fang Chen. 2013. Automatic cognitive load detection from face, physiology, task performance and fusion during affective interference. *Interacting with Computers* 26, 3 (2013), 256–268.

[18] Daniel Kahneman. 1973. *Attention and effort*. Vol. 1063. Citeseer.

[19] Jeffrey M Klingner. 2010. *Measuring cognitive load during visual tasks by combining pupillometry and eye tracking.* Ph.D. Dissertation. Stanford University Stanford, CA.

[20] Arthur F Kramer. 1991. Physiological metrics of mental workload: A review of recent progress. *Multiple-task Performance* (1991), 279–328.

[21] Krzysztof Krejtz, Andrew T Duchowski, Anna Niedzielska, Cezary Biele, and Izabela Krejtz. 2018. Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PLOS ONE* 13, 9 (2018), e0203629.

[22] Shiba Kuanar, Vassilis Athitsos, Nityananda Pradhan, Arabinda Mishra, and Kamisetty R Rao. 2018. Cognitive Analysis of Working Memory Load from EEG, by a Deep Recurrent Neural Network. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2576–2580.

[23] Jesus L Lobo, Javier D Ser, Flavia De Simone, Roberta Presta, Simona Collina, and Zdenek Moravek. 2016. Cognitive workload classification using eye-tracking and EEG data. In *Proceedings of the International Conference on Human-Computer Interaction in Aerospace*. ACM, 16.

[24] Caitlin Mills, Igor Fridman, Walid Soussou, Disha Waghray, Andrew M Olney, and Sidney K D'Mello. 2017. Put your thinking cap on: detecting cognitive load using EEG during learning. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. ACM, 80–89.

[25] Oskar Palinko, Andrew L Kun, Alexander Shyrokov, and Peter Heeman. 2010. Estimating cognitive load using remote eye tracking in a driving simulator. In *Proceedings of the 2010 Symposium on Eye-tracking Research & Applications*. ACM, 141–144.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[27] Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* 124, 3 (1998), 372.

[28] Christian Scharinger, Yvonne Kammerer, and Peter Gerjets. 2015. Pupil Dilation and EEG Alpha Frequency Band Power Reveal Load on Executive Functions for Link-Selection Processes during Text Reading. *PLOS ONE* 10, 6 (06 2015), 1–24. https://doi.org/10.1371/journal.pone.0130608

[29] Eva Siegenthaler, Francisco M Costela, Michael B McCamy, Leandro L Di Stasi, Jorge Otero-Millan, Andreas Sonderegger, Rudolf Groner, Stephen Macknik, and Susana Martinez-Conde. 2014. Task difficulty in mental arithmetic affects microsaccadic rates and magnitudes. *European Journal of Neuroscience* 39, 2 (2014), 287–294.

[30] Martin Spüler, Carina Walter, Wolfgang Rosenstiel, Peter Gerjets, Korbinian Moeller, and Elise Klein. 2016. EEG-based prediction of cognitive workload induced by arithmetic: a step towards online adaptation in numerical learning. *ZDM* 48, 3 (01 Jun 2016), 267–278. https://doi.org/10.1007/s11858-015-0754-8

[31] Zach Chuanzhong Tan, Bryan Reimer, Bruce Mehler, and Joseph F Coughlin. 2011. Detection of elevated states of cognitive demand in drivers in a naturalistic driving environment. In *Proceedings of the 2nd Annual International Conference on Advanced Topics in Artificial Intelligence (ATAI 2011)*. 24–25.

[32] Benjamin W Tatler, Roland J Baddeley, and Benjamin T Vincent. 2006. The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. *Vision Research* 46, 12 (2006), 1857–1862.

[33] Karl F Van Orden, Wendy Limbert, Scott Makeig, and Tzyy-Ping Jung. 2001. Eye activity correlates of workload during a visuospatial memory task. *Human Factors* 43, 1 (2001), 111–121.

[34] Glenn F Wilson and Christopher A Russell. 2007. Performance enhancement in an uninhabited air vehicle task using psychophysiologically determined adaptive aiding. *Human Factors* 49, 6 (2007), 1005–1018.

[35] Beste F Yuksel, Kurt B Oleson, Lane Harrison, Evan M Peck, Daniel Afergan, Remco Chang, and Robert JK Jacob. 2016. Learn piano with BACh: An adaptive learning interface that adjusts task difficulty based

on brain state. In *Proceedings of the 2016 CHI conference on Human Factors in Computing Systems*. ACM, 5372–5384.

Manuscript 3

Title:
Cross-task and Cross-participant Classification of Cognitive Load in an Emergency Simulation Game

Authors:
Tobias Appel, Peter Gerjets, Stefan Hoffmann, Korbinian Moeller, Manuel Ninaus, Christian Scharinger, Natalia Sevcenko, Franz Wortha, and Enkelejda Kasneci

Unpublished manuscript

Manuscript 3

Cross-task and Cross-participant Classification of Cognitive Load in an Emergency Simulation Game

Tobias Appel, *Student Member, IEEE,* Peter Gerjets, Stefan Hoffmann, Korbinian Moeller, Manuel Ninaus, Christian Scharinger, Natalia Sevcenko, Franz Wortha, Enkelejda Kasneci

Abstract—Assessment of cognitive load is a major step towards adaptive interfaces. However, non-invasive assessment is rather subject as well as task specific and generalizes poorly, mainly due to methodological limitations. Additionally, it heavily relies on performance data like game scores or test results. In this study we present an eye-tracking approach that circumvents these shortcomings and allows for generalizing well across participants and tasks. First, we established classifiers for predicting cognitive load individually for a typical working memory task (n-back), which we then applied to an emergency simulation game by considering the k most similar ones and weighting their predictions. Standardization steps helped achieving high levels of cross-task and cross-participant classification accuracy between 63.78% and 67.25% for the distinction between easy and hard levels of the emergency simulation game. These very promising results may pave the way for novel adaptive computer-human interaction across domains and particularly for gaming and learning environments.

Index Terms—Eye Tracking, Physiology, Intelligent Systems, Cognitive Model, Physiological Measures, Psychology, Adaptive and Intelligent Educational Systems.

1 Introduction

In many digital environments designed for learning, working or even for entertainment purposes there is a close link between users' current cognitive load and their affective experiences. An important impact of users' cognitive load on their affective states has been demonstrated repeatedly (e.g. [1], [2], [3]). For instance, in a learning context, it is highly relevant for users' subjective experiences that the instructions provided by a learning environment are neither over-straining (and thereby frustrating) for learners nor under-challenging (and thus boring) due to a lack of workload imposed [4]. However, imposing an optimal level of cognitive load onto learners that keeps them engaged and satisfied as well as in their zone of proximal development [5] might be a highly learner-specific issue that strongly depends on individual learners' prerequisites in terms of their prior knowledge and abilities. Thus, for calibrating learning experiences with regard to their cognitive demands and affective connotations in order to optimize the resulting learning outcomes technical support systems might be helpful that allow for a real-time adjustment of the cognitive-load level imposed by a learning environment.

Systems able to detect and properly react to a user's cognitive load in order to calibrate their affective and cognitive experiences would of course also offer considerable benefits for numerous other applications beyond learning environments - ranging from the workplace to the digital playground. For instance, in potentially stressful digital working environments (such as systems for surgical assistance, engine control or emergency management), in which errors might have serious and life-threatening consequences, individuals might also experience strong and fluctuating affective reactions related to their current level of cognitive (over-)load. Monitoring cognitive load in these contexts and providing respective feedback and support to users might not only help to avoid errors related to high cognitive load but also to improve the overall affective experience. For instance, (truck) drivers or other workers controlling complex engines may be prompted to take breaks or can be provided with individualized training when detected to be over-strained in specific situations (e.g. [6] or [7]). Other examples might be conceivable in the medical domain where surgeons might be relieved when necessary, or in aviation scenarios where pilots may be provided with support from their copilots or from assistance systems depending on their cognitive-load levels [8].

