ELSEVIER

Computer Methods and Programs in Biomedicine 00 (2022) 1–12

cmpb_logo.PNG

www.elsevier.com/cmpb

COLET: A Dataset for COgnitive workLoad estimation based on Eye-Tracking

Emmanouil Ktistakis[a,b,1,*], Vasileios Skaramagkas[a,1], Dimitris Manousos[a], Nikolaos S. Tachos[c], Evanthia Tripoliti[d], Dimitrios I. Fotiadis[c], Manolis Tsiknakis[a,e]

[a]*Institute of Computer Science, Foundation for Research and Technology Hellas (FORTH), GR-700 13 Heraklion, Greece*
[b]*Laboratory of Optics and Vision, School of Medicine, University of Crete, GR-710 03 Heraklion, Greece*
[c]*Biomedical Research Institute, FORTH, GR-451 10, Ioannina, Greece and the Dept. of Materials Science and Engineering, Unit of Medical Technology and Intelligent Information Systems, University of Ioannina, GR-451 10, Ioannina, Greece*
[d]*Dept. of Materials Science and Engineering, Unit of Medical Technology and Intelligent Information Systems, University of Ioannina, GR-451 10, Ioannina, Greece*
[e]*Dept. of Electrical and Computer Engineering, Hellenic Mediterranean University, GR-710 04 Heraklion, Crete, Greece*

Abstract

The cognitive workload is an important component in performance psychology, ergonomics, and human factors. Unfortunately, publicly available datasets are scarce, making it difficult to establish new approaches and comparative studies. In this work, COLET-COgnitive workLoad estimation based on Eye-Tracking dataset is presented. Forty-seven (47) individuals' eye movements were monitored as they solved puzzles involving visual search tasks of varying complexity and duration. The authors give an in-depth study of the participants' performance during the experiments while eye and gaze features were derived from low-level eye recorded metrics, and their relationships with the experiment tasks were investigated. The results from the classification of cognitive workload levels solely based on eye data, by employing and testing a set of machine learning algorithms are also provided. The dataset is available to the academic community.

Keywords:
Cognitive workload, Workload classification, Eye movements, Machine Learning, Eye-tracking, Affective computing

1. Introduction

The study of mental workload, also known as cognitive workload (CW), is a vital aspect in the areas of psychology, ergonomics, and human factors to understand the performance during a task or process (1). Despite the multitudinous and extended research in this area, there is no single definition to describe cognitive workload. Often we refer to cognitive workload as taskload i.e. the effort needed to perform a certain procedure. However, defining workload can be a rather subjective task depending on how different people with different experience and abilities can handle the same task (2; 3; 4; 5). So, a general definition of mental workload would be the product of factors that contribute to oneâs workload efficiency for a given task.

Numerous studies have concentrated on determining cognitive effort purely on the basis of ocular characteristics for various tasks, highlighting the need of further research (6). The majority of them provide binary categorization findings, indicating a high or low level of cognitive workload, with some obtaining highly accurate results (7; 8; 9; 10). However, there are just a few published attempts that focus on multi-class classification (high/medium/low), and the resulting performance is inferior (7; 11). The three-class classification is gaining popularity during the last years (12). A meta-analysis of the related studies has shown that the gaze extracted features that are better correlated with cognitive workload are blink rate, the diameter of the pupil, the duration of the blink, and the duration of fixations (13).

Our contribution in the field of computational biomedicine is a dataset to be used for the estimation of the cognitive workload level, based on eye-tracking (COLET) combined with both objective and subjective performance evaluation. The collection contains eye characteristics and movement recordings of 47 participants and their performance scores when solving puzzles related to visual search tasks. The recorded signals contain a number of metrics related to gaze positions, blinks, and pupil characteristics, enabling for the extraction and analysis of a broad variety of eye features, such as fixations and saccades. For the generation of the dataset, participants took part

*Corresponding author
URL: mankti2@gmail.com (Emmanouil Ktistakis)
[1]Authors contributed equally

in an experiment in which they executed a task under four different levels of difficulty, caused by time pressure and a supplementary task involving backward counting. Full details of the experimental protocol can be found in Section 3. After the conclusion of each task, we collected ratings from individuals in relation to a simplified version of the NASA task load index (NASA TLX) tool (14). The database is available to the academic community (15). To our knowledge, this is one of very few public databases with eye-tracking data obtained from mentally demanding visual search puzzles, that consists of such a high number of participants and can contribute towards the development and evaluation of modern human-computer interaction systems.

In this manuscript and in Section 2, eye features involved in cognitive processes are reported and subsequently, studies related to cognitive workload databases based on image as well as audio-visual stimuli are presented. Section 3 presents the experimental scenarios, stimuli selection, annotation explanation and equipment utilized whereas an overview of the experimental setup and the methods employed for the assessment of affect and personality traits is outlined. In Section 4, the low-level eye metrics analysis and algorithmic process for the extraction of eye and gaze features is explained. Afterwards, in Section 5, identified statistical correlations between mental workload and eye features are discussed in detail. Additionally, the strategy and structure of a machine learning approach for the identification of cognitive states is investigated and the results from the methods developed regarding the recognition of the states are presented. Furthermore, the results regarding the statistical and machine learning analysis are interpreted and a benchmarking upon our findings is performed in Section 6. Finally, the conclusions drawn from this study are recapitulated and discussed.

2. Related works

Due to the diversity of criteria for CW, there are several methods for quantifying it. No sensor can provide an accurate picture of how an individual responds to a task however the estimation of eye and gaze patterns and features can aid in determining workload levels (16; 13). The next paragraphs will discuss the most widely used and robust eye markers for assessing cognitive strain. These include fixations, eye movements, blinks, and pupil size measures. The blink rate, the diameter of the pupil, the duration of the blink, and the duration of fixations appear to be the most often employed eye-related metrics for investigating correlations with CW. Table 3 of our previous work (13) presents in detail the eye-related measurements and their relationship with increased CW.

2.1. Eye-tracking features involved in cognitive processes

The number of fixations and fixation duration have been demonstrated to increase as cognitive load increases during mentally demanding tasks such as surgical operations, simulated flight tasks (17; 18) and extracurricular activities (19; 20). It has been shown that mean fixation duration has a significant negative correlation with the level of cognitive load in simulated flight and driving tasks (21; 18; 22), and in video gaming (20), while maximum fixation duration also demonstrates the same behaviour (23). As far as the fixation rate is concerned, novices performed more fixations than experts in a surgical environment (17).

Saccades are the most often examined kind of eye movement in cognitive workload research (24). The average peak saccadic velocity is seen to increase in a positive linear fashion as the cognitive effort increases (25). Additionally, video game situations highlight the importance of saccadic velocity in determining the degree of cognitive burden. In (20), the saccadic peak velocity decreased while the speed of the game slowed down and increased rapidly when the game speed raised in proportion to the increase of the difficulty level. Furthermore, the average and maximum saccadic amplitudes have a moderately positive connection with the degree of difficulty (23). During the completion of simulated flying activities, saccade velocity and frequency rose in response to increasing time pressure and reduced in response to subject overload. The maximal workload was shown to be strongly linked with the average saccadic velocity and frequency peaks (18). When drivers were asked to do a secondary task while driving, a substantial rise in their saccade rate was seen as the task complexity rose (26).

Saccades may be used in conjunction with fixations as previously discussed and can also reflect the clinician's ability level, since beginner surgeons make more saccadic movements than intermediate surgeons (17). The difference in saccade amplitude between novices and experts failed to approach significance in a low cognitive load task in which participants were required to operate a training version of a military land platform (27). According to (24), certain trajectories (rapid and circular) result in greater gaze disparities of smooth pursuit eye movements when cognitive burden is present. In another investigation, eye-target synchronization during smooth pursuit eye movement improved in young normal volunteers subjected to intermediate cognitive strain (28).

Microsaccades are frequently employed to investigate cognitive processes. Microsaccade rate decreases with increasing task complexity in mental arithmetic tasks when fixating a central target (29; 30). In a more recent investigation, it was found that increasing cognitive load had no effect on microsaccade rate (31). Microsaccade amplitude appears to rise with task complexity (29), however the behavior of microsaccade peak velocity and amplitude is unknown.

Given the much greater prevalence of short blinks under situations of high visual load, blink length is a sensitive indication of cognitive effort (32). According to (20), blink length and frequency were likewise shown to be optimal at the lowest pace in the video game, whereas blink frequency declined as the mental strain grew. According to (33), blink rate reduced with increasing CW from low to medium, but did not alter anymore with further increasing cognitive load. Other authors have also shown a robust correlation between maximal blink duration and the amount of mistakes made during a high cognitive load arithmetic exercise (23). In another study, it was demonstrated that the blink rate decreases significantly from the rate observed at the resting state and is sensitive to the phases of microsurgical suture (34). In a driving scenario (35), the eye tracker demon-

strated rising blink rates concurrent with a rise in the difficulty of a secondary activity conducted in tandem with driving.

The pupil area is significantly related to the user's current work difficulty (36; 37) and mean pupil diameter has been demonstrated to correlate positively with cognitive effort across a variety of activities (33; 38; 9). Additionally, pupil diameter rises in proportion to the complexity of the user's ongoing secondary activity (35). Pupil size appears to rise according to the amount of effort required to complete intellectually demanding activities (39). In (23) where participants performed arithmetic tasks, maximum pupil dilation was substantially associated with the amount of mistakes made under conditions of high cognitive load. The increasing mental burden in (40) dilates the pupils, and because the participants are nearing overload, the saccade rate also increases. In (41), it is demonstrated that the topic and difficulty level of a text had no significant effect on pupil size measurements.

2.2. Eye-tracking databases for cognitive workload identification

Although eye movements have proven to be useful indicators of cognitive processes (13), only few authors have focused on the development of relevant databases. Amongst those available, MAMEM datasets (Phase 1 and Phase 2) (42) blend multimodal biosignals and eye tracking data collected within the context of human-computer interaction. The datasets contain eye tracking data from 34 people (18 able-bodied and 16 with motor impairments), as well as electroencephalography (EEG), galvanic skin response (GSR), and heart rate (HR) signals. The data were collected during engagement with a specially built interface for online browsing and manipulating multimedia material, as well as during fictitious mobility activities.