Beyond scenarios related to learning or working, gaming also seems to be a prime area for applying adaptive procedures based on cognitive load measurement in order

Tobias Appel was with LEAD Graduate School and Research Network, Tübingen, Germany. He is now with the Department of Human-Computer Interaction, University of Tübingen, Tübingen , Germany.
Peter Gerjets is with the Leibniz-Institut für Wissensmedien, Tübingen, Germany.
Stefan Hoffman is with Promotion Software GmbH, Tübingen, Germany.
Korbinian Moeller was with the Leibniz-Institut für Wissensmedien, Tübingen. He is now with the Centre for Mathematical Cognition, University of Loughborough, UK.
Manuel Ninaus is with the Leibniz-Institut für Wissensmedien, Tübingen, Germany.
Christian Scharinger is with the Leibniz-Institut für Wissensmedien, Tübingen, Germany.
Natalia Sevcenko is with the Leibniz-Institut für Wissensmedien, Tübingen, Germany and with Daimler AG, Stuttgart, Germany.
Franz Wortha was with LEAD Graduate School and Research Network, Tübingen, Germany. He is now with the Leibniz-Institut für Wissensmedien, Tübingen, Germany.
Enkelejda Kasneci is with the Department of Human-Computer Interaction, University of Tübingen, Tübingen, Germany.
Manuscript received TBA;

to optimize affective user experiences. For instance, in cases where an obstacle in a game is too difficult for a player to overcome, frustration may set in and the gaming experience may suffer. Contrarily, when a game is too easy in relation to users current abilities, gaming may also not be experienced as enjoyable. Both cases can be circumvented by adapting the degree of difficulty based on the player's current level of cognitive load. Thus, with accurate estimations of cognitive load during gaming, automatic adaptations might be enabled that prevent negative affective states (such as boredom, frustration, and stress) and enhance positive affects (such as engagement, joy, and satisfaction). A prime example of a desirable affective state in gaming in and many other scenarios of human-computer interaction is flow [9]. Flow is considered a positive affective state of optimal experience [10] that creates pleasure by balancing the challenge of the task at hand and the available capabilities of the user. Measuring a person's cognitive load online might help to adapt the levels of difficulty to a degree where it still constitutes enjoyable challenge but does not over-strain the user.

Usually, cognitive load is measured by using self-reports, such as the NASA task-load index (TLX) [11], or by obtaining performance metrics. These traditional approaches, however, have some drawbacks that render them impractical for systems aiming at real-time adaptations. In particular, filling out a questionnaire about the level of cognitive load currently experienced might strongly interfere with task performance and immersion and is therefore unsuitable for most applications. Moreover, questionnaire data are rather subjective and my be influenced by many factors with the current level of cognitive load being only one of them [12].Thus, for real-time adaptations to cognitive load levels, a more unobtrusive method would be required that does not interrupt the current task like questionnaires but offers a reliable, objective and continuous indirect online estimation of user's cognitive load.

Performance metrics such as test scores or task completion times are indirect and thus less interfering compared to questionnaires. However, they are usually only available at specific points in time and can not be measured continuously as would be required for real-time adaptations to cognitive load levels. For instance, in the case of digital learning environments, one would aim at measuring cognitive load levels during the learning process for adapting difficulty levels of learning materials and not only after a learning task is completed (e.g. by means of a test). Thus, the required continuous and unobtrusive cognitive-load monitoring can usually neither be provided by performance metrics nor by questionnaire data [13].

An alternative approach for assessing cognitive load is based on physiological measures. Cognitive load causes physiological reactions that can be measured by sensors [13], [14], [15]. The most reliable indicators are changes that occur in the brain, but measuring these changes is intrusive, hard to set up, and not feasible in broad real-world settings. In this context, methods like electroencephalography (EEG) or near-infrared spectroscopy (NIRS) require very specialized hardware and expertise to operate. Beres [16] gives a good overview of EEG measures, including their advantages and drawbacks - such as the number of required trials per experiment - that illustrate why these measures are rather unsuitable in the context of most real-life adaptive systems. Less intrusive sensors include heart rate monitors and devices for measuring skin conductance, which however seem to lack accuracy and/or validity for measuring cognitive load [17]. Finally, eye tracking measures such as eye-fixation features offer a good alternative to the aforementioned physiological signals. They do not require physical contact with participants, can be obtained in real-time, and have been comprehensively demonstrated to be associated with cognitive load [18]. When obtained by means of webcams, eye-tracking measures have the potential to become available to a broad audience across various application domains. Moreover, with the increasing integration of eye-tracking technology in VR, AR, and smart glasses [19], this physiological signal can also be measured in high quality in a variety of applications in the future [20].

One of the major limitations of physiological indicators such as eye-tracking data for measuring cognitive load consists in the difficulty to generalize measures across tasks and across participants [21], rendering cross-task and cross-participant predictions or real-time assessments virtually impossible. This drawback, however, is not limited to eye-tracking data or physiological measures in general but applies to many algorithms for real-time workload assessment as Heard et al. conclude in their meta-review [22]. System designed for real-time assessments usually need to be adapted to individual participants and/or specific tasks in order to yield reliable predictions. This usually requires data collection for lengthy calibration procedures for each participant rendering these systems time consuming and inconvenient for users. Even with individual calibration, generalizations to different tasks or applications are usually poor, resulting in a necessity for repeated calibrations for different tasks and/or applications. Currently, there is no satisfying general purpose classifier for cognitive load available.

Many researchers have worked on the problem of either cross-task or cross-participant estimations of cognitive load (see below for a more detailed discussion), but with limited success so far. Usually, although intra-participant results are good generalization results are limited or do not even exceed chance level (e.g. [23], [24], [25]).

In this article, we present a novel and intuitive approach how to remedy these methodological shortcomings. We show how a machine learning approach might be used for cognitive load detection based on eye-tracking data to allow for successful generalization across participants and tasks. We employ a schema of weighted votes that combines participant-specific classifiers into a composite classifier with a broader scope offering generalization ability across participants and tasks. Our method might thus be able to pave the ground for out-of-the-box solutions for adaptive human-computer interaction based on a reliable assessment and classification of users' cognitive load independent of the user and task at hand. As a result, users' affective experiences during human-computer interaction in contexts such as learning, working, or gaming might strongly benefit in terms of avoiding frustration, boredom, or stress and in terms of enhancing engagement, joy, and satisfaction. In line with this assumption we show that our cognitive-

load classification is significantly correlated with negative emotions such as stress and frustration.

2 Related Work

2.1 Adaptations based on cognitive load estimation

Cognitive load estimation usually is performed in a task and participant-specific way and has also been demonstrated in this context to allow for useful workload adaptations in learning environments or vehicel control tasks. Yuksel and colleagues created a brain-computer interface that adapted the difficulty of a musical learning task [26]. They measured cognitive load using fNIRS to decide when to increase difficulty. Their approach managed to significantly increase learning gains during piano lessons compared to a control group. However, the classifiers they used were participant-specific and were trained using a long training period consisting of 30 songs per participant.

Moreover, an aviation simulation was used by Wilson and Russel to provide real-time adaptive feedback [8]. They used a combination of EEG, respiration, and heart rate, but also eye-fixation behavior to realize adaptations during an uninhabited air vehicle task. Participant-specific artificial neural networks were trained to detect high cognitive load and adapt the task by slowing down simulated time when cognitive load was too high. In contrast to our approach, real-time adaptation was successfully realized only with participant- and task-specific classifiers.

Furthermore, Kelleher et al. developed a method that does not rely on EEG data, but rather on users' behavioural performance [27]. Their approach was able to distinguish between a hard puzzle and an easier one based on users' performance on the previous puzzles with an accuracy of 71% to 79%. A wide array of features derived from performance, user input, and user ratings was used to train random forests and predictions were made based on the last three puzzles the user was attempting to solve. While the results are promising, their method is still specific to their task and individual participants.

2.2 Cross-participant and cross-task approaches

While cognitive load estimation usually is performed in a task and participant-specific way, there are several studies that successfully implemented either cross-task or cross-participant approaches (but not both). In contrast to the method that we present in this article, most of them rely, at least partly, on EEG.

A very detailed assessment of mental workload is provided by Popovic et al. [28]. They classified different kinds of cognitive load (i.e., speech, fine motor, gross motor, auditory, visual and cognitive) using EEG and ECG. Their cross-participant classifier achieved 72.5% accuracy for cognitive load in a leave-one-participant-out cross-validation.

Another interesting approach was presented by Ke and colleagues [29]. They generalize from individual regression models to more general ones by applying a feature selection algorithm to EEG data recorded from a working memory task and a complex simulated multi-attribute task designed to evaluate operator performance and workload (see [30]). In a first step, they used two thirds of their data to systematically eliminate features with low cross-task correlations and then evaluated their feature set on the remaining validation set. They found a significant increase in performance of their regression model. Again, these results show cross-task- but not cross-participant-generalization. This makes them applicable in some situations, but still not as general as a many applications would demand.

Finally, Appel and colleagues successfully developed a machine learning approach for cross-participant classification of cognitive load [31] using eye-tracking data. For a working memory task they achieved an accuracy of 76.8% for offline classification and 70.4% for real-time online classification. A reworked version of their approach was used in a emergency simulation and showed promising results under noisy conditions similar to actual applications [32]. This updated version worked across different versions of the simulation, showing potential for a cross-participant and cross-task solution.