The EGTEA Gaze+ dataset (43) comprises almost 28 hours of video footage from 86 separate sessions including 32 people completing seven distinct food preparation activities. There are movies, eye-tracking data, action annotations, and hand masks included in the dataset. It is a supplement to the previously released GTEA Gaze+ dataset (44).

In USC CRCNS Dataset (45), the authors used abrupt transitions to convert continuous video clips into clip parts (jump cuts). 16 subjects had their saccadic motions recorded as objective behavioral markers of attentional choices. By assessing the agreement between human attentional selection and prediction produced by a neurally grounded computational model, they were able to measure the usage of perceptual memory across viewing circumstances and across time. Additionally, MIT CVCL Search Model Database (46) comprises of eye-tracking recordings from 14 participants while performing person detection tasks. The ground-truth eye movement data were used to evaluate three computational models for search guidance based on saliency, target features, and scene context respectively.

Despite their major contribution, each of the aforementioned databases has distinct drawbacks. The methodologies used suffer from a limited number of available eye and gaze measurements. This is especially critical when examining relationships between eye movements and cognitive states, since some measures, such as blink duration and saccadic velocity, play a vital role in the estimation of increased cognitive workload (13). Furthermore, the above mentioned datasets do not primarily target to study the alterations of ocular movements in relation to cognitive load variations which are measured precisely based on the NASA-TLX index.

The COLET database, presented in our work, explores the possibility to analyze CW levels induced by visual search puzzles along with secondary tasks performed from different users. To the best of our knowledge, this is the first eye-tracking based dataset available to the academic community that combines eye and gaze movements signals along with performance assessment from the users, thus studying the effect and the correlation of eye movements with increased mental demand. COLET is based on visual stimuli and supplementary tasks specifically selected to affect cognitive effort and is the largest eye-tracking database in terms of the number of participants as well as the quantity of eye and gaze metrics currently available.

3. Methodology

In this Section, we describe the protocol followed for generating the dataset and every material that was used in the study.

3.1. Participants

The experimental protocol (110/12-02-2021) was submitted and approved by the Ethical Committee of the Foundation for Research and Technology Hellas (FORTH).

Exclusion criteria for the participants included: any known ocular disease, spectacle-corrected binocular visual acuity in 80 cm worse than 0.10 logMAR (0.8 decimal acuity equivalent), clinically significant abnormal phorias.

Fifty six (56) individuals volunteered for the study and nine of them were excluded: Seven due to the exclusion criteria and two because of poor quality recordings. Thus, analysis was performed for the remaining forty-seven (47) participants (26 female, 21 male). Their mean age was 32±8 years (range: 18-47 years), their mean education level was 17±2 years (range:12-21 years) and their mean binocular visual acuity at 80 cm was -0.10±0.08 logMAR (range: 0.10-(-0.29) logMAR).

3.2. Materials and Setup

A set of 21 images of indoor scenes was chosen from the free database âIndoor scene recognitionâ (47). A grid was added to each image, thus dividing it into nine (9) equal squares as it is shown in Fig. 3. Each image was selected so as a specific object was present on some of the 9 squares, thus looking like a CAPTCHA puzzle. The images were presented on a computer screen (LCD, 24″, 1280x720) at 80cm distance from the participant as it is shown in Fig. 1.

Eye-tracking measurements were recorded using the Pupil Labs "Pupil Core" eye-tracker. Recordings were binocular with 240 Hz sampling frequency, accuracy 0.60° and precision 0.02. All measurements were performed with the participants seated

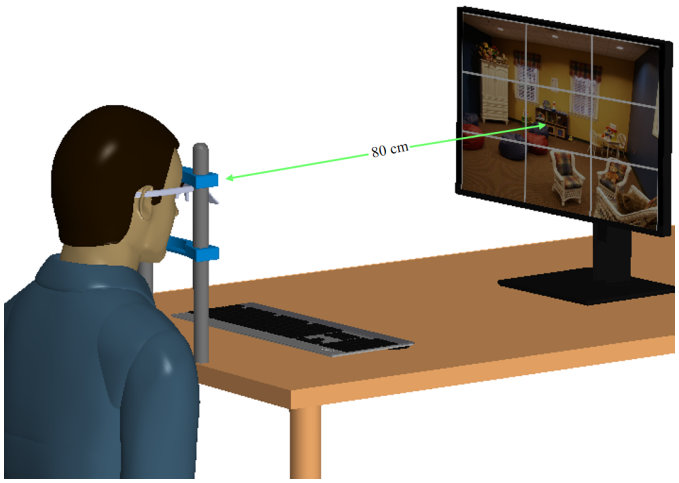

Figure 1: Graphical representation of the experimental setup.