3 Experimental setup

We collected eye-tracking data from two different tasks: (1) an N-back task (a standardized working memory task inducing a controlled level of cognitive load), and (2) a computer simulation, that represented a real-life application. We aimed to use data from the first to estimate cognitive load in the latter.

3.1 N-back task

The N-back task [33] is commonly used to induce cognitive load and to measure working memory capacity. Participants are presented with a randomly generated sequence of letters and have to press one of two buttons to indicate whether the currently presented letter is the same as N letters before. N modulates the difficulty of the task, because a larger N means that more letters have to be held in memory and compared to the actual one presented. With regard to working-memory demands, the N-back task requires to keep a string of N letters active in memory, compare the first letter of the string to the current trial, decide on the correct button, and update the memorized string by deleting the first letter of the string and adding the letter of the current trial to it. 0-back can be used as a control condition where participants have to compare the current stimulus with a constant that was presented at the very beginning. Letters are randomly chosen from the set $L = \{C, F, H, S\}$ and are presented for 0.5 seconds, followed by a black screen shown for 1.5 seconds. A schematic overview is provided in Figure 1.

Participants first received instructions for the task and had to perform a short training until they achieved an accuracy of 60%. They then completed two critical blocks, each comprising three difficulty levels: 0-back, 1-back, and 2-back. Each level of a block consisted of 154 trials; the order of levels in a block was randomized.

We used the "N" of the N-back task as an experimental manipulation of cognitive load (within participants design) and focused on the difference between 0-back and 2-back conditions.

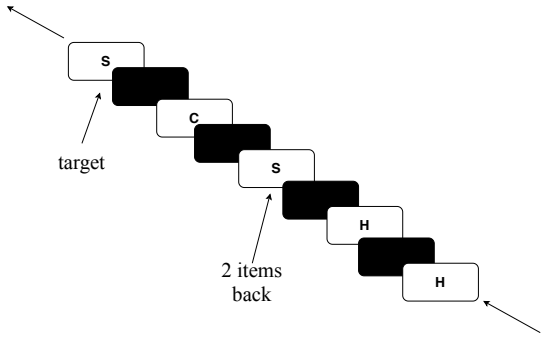


Fig. 1. Overview of the N-back task as illustrated for $N = 2$

3.1.1 Participants

28 students (*mean age* = 24.71, *SD* = 4.12, 14 females) from the University of Tübingen were recruited for the N-back task. Data of one participant was discarded due to problems with the eye-tracking recordings resulting in too little usable data.

The experiment was approved by the local ethics committee and all participants gave written informed consent at the beginning of the experiment. Participants received monetary compensation at the end of the experiment. All were right-handed and German native speakers.

3.1.2 Apparatus

We used a RED250 eye tracker from SensoMotoric Instruments (SMI) in combination with the SMI Experiment Center software (version 2.7.13) for the recording of eye movements and pupil-related features. Calibration was performed with SMI's built-in 9-point calibration. All eye-tracking data were recorded at 250Hz in a laboratory setting with illumination held constant in individual sessions.

During the task a chin-rest was used to ensure stable head position and constant viewing distance. Stimuli were presented on a 22-inch monitor with a resolution of 1,680 $\times$ 1,050 px using Arial font with a size of 25. All letters were presented in gray on a black background.

3.2 *Emergency* simulation task

The *Emergency* simulation task was based on the commercially available simulation *Emergency* by Promotion Software GmbH [34]. Participants had to coordinate emergency personnel consisting of firefighters, paramedics, and ambulances involved in responding to different scenarios (e.g., a car crash or burning buildings). The simulation can be used as a training tool for emergency management tasks as well as for entertainment purposes and thus covers aspects of digital environments for learning, working, and gaming.

The simulation started with a tutorial that introduced participants to the handling of the simulation. After the tutorial was completed successfully, three scenarios were presented: A car crash, burning buildings, and a train crash. Each scenario had three levels of increasing difficulty (easy, medium, and hard version of the scenario). Scenarios had to be completed in the same order of easy to difficult levels and scenarios (i.e., from car crash to train crash) by all participants. Scenarios and difficulty levels differed in the number of sub-tasks to be completed as well as the available numbers of emergency personnel and their composition. These manipulations were calibrated by Promotion Software GmhH for the purposes of this study in order to optimally manipulate the levels of cognitive load imposed onto participants. Descriptive statistics can be found in Table 1.

The fixed order of scenarios and difficulty was chosen deliberately for this task. While a randomized order would be ideal to avoid confounds, it is hard to implement in a task that involves learning and skill acquisition. Participants who complete easy parts of the simulation first gain proficiency fast and scenarios that may have been difficult for them in the beginning become more and more easy. On the other hand, when participants are confronted with very difficult scenarios first, they may be overwhelmed which hinders or even prevents learning. This expertise change over time and its dependency on the order of task presentation necessitated a fixed order. Gerjets et al. also recommend a fixed order from simple to complex for learning tasks for this exact reasons [4]. Moreover, an ascending order of difficulty is in line with the way learning materials and games are commonly structured, making it a better showcase for the application of our method. A further important aspect is that we aim to use data from the N-back task to estimate cognitive load in the *Emergency* simulation so that the training data are not affected by a confounding of difficulty level and time.

In the simulation certain sub-tasks could only be performed by specific emergency personnel and with varying degrees of efficiency. Therefore, especially in scenarios that did involve fire, planning activities were essential for successful completion of a mission. For instance, as fire can spread to nearby buildings and also hurt emergency personnel, prioritisation of which fires to put out first was crucial. Putting out fires could be done by firetrucks but also by firefighters alone, however, with firefighters being considerably slower at performing the task. Moreover, firefighters might be required for cutting trapped victims free. In general, more difficult levels involved more emergency personnel units to be coordinated and more sub-tasks, posing higher demands on planning, prioritisation, monitoring, and information updating.

After each level, participants were asked to indicate their subjective cognitive load and also their affective experiences based on a modified version of the NASA-TLX questionnaire [11]. The questionnaire contained scales for positive and negative emotions, mental and temporal demand, effort, frustration, and stress, as well as a measure of seriousness that were each rated on a scale of 0 to 100. Participants' responses were used to evaluate the validity of our approach and the relation of our envisioned cognitive-load classification to affective experiences.

For this task, we used the difficulty levels within each scenario as manipulation of cognitive load (within participants design) and focused on the difference between the easy and hard version of each scenario.

Fig. 2. This image is taken from the third scenario "train crash" and shows a typical situation in *Emergency*. Paramedics and ambulances can be seen at the bottom of the screen, as well as the firefighters' trucks.

Scenario	difficulty	units	sub-tasks	fire	time	completed	cog. demand	temp. demand	effort
scenario 1	easy	6	11	no	300s	83.33%	35.85	32.43	33.29
scenario 1	hard	12	20	no	300s	83.33%	42.14	40.29	38.14
scenario 2	easy	9	18	yes	450s	97.22%	39.00	23.14	32.57
scenario 2	hard	15	26	yes	450s	36.11%	51.14	56.57	52.71
scenario 3	easy	10	24	yes	600s	97.22%	43.43	31.57	39.43
scenario 3	hard	16	44	yes	600s	33.33%	59.00	64.57	61.86

TABLE 1
Descriptive statistics of the three scenarios of the Emergency simulation game.

3.2.1 Participants

The Emergency simulation was completed by 47 participants (*mean age* = 24.6, *SD* = 6.3, 33 females). There was no overlap between the participants of the *Emergency* simulation and the participants of the N-back task. Seven participants had to be excluded due to problems with eye-tracking recordings. Another 2 were excluded because they reported that they did not take the experiment seriously. Finally, 2 participants had a very high number of missing data and consequently did not provide enough usable data for all scenarios, rendering their data partly unusable. The data of the remaining 36 participants were included in the further analyses.

We deliberately included participants with noisy data or poor tracking ratios (i.e., time spans with invalid data caused by the pupil not being detected reliably). This renders the data more realistic with closer resemblance to data one would expect in an online-scenario of a real-world application.

The experiment was approved by the local ethics committee and all participants gave written informed consent at the beginning of the experiment. All participants were right-handed, German native speakers and received monetary compensation at the end of the experiment.

3.2.2 Apparatus

The eye-tracking setup was the same as for the N-back task, featuring a RED250 eye tracker from SensoMotoric Instruments (SMI) in combination with the SMI Experiment Center software (3.7.60). Calibration was performed with SMI's built-in 9-point calibration and the recording frequency was set to 250Hz. Data recording was performed in a laboratory setting with illumination held constant in individual sessions.