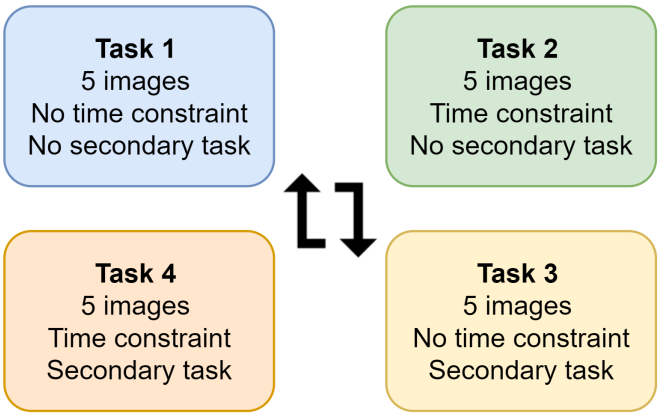

Figure 2: Two-by-two factorial design of the experimental study.

on a chair with their head stabilized by means of a chin and head rest to minimize head movements.

Standardised logMAR acuity was measured with the European-wide standardized logMAR charts (48). Stereopsis was evaluated with the cover test.

Recordings were performed under controlled, photopic lighting conditions, which were achieved with the room lights on. Illuminance at cornea when screen was off, was 400 lux and when on, in blank screen, it was 450 lux.

3.3. Experimental Procedure

In the beginning, a binocular visual acuity test at distance of 80 cm and a stereopsis test were conducted. Subsequently, all participants read and signed an Information Consent Form. Subsequently and were led to the experiment room. All the necessary measures for the protection of the participants and the research team from the SARS-CoV-2 pandemic and expansion of the corona-virus were applied.

Following that, participants were invited to complete some demographics (age, education level) on a computer screen. A test with a random image was conducted next in order the participant to get familiar with the process. After the test the main part of the study commenced.

The study used a two-by-two factorial design, with the two variables being Time Constraint (with or without) and Tasking (single or multi task). The interaction of these variables resulted in the establishment of four experimental task conditions as shown in Fig. 2. Time constraints were imposed by instructing participants to finish the assignment ”as quickly as possible,” whereas ”no time constraints” were introduced by instructing participants to complete the activity ”at a comfortable pace”. Each task consisted of 5 random images/trials. The tasks were presented in a randomised order.

At any time during the study, participants could request that the process be stopped and their data deleted.

The study’s primary objective was the subjects to perform a visual search task based on a CAPTCHA-like test. The participants were shown the pictures and asked to complete CAPTCHA-style puzzles. The dual task required participants to conduct an interference task, i.e., backward counting from 1000 by subtracting 4, while performing the primary visual search task. An image from the ones used in the experimental procedure is shown in Fig. 3. Throughout the task completing procedure, a member of the research group monitored the gaze tracker’s output on a second screen in case any anomalies in the recordings happened or the participant needed further assistance.

Figure 3: A sample trial/image of the CAPTCHA test. Instructions: âChoose the squares in which pouffes are located”.

3.4. Cognitive Workload assessment

A variety of measures to evaluate cognitive workload have been used and they can be divided in four categories; subjective measures, performance measures, psychophysiological measures and analytical measures (49). Rating scales and questionnaires may be considered as subjective measures, reaction time and number of errors may be considered as performance measures and any eye movement, cardio and brain related measure may be considered as psychophysiological measure. Analytical measures are the ones that are derived from mathematical models and task-analytic techniques.

In this study, subjective, performance and physiological measures were used to evaluate cognitive workload.

3.4.1. Subjective measures

At the end of each task the participants were asked to complete a simplified version of the NASA-TLX questionnaire, a subjective workload assessment tool (14). It consists of six subscales and originally it derives an overall workload score based on a weighted average of ratings on these subscales. The six subscales are:

1. Mental Demand: The level of mental and perceptual engagement that was necessary for the completion of the task.
2. Physical Demand: The level of physical activity that was necessary for the completion of the task.
3. Temporal Demand: The time pressure felt due to the pace at which the tasks or task elements occurred.
4. Performance: The level of success resulted from the completion of the task.
5. Effort: The level of difficulty to work (both mentally and physically) for the achieved performance level.
6. Frustration: The level of stress or irritation felt during the task vs the level of pleasantness or calmness.

A simplified version of NASA-TLX, called NASA RTLX, was proposed by (50) and has been used widely since (51; 52; 53). In NASA RTLX, the pairwise comparisons of the subscales that are used in the original NASA TLX for weighing, are omitted. In the present study, the simplified version of NASA RTLX was used. Apart from the ratings of the six subscales, the mean value of the six ratings was also evaluated. The higher the mean value, the higher the experienced cognitive workload (14).

3.4.2. Performance measures

During all trials and tasks, the number of mistakes and missed correct squares was measured. Additionally, the time that was needed to complete a trial was measured and it is referred to as Reaction Time. An attempt to combine speed and error is the Inverse Efficiency Score (IES) (54) which is widely used. For a given participant, IES is given by the mean reaction time (RT) in a particular condition divided by the percentage of correct answers (PC) (55). PC was calculated as the number of the correct answers divided by the sum of the correct, the wrong and the missed answers.

3.4.3. Physiological measures

Physiological measures were derived from the eye tracker raw data. They are in total 28 measures and they are fixation, saccade, blink and pupil related, including skewness, kurtosis and coefficient of variation (CV) for every eye feature. Skewness is a measure of symmetry of a distribution, while kurtosis is a measure of whether a distribution is heavy-tailed or light-tailed relative to a normal distribution. Coefficient of variation shows the extent of variability in relation to the mean of the population and is a dimensional number. It is defined as the ratio of the standard deviation to the mean of a population. For ease of reading, it will be referred to as Variation. For a detailed presentation of the features studied, is given in Table 2.

4. Data analysis methodology

The computational procedure that was followed to compute and evaluate the eye and gaze related features from the raw data acquired by the gaze tracking device is described in detail below.

4.1. Raw data processing

When collecting data, some noise is typically present due to eye blinking and failure to capture corneal reflections (i.e., signal loss). The gaze tracker's output includes the gaze positions (x,y coordinates), blink timings (start and end times), and pupil diameter in mm. These measures include a variety of sources of noise, including the eye-tracker and the participants. Filtering and denoising are used to eliminate this undesired volatility in eye movement data (56).

The raw gaze coordinates in normalized pixels form are converted to degrees of visual angle, and the instantaneous sample-to-sample gaze movement between two consecutive gaze locations is determined, resulting in the computation of the angular velocity at the specified sampling frequency F_s. To decrease velocity noise, we used a five-tap velocity filter whose form was modified in response to a defined velocity peak value during a saccade (57).

4.2. Fixation and Saccade Detection

In this work, fixations and saccades are identified based on the Velocity-Threshold Identification (I-VT) algorithm (58) due to its superiority when considering sample-by-sample comparisons (59). Additionally, we introduced an additional minimum time criteria to assess the duration of the fixations. According to the method, a defined velocity threshold determines a gaze point as a fixation or saccade. Then, consecutive fixation points are collapsed into fixation groups based on the duration threshold. In the I-VT algorithm the velocity threshold for saccade detection was set to 45 deg./sec, as in Andersson et al. (2016) (59). In addition, the minimum fixation duration threshold were determined at 55 msec (60).

4.3. Pupil and Blink Detection

The effect of a certain factor on pupil size is hard to evaluate, since pupil diameter and its variation is highly dependent on multiple factors that need to remain fixed, such as lighting conditions (61; 62; 63) and the adapting field size (64; 65). In our study, the luminance of each image may be a factor of pupil size change.

The current study's experimental design was chosen to minimize this impact as much as possible. Initially, the lighting settings of the room were configured to be photopic, ensuring that the effect of brightness shifts of the images was minimal. For

the same purpose, the screen was positioned at a distance of 80 cm away from the participant.

In order to evaluate whether eventually the influence of the image luminance was sufficiently low, linear regression analysis between mean pupil diameter of each participant and the V component of the HSV color space of each image was carried out. The idea was based on a recent attempt to remove the movie luminance effect by extracting the estimated pupil diameter based on the V component of the HSV color space, from the recorded pupil diameter (66). Among the 47 participants, only 2 showed correlation; one moderate (r=0.442, p=0.034) and one strong (r=0.635, p=0.003).

The results of linear regression analysis were considered satisfactory since no strong correlation between pupil diameter and the V component was found in any of the participants. Thus, all analysis on pupil diameter was carried out with the pupil diameter obtained from the gaze tracker.

The eye-tracker-derived pupil diameter and blink timings aid the extraction of additional pupil and blink-related features. The pupil recognition algorithm locates the black pupil in the infrared lighted eye camera frame (67). Because the algorithm is not influenced by corneal reflection, it is suitable for persons who use contact lenses or spectacles. The start and finish periods of blinks are determined using a confidence threshold matching to the effective detection of the pupil region.

4.4. Feature selection and model training

Table 2 summarizes the 28 eye and gaze features collected from fixation, saccade, blinks, and pupil characteristics. After the features were extracted and normalized using a MinMaxScaler function, we built a correlation matrix to study which are highly correlated with each other. Then, we derived the most dominant features for every class (see Section 5.2) based on the ANOVA repeated measures analysis performed in Section 5.1.

In total, 8 classifiers were trained and tested during the classification procedure. More specifically: Gaussian Naive Bayes (GNB), Random Forest (RF), Linear Support Vector Machine (SVM), Ensemble Gradient Boosting (EGB), K-Nearest Neighbor (k-NN), Bernoulli Naive Bayes (NB), Logistic Regression (LR) and Decision Trees (DT). We studied the behavior of a variety of well-known and extensively used classifiers in comparison to the relevant literature. To fine tune the hyperparameters of each classifier we performed a RandomSearch iterating 1000 times through training data to find the combination of parameters that maximizes the overall performance and accuracy. We split the data into training and testing, with the number of the test data being 20% of the total number of examples. We evaluated the models using the metrics of accuracy and f1-score. Furthermore, we validated the models using a k-fold cross-validation (k=5).

5. Results

In this section, the results of the study are presented. The statistical analysis is shown and the Machine Learning analysis which has been performed to identify any relation between eye features and cognitive workload tasks and levels.