For the *Emergency* simulation a laptop with a 16-inch screen driven at 1920 x 1080 px resolution was used. This task did not involve a chin-rest as to closer mimic a real-world learning or gaming situation.

4 Features used for Classification

Eye-fixation behavior is strongly influenced by presented stimuli as their structure and appearance guide the users attention (e.g., Rayner, for a review [35], [36]). Therefore, our

approach relies on eye-related features that were chosen because they are either independent of the stimulus structure or only marginally depended on it. More specifically, we did not rely on saccades, areas of interest, or the coordinates of fixations.

The feature extraction process for a chosen share of data always followed the same procedure. First we extracted 7 features which will be described later in this section and then normalized them using a participant-specific baseline to allow for cross-participant comparisons. Baseline in this context refers to the features of a specific part of the data. For the N-back task this baseline was taken from the instruction phase, while for the Emergency simulation we used the tutorial phase as baseline.

Normalization was performed at the participant level and involved subtracting the baseline from the segment's features and dividing by it. As a consequence, all features that we used reflected relative changes from the individual participant's baseline.

Detection of fixations, saccades, and blinks used SMI's built-in event detection. For fixations this is a dispersion-based algorithm with a maximum dispersion of $2 - 3°$ (depending on the distance between screen and user) and a minimum fixation duration of 80ms. Blinks are defined via the gaze and pupil signal. Gaze coordinates of $(0, 0)$ or the pupil being zero or outside a dynamic computed validity range is interpreted as a blink. Blinks of less than 70ms are discarded. SMI's default algorithm interprets anything that is between between two fixations or a blink and a fixation as a saccade.

4.1 Pupil-related features

Pupil diameter has been used to measure cognitive load for several decades. An increase in cognitive load leads to decreased parasympathetic activity in the peripheral nervous system, which in turn leads to an increase in pupil diameter [37]. This effect was observed consistently within a task, between tasks, and between individuals [38]. Various studies have successfully replicated this relationship within a wide range of settings, including short-term memory, language processing, reasoning, perception, as well as sustained and selective attention [18]. Pupil diameter has also successfully been used to detect cognitive load in a variety of scenarios, including driving [39], during low visual load tasks [40], route planning with maps [41], and simultaneous interpreting [42]. Furthermore, it was successfully used to differentiate expertise closely related to cognitive load [43]

We applied preprocessing steps to improve data quality of the pupil signal. First, we removed periods that were marked as blinks, as well as the 100ms right before and after a blink. During these phases, the pupil could not be detected reliably and as a consequence measurements of pupil diameter would suffer from reduced accuracy. We furthermore removed implausible pupil values (e.g., values of 0mm or less, as well as values greater than 10mm). Finally, we linearly interpolated small gaps of less than 50ms (12 data points at sampling rate 250Hz) and applied a median filter to reduce noise.

We selected the median of the pupil diameter as the main pupil feature, as it is more robust to outliers than the mean, in particular for short sampling periods. Moreover, we utilized the maximum pupil diameter as a feature to capture spikes in the pupil signal. We expected to see an increase in both median and maximum pupil diameter with increasing cognitive load.

Moreover, we employed the Index of Cognitive Activity (ICA) as proposed by Marshall [44], [45]. It reflects rapid spikes in the pupil signal that are caused by cognitive load. For this part of the analysis, we did not apply the interpolation and filter preprocessing as it may remove those spikes needed for the ICA. A higher degree of cognitive load was supposed to result in increased ICA.

As an additional, more exploratory feature, we included the standard deviation of the pupil diameter. According to the ICA, cognitive load can cause fluctuations and rapid spikes in pupil diameter. Based on this assumption, we expected a higher standard deviation of pupil diameter for higher cognitive load.

4.2 Blinks

Cognitive load influences the frequency and duration of blinks [46], [47]. Thus, increasing task difficulty was expected to increase the frequency of blinks, while increasing visual demands should lower the amount of blinks [48]. We used blink frequency as a feature and expected it to increase with cognitive load in both the N-back task and the *Emergency* simulation.

4.3 Fixations

Fixations describe a stable gaze on the same location usually lasting between 200 ms and 350 ms [35]. Frequency of fixations is influenced by several factors. Time pressure induced by high task demands tends to increase the number of fixations while reducing their duration [49]. We expected to observe the same pattern for higher levels of cognitive load in our study. Consequently, we used the number of fixations per second as a feature.

4.4 Microsaccades

Microsaccades are small involuntary eye movements that may occur during a fixation and are associated with cognitive load. Studies reported an increase in microsaccade frequency in visually demanding tasks [50], whereas non-visual tasks (e.g, auditory tasks or mental arithmetic) seemed to reduce their frequency [51], [52], [53].

We used the method suggested by Krejtz and colleagues [53] to detect microsaccades, which relies on thresholds to find small ballistic sequences in an otherwise fixed gaze. Instead of focusing on amplitude or velocity we use microsaccade frequency as a feature, as the former two would require a higher sampling rate than 250Hz to be reliable. Because both tasks involved visual presentation, we expected an increase in microsaccade frequency with rising cognitive load.

5 Cognitive Load Detection Method

The core of our approach was strongly inspired by Appel et al. [31], [32]. The fundamental idea was to train participant-specific classifiers for low and high cognitive load based on

data from a N-back task and use their weighted predictions on the Emergency simulation. Participants that were similar during baseline periods are weighted stronger as we expect their physiology to change under cognitive load in a similar way.

5.1 Within-task and within-participant classification

Participant-specific classifiers were trained on N-back data. As the N-back is a standard working-memory updating task that is recorded under laboratory conditions, we expected it to reflect characteristic physiological changes caused by cognitive load and to allow for generalization from this task to the *Emergency* simulation as described in Section 5.2.

To train a classifier that can differentiate between high and low cognitive load, we needed data from periods of high cognitive load and periods of low cognitive load during the training phase. We used the N-back task as foundation for single-participant classifiers and considered the 0-back condition to reflect low cognitive load and the 2-back condition to represent high cognitive load. 25 non-overlapping samples with a length of 4s each were randomly selected from both conditions, yielding 50 samples per participants that were used for training the individual classifier. We rejected samples with more than 50% missing values in the pupil signal and resampled to ensure that each sample contained enough information to be useful for training. These numbers represented a balanced compromise between sample size and sample length.

For each of the samples we extracted the features described in Section 4. All features were then z-transformed using individual means and SDs for standardization. This scaling improved inter-participant comparability considerably and should thus help applying classifiers across participants.

Finally, we trained a forest of 1000 extremely randomized trees (Extra-Trees) [54] per participant to distinguish between high and low cognitive load based on that individual participants' samples. Extra-Trees had the advantage of providing not just a decision into classes, but also class probabilities between 0 and 1. This enabled us to form a continuous scale instead of a dichotomous decision, adding further information. The output was a number between 0, when the classifier was absolutely certain that a sample was collected under low cognitive load, and 1 in case of high cognitive load. In addition, Extra-Trees tended to not overfit as fast as other classification methods allowing more features in conjunction with less samples. Moreover, Extra trees seemed appropriate for the goal of real-time classification of cognitive-load levels as they can be trained and evaluated fast. The use of 1000 trees was empirically determined, even though fewer trees may work as well. In case computation time is an issue, less trees may be chosen.

We used the Extra-Tree implementation provided by the Python toolbox scikit-learn [55].

5.2 Cross-participant and cross-task approach

In a next step, we combined the single-participant classifiers trained with data from the N-back task to form a composite classifier that can be applied across participants and tasks. The fundamental idea of our approach was to apply the classifiers trained on N-back data to the Emergency simulation, but to weigh their contribution to the final prediction according to how similar their baselines were. In this way, participants from the N-back taks that were more similar regarding their physiological features and behavioral parameters to a participant from the *Emergency* task were given higher weights in the final prediction. Adding this weighing substantially increased the accuracy of the combined classifier.

To verify cross-task capability, sample data from the *Emergency* task was needed. Therefor, we randomly sampled 25 segments of length 4s from the easy and hard version of each scenario of *Emergency*, resulting in 50 samples per scenario and participant. Again, these numbers represented a compromise between sample length and sample number. From these samples we extracted the features in the same way as we did with the N-back data including the normalization using the baseline and z-transformation. Segments extracted from the easy version of a scenario represented low cognitive load, while segments from the hard version represented high cognitive load.

For baseline comparison, features were normalized across all participants to have a mean of 0 and standard deviation of 1 as to not inflate the importance of features that are on a larger scale. There were, for instance, a lot fewer blinks within one second than there were fixations and the pupil diameter in millimeters was a lot larger compared to the number of microsaccades per second.