5.1. Statistical Analysis

5.1.1. Cognitive Workload induction

A repeated measures ANOVA determined that all NASA RTLX subscales and mean NASA differed statistically significantly among the different tasks. Post hoc analysis with a Bonferroni adjustment revealed that mean NASA was statistically significantly different between all pairs of tasks (p<0.014) and it was getting gradually higher when moving from Task 1 to Task 4 (Fig. 4A). Mental demand was statistically significantly higher in task 2 compared to task 1 (11.0 (95% CI, 4.0 to 18.0), p<0.0001) and even higher in task 3 (30.2 (95% CI, 19.9 to 40.5), p<0.0001). Mental demand in task 4 was not statistically significantly different from the mental demand in task 3 (6.0 (95% CI, -1.1 to 13.2), p=0.143). Temporal demand was higher in task 2 compared to task 1 (22.5 (95% CI, 10.5 to 34.6), p<0.0001) and higher in task 4 compared to task 3 (22.6 (95% CI, 12.1 to 33.1), p<0.0001), as expected. Performance and Frustration were statistically significantly different only between single (1 and 2) and double (3 and 4) tasks, while Effort was statistically significantly different between all pairs apart from task 3 and task 4 (3.0 (95% CI, -5.5 to 11.5), p=1.000). Finally, post hoc analysis did not show any statistically significant difference in Physical Demand between any pair of tasks (p>0.102). Table 1 shows mean values of all NASA subscales in all tasks.

Table 1: NASA RTLX scores

Subscale	Task 1	Task 2	Task 3	Task 4
Mental Demand	25.6	36.7	66.9	72.9
Physical Demand	16.0	19.6	26.5	27.3
Temporal Demand	19.0	41.5	37.3	59.9
Performance	16.6	23.2	43.0	45.6
Effort	26.7	36.7	63.8	66.8
Frustration	12.5	17.4	31.6	40.4
Mean	19.4	29.2	44.8	52.2

Repeated measures ANOVA also showed that the mean number of mistakes per task ($F(3, 141) = 4.354$, $P = 0.006$), the total time needed to complete a task (Reaction Time, RT) ($F(2.371, 111.445) = 49.878$, $P < 0.0001$) and the Inverse Efficiency Score (IES) ($F(2.062, 96.905) = 40.443$, $P < 0.0001$) differed statistically significantly among the different tasks. Post hoc analysis with a Bonferroni adjustment revealed that the number of mistakes was statistically significantly lower in task 1 compared to the the number of mistakes in tasks 3 and 4 (p<0.026), while no difference was found between the rest of the tasks. RT was statistically significantly lower in task 2 compared to task 1 (10.355 (95% CI, 4.46 to 16.25) sec, p<0.0001) and lower in task 4 compared to task 3, without reaching significance though (11.74 (95% CI, -0.78 to 24.25) sec, p=0.078). RT was also statistically significantly higher is tasks 3 and 4 compared to tasks

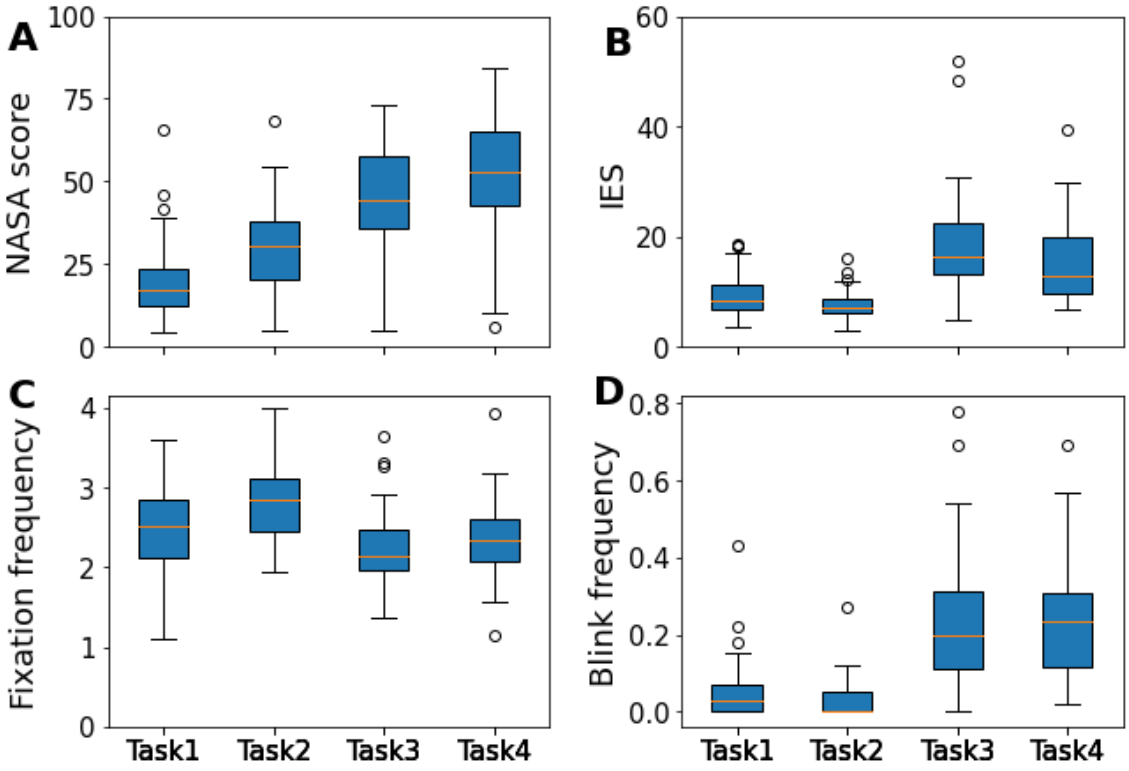

Figure 4: Box plots for A) Mean NASA score B) IES C) fixation frequency and D) blink frequency in each Task.