The procedure can be described as follows: Let p be a participant of *Emergency* whose cognitive load we want to classify, s_p a sample of p characterized by a set of features, and P the set of participants of the N-back task. Every $p \in P$ has a classifier c_p that predicts a value between 0 and 1 for $sample_p$. We combined these predictions according to the following equations:

$$sim(p, other) = \frac{1}{\sum acc_{c_{other}} w_{c_{other}} |base_p - base_{other}|}$$

$$pred(s) = \frac{\sum_{other \in P_n} sim(x, other) pred_{c_{other}}(sample_x)}{\sum_{other \in P_n} sim(x, other)}$$

$sim(p, other)$ refers to the baseline similarity between participants p and $other$, $w_{c_{other}}$ to the normalized feature weights of classifier c_{other}, $acc_{c_{other}}$ to the cross-validated accuracy that classifier c_{other} achieved on participant $other$, and $pred_{other}$ to the prediction of classifier c_{other}. This means that we drew a prediction for cognitive load from each classifier c_{other} and weighted these predictions according to how similar the baselines of participants p and $other$ are. Additionally, we factored the feature weights of classifier c_{other} into the similarity, giving a higher weight to more important features, and also taking into account how well the classifier performed on its specific participant.

Dividing by the sum of all similarities normalized these similarities and ensured that the prediction's final result was within the interval of $[0, 1]$. P_n refers to a subset of P, that is restricted to the n participants with the highest similarity. The choice of a smaller n can help to reduce computational costs in case there are a lot of classifiers available from the

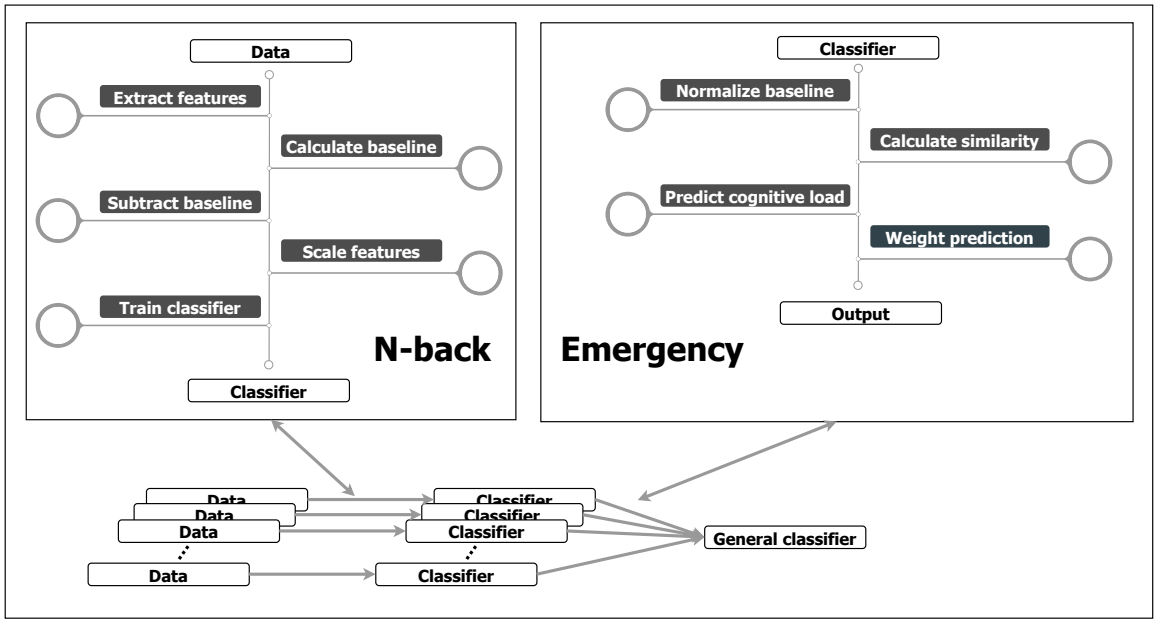


Fig. 3. Overview of our method, showing how we train individual classifiers and apply them to new data.

N-back task. We employed $n = 5$ to highlight that it does not take a lot of participants to get accurate results.

Figure 3 shows a schematic overview of our method and Algorithm 1 presents pseudo-code of our cross-participant and cross-task classification. Both serve to illustrate our approach.

6 Results

6.1 Within-task

As a frame of reference, we did not only analyze results for cross-task and cross-participant classification but also for within-task and within-participant classification. All results reported for within-participant classification were obtained based on a 10-fold cross-validation to avoid overfiting a classifier to a specific participant and thereby artificially inflating classification accuracy.

Within-task accuracy for the *Emergency* task is reported for each scenario individually and is based on random samples with a length of 4 seconds that were extracted from the easy and hard version, respectively. Feature extraction was performed in the same way as described in detail for the N-back samples.

In case a participant reported that the easy version of a scenario was experienced to be more difficult than the hard version, that scenario was excluded from the results of that participant. "More difficult" refers to the average rating of cognitive demands, temporal demands, and effort as reported in the NASA TLX. For 10 participants, the first scenario had to be excluded for this reason and for 2 the second scenario. In the easy version of the first scenario, participants had their first real interaction with the simulation and it is therefore likely that they experienced it as more difficult than the hard version, because by that time they already were familiar with the simulation.

Table 2 shows the detailed accuracy scores for the N-back task and emergency task, respectively.

	within-participant	cross-participant
N-back	79.55%	75.81%
Emergency, Scenario 1	71.91%	71.41%
Emergency, Scenario 2	71.98%	69.34%
Emergency, Scenario 3	68.91%	67.16%

TABLE 2
Detailed accuracy of within-task classification.

It is notable that the results for the N-back task were slightly better than those obtained for the Emergency task. Partly, this may be because of the experimental setup. The N-back task was recorded with participants using a chin-rest, which helped to improve quality of the eye-tracking data in general and reliability of the pupil measurements in particular. Furthermore, difficulty remained constant over the course of one level, whereas situational difficulty varied during levels of the emergency simulation. The fact that we took random samples from easy and difficult versions of the simulation may thus have led to samples not reflecting the exact same degree of cognitive load, even within one participant. This, in turn, added variance to the features and made the labels "easy" and "difficult" less distinct for *Emergency* than for the N-back.

Comparing the drop in accuracy caused by the shift from within-participant to cross-participant classification, one can see that the drop was more pronounced for the N-back task. This was likely due to the fact that we had a larger number of participants for the Emergency simulation, meaning that it was more likely to find good matches during the baseline comparison.

Algorithm 1 Pseudocode outlining our method for cognitive load detection

$P \leftarrow$ set of all participants in N-back task
$d_{p,i} \leftarrow$ data of participant p taken from the ith scenario of Emergency
$base_P \leftarrow$ normalized baseline of participant p
$c_P \leftarrow$ N-back-trained classifier of participant p
$n \leftarrow$ number of neighbours to consider
for $other \in P$ **do** ▷ calculate distances between participants' baselines and make predictions
  $acc \leftarrow accuracy(c_{other})$
  $w \leftarrow featureweights(c_other)$
  $w \leftarrow \frac{w}{\sum w}$
  $dist(p, other) \leftarrow \sum acc\ w|base_p - base_{other}|$
  for $i \in \{1, 2, 3\}$ **do**
    $prediction_{other,p,i} \leftarrow c_{other}.predict(d_{p,i})$ ▷ prediction for each sample of $d_{p,i}$
  end for
  $dist \leftarrow \frac{dist}{\sum dist}$ ▷ normalize distances to sum to 1
  $sim \leftarrow \frac{1}{dist}$ ▷ get similarity from the distance
  $P_n \longleftarrow \{y \in P | y\ amongst\ n\ most\ similar\}$ ▷ n participants with highest similarity to p
  for $i \in \{1, 2, 3\}$ **do**
    for $sample \in d_{p,i}$ **do**
      $out_p[i, sample] \longleftarrow \frac{\sum_{other \in P_n} sim(x,other) pred_{c_{other}}(sample)}{\sum_{other \in P_n} sim(p,other)}$
    end for
  end for
end for

6.2 Cross-task

Cross-task and cross-participant results were obtained by applying classifiers trained on N-back data to samples from the *Emergency* simulation following the approach described in Section 5.2. Classification accuracy is summarized in Table 3 and ranged between 63.78% and 69.25%.

Scenario	accuracy
Scenario 1	69.25%
Scenario 2	63.78%
Scenario 3	64.02%

TABLE 3
Accuracy of classifiers trained with N-back data and applied to data gathered during the emergency simulation.