1 and 2 (p<0.0001). Post hoc analysis showed that the Inverse Efficiency Score (IES) was statistically significantly different among all pairs, apart from between task 3 and task 4 (3.285 (95% CI, -0.804 to 7.374) sec, p=0.191) (Fig. 4B).

Based on the subjective and performance measures, it becomes evident that cognitive workload is increased as one moves from Task 1 to Task 4. Among all measures, mean NASA score seems to be the measure that can better distinguish among the Tasks. Thus, from now on, cognitive workload (CW) will be considered as follows: $CW_{Task1} < CW_{Task2} < CW_{Task3} < CW_{Task4}$.

5.1.2. Eye feature analysis

A two-way repeated measures ANOVA was performed for all eye features. The two factors were Tasking (single or multi) and Time (with or without time constraint). There was a significant main effect of Tasking on seventeen (17) and a significant main effect of Time on seven (7) out of 28 features. There was also a significant interaction between Tasking and Time in 3 features. ANOVA results are presented in Table 3 and mean values of all features across Tasks, in Table 2. Post hoc analysis with a Bonferroni adjustment was also performed and mean differences of the features between two levels of each factor are presented in Table 3. Indicatively, box plots of fixation frequency and blink frequency are shown in Fig. 4C and 4D.

5.2. Machine learning analysis

In this section, we attempt to identify relations between fixation, saccade, blink and pupil related eye features and the CW tasks and levels. Therefore, the four tasks that the participants engaged with during the experimental procedure are noted as **T1, T2, T3, T4** for tasks 1, 2, 3 and 4, respectively. Moreover, an additional machine learning analysis was performed based on the subjective annotation as extracted from the mean NASA RTLX scores per task given by the participants. In the same manner, the outcome measure regarding CW levels can have three values: low, medium and high (68). Specifically, the divided CW instances low, medium and high are marked as class **C0, C1** and **C2**, respectively, where each class amounts to the one third of the total NASA-TLX mean score.

The results of the classification study using COLET are presented in Fig.5 and 6. Each of the two circles contains in its inner circles the classes of the respective classification attempt, the sample size for each class, and the three best performing classifiers in terms of their accuracy and f1-score. The highest accuracy for each classification attempt is highlighted in color.

From the results presented in Fig. 5, SVM classifiers classified correctly over 90% of the cases. Specifically, T2 was distinguished from T3 and T4 at percentages of 93 and 98%, respectively, demonstrating the effect of the secondary task (backwards counting) to the differentiation of the tasks. In the same manner, in T1/T3 and T1/T4 binary classification problems, the accuracy rate of GNB and k-NN classifiers was found to be 81 and 86%, respectively. Additionally, time pressure factor played a critical role in the discrimination among T1 and T2 classes with NB achieving 80% accuracy.

Interestingly, the backwards counting which was common for T3 and T4 seems to have outweighed their difference which

Table 2: Mean values and standard deviations of all features for the four Tasks.

	Task 1		Task2		Task 3		Task 4	
Feature	**Mean**	**Std.**	**Mean**	**Std.**	**Mean**	**Std.**	**Mean**	**Std.**
Fixation frequency (fix./sec)	2.50	0.53	2.80	0.48	2.24	0.47	2.35	0.50
Fixation duration (msec)	272.65	65.97	244.43	46.06	269.32	63.63	254.13	42.21
Variation	0.85	0.14	0.79	0.15	0.87	0.14	0.86	0.13
Skewness	2.49	0.87	2.26	0.67	2.49	0.84	2.57	1.04
Kurtosis	8.65	6.94	6.49	4.45	8.78	8.32	9.74	11.13
Saccade frequency (sac./sec)	1.71	0.85	1.85	0.90	3.40	2.56	3.35	2.55
Saccade amplitude (deg.)	14.10	0.92	13.96	0.99	14.15	1.45	14.06	1.29
Variation	0.12	0.06	0.10	0.05	0.23	0.11	0.22	0.10
Skewness	-0.39	1.55	-0.26	1.52	0.39	1.70	-0.07	1.70
Kurtosis	4.62	6.81	3.13	6.88	4.88	6.10	5.42	7.12
Saccade velocity (deg./sec)	146.41	68.55	132.91	68.85	261.13	103.09	267.90	113.34
Variation	0.77	0.30	0.61	0.28	0.75	0.19	0.75	0.21
Skewness	2.06	1.14	1.95	1.07	0.88	1.12	0.86	1.30
Kurtosis	5.08	6.06	4.84	5.28	0.72	3.68	1.10	4.67
Peak saccade velocity (deg./sec)	216.89	89.26	204.14	90.82	350.49	114.24	357.25	126.08
Variation	0.77	0.21	0.66	0.20	0.70	0.18	0.69	0.21
Skewness	1.64	1.10	1.60	1.16	0.51	1.08	0.45	1.16
Kurtosis	3.27	5.41	3.81	5.82	-0.14	3.02	0.04	2.61
Saccade duration (msec)	15.33	3.27	15.39	2.61	19.05	4.59	19.16	3.51
Variation	0.97	0.37	0.92	0.42	1.25	0.35	1.38	0.70
Skewness	3.00	1.39	2.88	1.93	2.98	1.04	3.39	2.23
Kurtosis	12.41	10.51	12.96	16.56	11.82	9.46	17.93	39.80
Blink frequency (blinks/sec)	0.05	0.07	0.03	0.05	0.24	0.17	0.24	0.17
Blink duration (msec)	205.54	48.35	200.31	47.53	229.00	40.48	212.29	31.97
Pupil diameter (mm)	3.49	0.60	3.66	0.61	3.80	0.69	3.82	0.71
Variation	0.05	0.04	0.05	0.03	0.07	0.04	0.06	0.03
Skewness	-0.30	0.53	-0.41	0.59	-0.16	0.82	-0.46	0.86
Kurtosis	1.02	4.25	1.04	2.22	2.34	3.30	3.70	6.16