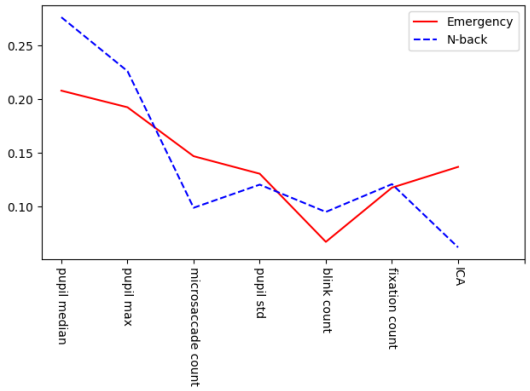


Fig. 4. Average weights for Emergency and N-back tasks

As expected, applying N-back classifiers to Emergency data led to a slight drop in classification accuracy as it represented a classification across different participants and tasks. Using classifiers from "wrong" participants introduced a certain error as the classifier did not match the participant. The same holds true for the application across tasks. Moreover, as Figure 4 shows, feature weights also differed between the two tasks introducing yet another source of error.

The main difference in feature importance was observed for ICA and microsaccades. Both carried considerably more importance in the Emergency simulation than they did for the N-back task. This was in line with results of Fairclough and colleagues who found the ICA to not be significantly sensitive to isolated working memory tasks like the N-back task [56]. A possible explanation for the difference in microsaccade importance may be the different nature of the task. Emergency was a lot more visually demanding and required more widely distributed attention. This fits with findings from Duchowski and colleagues [57] that ambient visual search increases the number of microsaccades. The importance of the remaining features was slightly higher for the N-back task because ICA and microsaccades were less important and the importance of all features sums up to 1.

For scenario 1 classification accuracy decreased from 71.91% from within-participant, within-task application to 69.25% in cross-participant, cross-task classification. A possible explanation for this good performance may be that participants did not yet have experience with the task (i.e., they all started at the same point). This "neutral" conditions with regards to the experience and skills acquired may be similar to the N-back task, thus leading to less loss caused by cross-task application of classifiers.

Scenario 2 showed the most pronounced decrease, from 71.98 to 63.78%. The major error source was the cross-task application, as cross-participant results differed only slightly

from the within-participant ones for Emergency. It seems likely that the structure of this specific scenario lead to a feature distribution that differed the most from the N-back task's features, resulting in this decrease in accuracy.

An accuracy loss of less than 5 percentage points – from 68.91% to 64.02% – could be observed for scenario 3. These results were very similar to the second scenario and likely have a similar cause: cross-task application.

To verify our prediction not only on a binary level, we considered participants' questionnaire data and their correlation with our predicted continuous cognitive load. Table 4 shows Pearson correlations between cognitive load predicted by our algorithm and self-report scores. Self-reports were normalized on participant level to account for individual differences of scale. As a frame of reference, we furthermore included correlations between the questionnaire's different sub- scales.

Predictions made by our algorithm showed a significant correlation with self-reports. They correlated at 0.399 with self-reported cognitive demands, at 0.459 with reported temporal demands, and at 0.484 with the effort subjectively experienced by participants. The high correlation with perceived effort is a strong indicator for the validity of our predictions.

7 Discussion

We applied a machine learning approach to the classification of cognitive load based on eye-tracking data and investigated how this approach generalizes across participants and tasks. Our results indicate a robust approach that yields good classification accuracy of 63.78% to 69.25% across participants and tasks. This is significantly above chance level and is comparable to eye-tracking based classification results for cognitive load in other scenarios. For instance, Hogervost et al. [58] reported roughly 68% accuracy in the distinction between level 0 and level 2 for the N-back task purely based on eye-related features. However, their classification algorithm was trained for each participant individually, within a specific task, and using intervals that were 50s long – all limitations that our approach does not have. Additionally, cognitive load predictions yielded by our method correlate at $r = 0.484$ with participants' self-reported invested effort, which provides an indicator of validity on a second level.

When taking a closer look at our data set, the robustness of our approach seems noteworthy. We considered noisy data in our analyses and, in the case of the *Emergency* game, used the tutorial as an active baseline instead of a neutral fixation cross. Furthermore, *Emergency* is not a well-controlled laboratory task, but a complex emergency simulation game, which requires participants to identify what to do with the right emergency personnel under time constraints in an environment that adaptively reacts to players actions (e.g., spreading fires when not extinguished by fire fighters). As such, the present results seem promising as they indicate the validity of our approach even when applied to a real-world scenario with limited baseline options and complex interactions.

Moreover, due to the dynamic nature of the *Emergency* simulation, cognitive load was not constant over the course of one level. Closer inspection of the predictions generated by our algorithm revealed that participants seem to start each level with rather high predicted load which quickly dropped after a first orientation phase of about 20-30s. Towards the end there was also a clear difference between participants that successfully finished a level and those who did not. When participants realized they will finish on time, predicted cognitive load dropped considerably, whereas it rose when they became aware that they could not finish on time. This uneven distribution of cognitive load adds to the error rates that we report. Therefore, our predictions may actually be even more accurate then what is reported, because we had to rely on the overall task difficulty of a specific level as an indicator of cognitive load instead of a more direct measure (e.g., derived from interaction metrics). Generalizing a difficulty level by labeling all samples from this level as "high cognitive load" even though there were probably periods of lower cognitive load within the same time-frame possibly introduced a kind of artificial error.

Additionally, the nature of our features and method are very versatile. All features we used are either aggregated over the whole length of the segment or are calculated per second. As a result, length of segments can be adjusted at will. Longer segments are less noisy, but shorter segments better capture the cognitive load at a certain point in time. Pre-trained classifiers may be applied independent of segment length, making our approach more flexible. The same holds true for the number of classifiers that are used. When computation time is a constraint, less classifiers may be used for prediction, as n – the number of closest classifies during baseline comparison – can be adjusted at will.

7.1 Limitations

There are, however, also some limitations to our approach. First, it only works reliably when features we are z-transformed the features onat the participant level. This means that we can only reliably analyze data in hindsight and when there are periods of low and high cognitive load. One may nevertheless use the presented approach in real-time scenarios, but then it has to be considered that workload predictions might not be reliable in the very beginning might not be reliable, but will improve over time as more data becomes available and more variation in cognitive load is observed.

Moreover, one of the reasons why our approach works successfully may also be considered a drawback, namely: baseline comparisons. When the baseline for two tasks is recorded under different conditions problems might arise. For instance, cognitive load may be different in a baseline obtained while looking at a fixation cross as compared to a baseline extracted from completing a tutorial. Using the suggested process of matching participants for cross-participant and cross-task classification, this may lead to a sub-optimal distance metric and consequently an inappropriate weighing of predictions. Ideally, all baselines should evoke the same degree of cognitive load for baseline distances to operate best.

Furthermore, as our analysis of the feature weights showed, a complex simulation game like such as Emergency does not evoke the exact same physiological responses as a

measure	prediction	pos. emotions	neg. emotions	cog. demand	temp. demand	effort	frustration	stress	seriousness
prediction	1.000								
pos. emotions	-0.111	1.000							
neg. emotions	0.186**	0.037***	1.000						
cog. demand	0.399***	-0.134***	0.212***	1.000					
temp. demand	0.459***	-0.175***	0.285***	0.633***	1.000				
effort	0.484***	-0.203***	0.283***	0.701***	0.758***	1.000			
frustration	0.294***	-0.338***	0.443***	0.474***	0.546***	0.607***	1.000		
stress	0.404***	-0.184***	0.309***	0.551***	0.622***	0.657***	0.573***	1.000	
seriousness	0.012	0.175***	-0.112**	-0.009	-0.031	-0.023	-0.058	-0.016	1.000

TABLE 4

*Significant at $p \leq 0.05$, **Significant at $p \leq 0.01$, ***Significant at $p \leq 0.001$

laboratory working- memory task like such as the N-back task. Our results indicate that in particular the importance of the ICA and of microsaccades, in particular, seems to depend on the task at hand. This indicates that – although successful cross-task is possible - there may not be a classifier that works optimally for all tasks. A possible solution for this problem might could be to have classifiers trained on data of different tasks and also add these tasks to the baseline comparison. This way, classifiers trained on tasks that are similar in nature will be preferred.

Our study relied on visual stimuli and did not include other modalities like such as auditory tasksstimuli. However, at least for working memory tasks important features seem to react alike andbe useful independently of presentation modality. For instance, Kahneman demonstrated extensively and with many different stimuli, tasks, and modalities showed thean effect of working memory load on pupil diameter [38]. and Rrecent research indicates yielded similar findings also for microsaccades [51] - at least for auditory and visual stimuli.

8 Conclusion

In summary, we evaluated a cross-participant as well as cross-task classification algorithm that yields good accuracy. Combined with the robustness of our method and its non-invasive nature this article - despite its limitations - provides a promising step towards out-of-the-box solutions for adaptive human-computer interaction based on the assessment and classification of users' cognitive load by means of eye-tracking data.

Acknowledgments

This research was funded by the LEAD Graduate School and Research Network [GSC1028], a project of the Excellence Initiative of the German federal and state governments. Tobias Appel and Franz Wortha were doctoral students of the LEAD Graduate School and Research Network.

References

[1] M. S. Hussain, R. A. Calvo, and F. Chen, "Automatic Cognitive Load Detection from Face, Physiology, Task Performance and Fusion During Affective Interference," *Interacting with Computers*, vol. 26, no. 3, pp. 256–268, 06 2013. [Online]. Available: https://doi.org/10.1093/iwc/iwt032

[2] J. M. Rose, F. D. Roberts, and A. M. Rose, "Affective responses to financial data and multimedia: The effects of information load and cognitive load," *International Journal of Accounting Information Systems*, vol. 5, no. 1, pp. 5–24, 2004.

[3] K. Gasper and J. Hackenbracht, "Too busy to feel neutral: Reducing cognitive resources attenuates neutral affective states," *Motivation and Emotion*, vol. 39, no. 3, pp. 458–466, 2015.

[4] P. Gerjets, C. Walter, W. Rosenstiel, and M. Bogdan, "Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain-computer interface approach," 01 2014.

[5] S. Chaiklin, "The zone of proximal development in vygotsky's analysis of learning and instruction," *Vygotsky's educational theory in cultural context*, vol. 1, pp. 39–64, 2003.

[6] L. Zhang, J. Wade, D. Bian, J. Fan, A. Swanson, A. Weitlauf, Z. Warren, and N. Sarkar, "Cognitive load measurement in a virtual reality-based driving system for autism intervention," *IEEE Transactions on Affective Computing*, vol. 8, no. 02, pp. 176–189, apr 2017.

[7] A. Yuce, H. Gao, G. L. Cuendet, and J. Thiran, "Action units and their cross-correlations for prediction of cognitive load during driving," *IEEE Transactions on Affective Computing*, vol. 8, no. 02, pp. 161–175, apr 2017.

[8] G. F. Wilson and C. A. Russell, "Performance enhancement in an uninhabited air vehicle task using psychophysiologically determined adaptive aiding," *Human Factors*, vol. 49, no. 6, pp. 1005–1018, 2007.

[9] M. Csikszentmihalyi, S. Abuhamdeh, J. Nakamura, *et al.*, "Flow," 1990.

[10] K. Kiili, A. Lindstedt, and M. Ninaus, "Exploring characteristics of students' emotions, flow and motivation in a math game competition," 01 2018.

[11] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Advances in Psychology*. Elsevier, 1988, vol. 52, pp. 139–183.

[12] R. D. McKendrick and E. Cherry, "A deeper look at the nasa tlx and where it falls short," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 62, no. 1, pp. 44–48, 2018. [Online]. Available: https://doi.org/10.1177/1541931218621010

[13] P. Antonenko, F. Paas, R. Grabner, and T. Van Gog, "Using electroencephalography to measure cognitive load," *Educational Psychology Review*, vol. 22, no. 4, pp. 425–438, 2010.

[14] C. S. Ikehara and M. E. Crosby, "Assessing cognitive load with physiological sensors," in *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, Jan 2005, pp. 295a–295a.

[15] A. Jimenez-Molina, C. Retamal, and H. Lira, "Using psychophysiological sensors to assess mental workload during web browsing," *Sensors*, vol. 18, no. 2, p. 458, 2018.

[16] A. M. Beres, "Time is of the essence: A review of electroencephalography (eeg) and event-related brain potentials (erps) in language research," *Applied psychophysiology and biofeedback*, vol. 42, no. 4, pp. 247–255, 2017.

[17] L. M. Naismith and R. B. Cavalcanti, "Validity of cognitive load measures in simulation-based training: a systematic review," *Academic Medicine*, vol. 90, no. 11, pp. S24–S35, 2015.

[18] J. Beatty, "Task-evoked pupillary responses, processing load, and the structure of processing resources." *Psychological Bulletin*, vol. 91, no. 2, p. 276, 1982.

[19] V. Clay, P. Koenig, and S. Koenig, "Eye tracking in virtual reality," *Journal of Eye Movement Research*, vol. 12, 04 2019.

[20] E. Bozkir, D. Geisler, and E. Kasneci, "Person independent, privacy preserving, and real time assessment of cognitive load using eye tracking in a virtual reality setup," in *The IEEE Conference on Virtual Reality and 3D User Interfaces (VR) Workshops*, mar 2019.

[21] J. L. Lobo, J. D. Ser, F. De Simone, R. Presta, S. Collina, and Z. Moravek, "Cognitive workload classification using eye-tracking

and eeg data," in *Proceedings of the International Conference on Human-Computer Interaction in Aerospace*. ACM, 2016, p. 16.

[22] J. Heard, C. E. Harriott, and J. A. Adams, "A survey of workload assessment algorithms," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 5, pp. 434–451, Oct 2018.

[23] C. L. Baldwin and B. Penaranda, "Adaptive training using an artificial neural network and eeg metrics for within- and cross-task workload classification," *NeuroImage*, vol. 59, no. 1, pp. 48 – 56, 2012, neuroergonomics: The human brain in action and at work. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S105381191100824X

[24] C. Walter, S. Schmidt, W. Rosenstiel, P. Gerjets, and M. Bogdan, "Using cross-task classification for classifying workload levels in complex learning tasks," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, Sep. 2013, pp. 876–881.

[25] M. Spüler, T. Krumpe, C. Walter, C. Scharinger, W. Rosenstiel, and P. Gerjets, "Brain-computer interfaces for educational applications," in *Informational Environments*. Springer, 2017, pp. 177–201.

[26] B. F. Yuksel, K. B. Oleson, L. Harrison, E. M. Peck, D. Afergan, R. Chang, and R. J. Jacob, "Learn piano with bach: An adaptive learning interface that adjusts task difficulty based on brain state," in *Proceedings of the 2016 CHI conference on Human Factors in Computing Systems*. ACM, 2016, pp. 5372–5384.

[27] C. Kelleher and W. Hnin, "Predicting cognitive load in future code puzzles," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: ACM, 2019, pp. 257:1–257:12. [Online]. Available: http://doi.acm.org/10.1145/3290605.3300487

[28] D. Popovic, M. Stikic, T. Rosenthal, D. Klyde, and T. Schnell, "Sensitive, diagnostic and multifaceted mental workload classifier (physioprint)," vol. 9183, 08 2015.

[29] Y. Ke, H. Qi, F. He, S. Liu, X. Zhao, P. Zhou, L. Zhang, and D. Ming, "An eeg-based mental workload estimator trained on working memory task can work well under simulated multi-attribute task," *Frontiers in Human Neuroscience*, vol. 8, p. 703, 2014. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnhum.2014.00703

[30] Y. Santiago-Espada, R. R. Myer, K. A. Latorella, and J. R. Comstock Jr, "The multi-attribute task battery ii (matb-ii) software for human performance and workload research: A user's guide," 2011.

[31] T. Appel, C. Scharinger, P. Gerjets, and E. Kasneci, "Cross-subject workload classification using pupil-related measures," in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, 2018, pp. 1–8.

[32] T. Appel, N. Sevcenko, F. Wortha, K. Tsarava, K. Moeller, M. Ninaus, E. Kasneci, and P. Gerjets, "Predicting cognitive load in an emergency simulation based on behavioral and physiological measures," in *2019 International Conference on Multimodal Interaction*, 2019, pp. 154–163.

[33] W. K. Kirchner, "Age differences in short-term retention of rapidly changing information." *Journal of experimental psychology*, vol. 55, no. 4, p. 352, 1958.

[34] P. S. GmbH. (1999) World of emergency. [Online]. Available: [https://www.world-of-emergency.com/?lang=en

[35] K. Rayner, "Eye movements in reading and information processing: 20 years of research." *Psychological Bulletin*, vol. 124, no. 3, p. 372, 1998.

[36] ——, "Eye movements and attention in reading, scene perception, and visual search," *The quarterly journal of experimental psychology*, vol. 62, no. 8, pp. 1457–1506, 2009.

[37] A. F. Kramer, "Physiological metrics of mental workload: A review of recent progress," *Multiple-task Performance*, pp. 279–328, 1991.

[38] D. Kahneman, *Attention and effort*. Citeseer, 1973, vol. 1063.

[39] O. Palinko, A. L. Kun, A. Shyrokov, and P. Heeman, "Estimating cognitive load using remote eye tracking in a driving simulator," in *Proceedings of the 2010 symposium on eye-tracking research & applications*. ACM, 2010, pp. 141–144.

[40] S. Chen and J. Epps, "Using task-induced pupil diameter and blink rate to infer cognitive load," *Human–Computer Interaction*, vol. 29, no. 4, pp. 390–413, 2014.

[41] P. Kiefer, I. Giannopoulos, A. Duchowski, and M. Raubal, "Measuring cognitive load for map tasks through pupil diameter," in *The Annual International Conference on Geographic Information Science*. Springer, 2016, pp. 323–337.

[42] K. G. Seeber, "Cognitive load in simultaneous interpreting: Measures and methods," *Target. International Journal of Translation Studies*, vol. 25, no. 1, pp. 18–32, 2013.

[43] N. Castner, T. Appel, T. Eder, J. Richter, K. Scheiter, C. Keutel, F. Hüttig, A. Duchowski, and E. Kasneci, "Pupil diameter differentiates expertise in dental radiography visual search," *PLOS ONE*, vol. 15, no. 5, pp. 1–19, may 2020. [Online]. Available: https://doi.org/10.1371/journal.pone.0223941

[44] S. P. Marshall, "Method and apparatus for eye tracking and monitoring pipil dilation to evaluate cognitive activity," Patent US 006 090 051A, 07 18, 2000. [Online]. Available: https://patentimages.storage.googleapis.com/pdfs/9171d27ab488a900c7db/U

[45] ——, "The index of cognitive activity: measuring cognitive workload," in *Proceedings of the IEEE 7th Conference on Human Factors and Power Plants*, 2002, pp. 7–5–7–9.

[46] S. Benedetto, M. Pedrotti, L. Minin, T. Baccino, A. Re, and R. Montanari, "Driver workload and eye blink duration," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 14, no. 3, pp. 199 – 208, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S136984781000094X

[47] K. F. V. Orden, W. Limbert, S. Makeig, and T.-P. Jung, "Eye activity correlates of workload during a visuospatial memory task," *Human Factors*, vol. 43, no. 1, pp. 111–121, 2001, pMID: 11474756. [Online]. Available: http://dx.doi.org/10.1518/001872001775992570

[48] M. Á. Recarte, E. Pérez, Á. Conchillo, and L. M. Nunes, "Mental workload and visual impairment: Differences between pupil, blink, and subjective rating," *The Spanish journal of psychology*, vol. 11, no. 2, pp. 374–385, 2008.

[49] M. De Rivecourt, M. Kuperus, W. J. Post, and L. J. Mulder, "Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight," *Ergonomics*, vol. 51, no. 9, pp. 1295–1319, 2008.

[50] S. Benedetto, M. Pedrotti, and B. Bridgeman, "Microsaccades and exploratory saccades in a naturalistic environment," *Journal of Eye Movement Research*, vol. 4, no. 2, pp. 1–10, 2011.

[51] X. Gao, H. Yan, and H.-j. Sun, "Modulation of microsaccade rate by task difficulty revealed through between-and within-trial comparisons," *Journal of Vision*, vol. 15, no. 3, pp. 3–3, 2015.

[52] E. Siegenthaler, F. M. Costela, M. B. McCamy, L. L. Di Stasi, J. Otero-Millan, A. Sonderegger, R. Groner, S. Macknik, and S. Martinez-Conde, "Task difficulty in mental arithmetic affects microsaccadic rates and magnitudes," *European Journal of Neuroscience*, vol. 39, no. 2, pp. 287–294, 2014.

[53] K. Krejtz, A. T. Duchowski, A. Niedzielska, C. Biele, and I. Krejtz, "Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze," *PLOS ONE*, vol. 13, no. 9, p. e0203629, 2018.

[54] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006.

[55] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[56] S. H. Fairclough, L. J. Moores, K. C. Ewing, and J. Roberts, "Measuring task engagement as an input to physiological computing," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, Sep. 2009, pp. 1–9.

[57] A. Duchowski, K. Krejtz, J. Zurawska, and D. House, "Using microsaccades to estimate task difficulty during visual search of layered surfaces," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2019.

[58] M. A. Hogervorst, A.-M. Brouwer, and J. B. F. van Erp, "Combining and comparing eeg, peripheral physiology and eye-related measures for the assessment of mental workload," *Front Neurosci*, vol. 8, p. 322, Oct 2014, 25352774[pmid]. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4196537/

Tobias Appel received his Bachelor's and Master's degree in Computer Science from the University of Tübingen in 2014 and 2017 respectively. As of 2017 he is pursuing his Ph.D. in Computer Science at the LEAD Graduate School and Research Network and the University of Tübingen. His research focuses on the evaluation of cognitive load based on Eye Tracking and other physiological sensors. In his research he relies on machine learning to realize cross-participant and cross-task solutions.

Peter Gerjets finished his diploma in psychology at the University of Goettingen in 1991. From 1991 to 1995 he was a Research Associate at the University of Göttingen where he received his Ph.D. in 1994. Afterwards he has been working as Assistant Professor at the Saarland University in Saarbrücken where he finished his habilitation in 2002 before taking over his current position at the University of Tübingen. Since 2002 he has been working as principal research scientist at the Knowledge Media Research Center and beside as full professor for research on learning and instruction at the University of Tübingen. He was honoured with the Young Scientist Award of the German Cognitive Science Society in 1999 and served in the editorial boards of the Journal of Educational Psychology, Educational Re-search Review, Computers in Human Behavior, Metacognition and Learning, and Educational Technology, Research, and Development. His current research focuses on multimodal and embodied interaction with digital media as well as on learning from multimedia, hypermedia, and the Web. He is a member of DGPs, APS, and EARLI and served as coordinator of the EARLI Special Interest Group 6: Instructional Design.

Stefan Hoffmann Stefan Hoffmann is a developer of video games for more than 30 years now, with a strong focus on Serious Games and research projects with Serious Games and/or Gamification. He works for Serious Games Solutions, a division of Promotion Software GmbH, the software developer behind the "Emergency Lernspiel" (Learning game). He developed games or gamified apps for mobile platforms, browser and HoloLens. His focus is on game design and project management.

Korbinian Moeller received his PhD in Psychology from University of Tübingen, Germany, in 2010. Currently he is Professor of Mathematical Cognition at Loughborough University, United Kingdom. His research interests include neurocognitive foundations of mathematical cognition and developmental psychology in the context of educational games and learning analytics.

Manuel Ninaus received his PhD in Psychology from University of Graz, Austria, in 2015. Currently he is PostDoc at the Leibniz-Institut für Wissensmedien in Tübingen, Germany. He is elected board member of the Serious Games Society. His research interests include neuroscience and educational psychology in the context of educational games and learning analytics.

Christian Scharinger Christian Scharinger received the PhD degree in cognitive science from the University of Tuebingen in 2015. He currently works at the Multimodal Interaction Lab of the Knowledge Media Research Center Tuebingen. He has a profound expertise in (neuro-) physiological measures like eye-tracking and EEG. In his research he tries to combine basic and applied research areas. His research interests comprise working memory, executive functions, learning, hypertext reading, web searching, and multimedia, with a strong focus on physiological measures of cognitive load.

Natalia Sevcenko Natalia Sevcenko is currently pursuing her doctorate in psychology at the LEAD Graduate College and Research Network at the University of Tübingen. She received her bachelor's and master's degrees in psychology in 2015 and 2017 respectively at the Eberhard-Karl-University of Tübingen. Her research interests lie in the field of human-machine interaction, especially in the area of behavior and sensor-based measurement of cognitive states of operators and their relation to learning outcomes and personal predisposition.

Franz Wortha Franz Wortha is currently pursuing his Ph.D. in Psycholgoy at the LEAD Graduate School and Research Network at the University of Tübingen. He received his Bachelor's degree in Industrial Engineering from the University of Applied Sciences in Dresden in 2013 and his Master's degree in Psychology from Technische Universität Dresden in 2016. His research interests lie in self-regulated learning with a focus on metacognitive and emotional processes and their relation to learning outcomes and personal predisposition.

Enkelejda Kasneci is a Professor of Computer Science at the University of Tübingen, Germany, where she leads the Human-Computer Interaction Lab. As a BOSCH scholar, she received her M.Sc. degree in Computer Science from the University of Stuttgart in 2007. In 2013, she received her PhD in Computer Science from the University of Tübingen. For her PhD research, she was awarded the research prize of the Federation Südwestmetall in 2014. From 2013 to 2015, she was a postdoctoral researcher and a Margarete-von-Wrangell Fellow at the University of Tübingen. Her research evolves around the application of machine learning for intelligent and perceptual human-computer interaction. She serves as academic editor for PlosOne and as a TPC member and reviewer for several major conferences and journals.