Fixation duration, saccade amplitude and saccade duration are median values. Saccade velocity, peak saccade velocity, blink duration and pupil diameter are mean values. Fix./sec: number of fixations per second, sac./sec: number of saccades per second, deg.: degrees of visual angle during a saccadic movement, deg./sec: degrees of visual angle during a saccadic movement per second, blinks/sec: number of blinks per second.

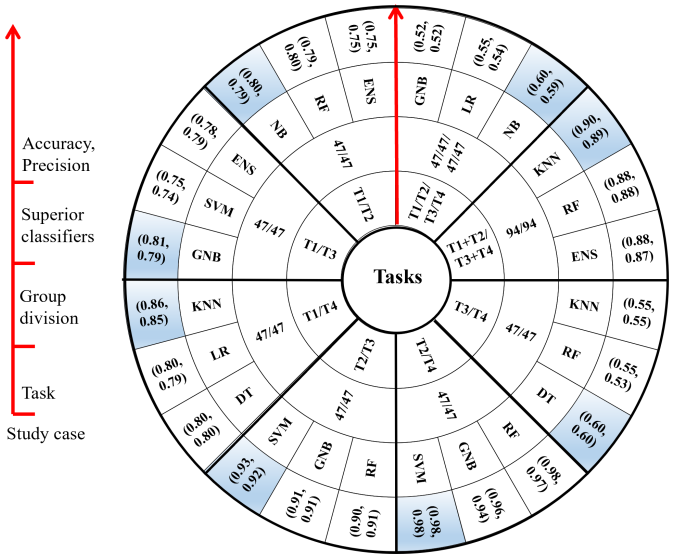

Figure 5: Superior algorithms in classifying the four tasks of cognitive workload.

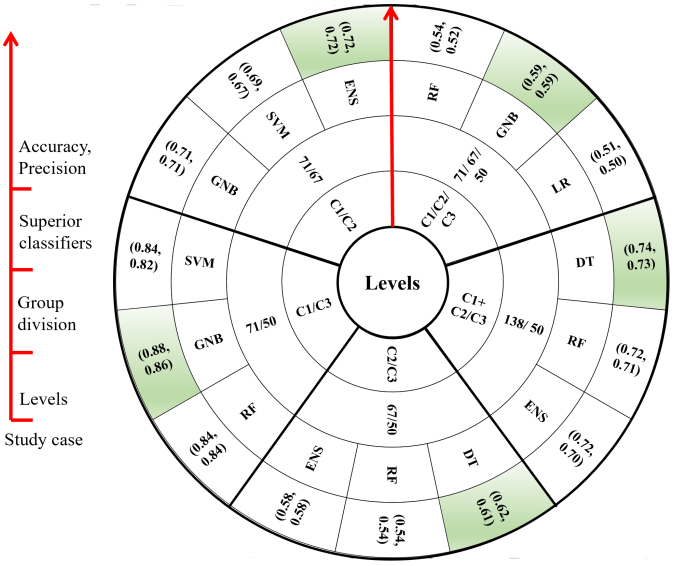

Figure 6: Superior algorithms in classifying three levels of cognitive workload.

was the time pressure, as the highest accuracy percentage was observed from DT classifier at 60%. Moreover, the addition of the secondary task leads to the effective identification of T3 and T4 instances from T1 and T2 with 90% prediction accuracy achieved from k-NN classifier, whereas RF and ENS accuracy rates remaining close enough. On the contrary, the multi-class problem regarding the synchronous classification of all four tasks with the NB classifier, decreased the accuracy to 60%.

The results of our attempt to predict the levels of CW based on the mean scores of the NASA-TLX are presented in Figure 6. Almost 9 out of 10 C1 and C3 examples were classified correctly by the GNB classifier. On the contrary, the identification of C2 from the rest two classes proved particularly challenging for the classifiers. Specifically, C1 and C2 examples were classified correctly from the ENS model at 72% accuracy rate, while the 62% of C2 and C3 cases were predicted properly by the DT classifier.

The DT classifier was proven superior in correctly identifying high CW (C3) from the other two levels based on the NASA-TLX mean score with 74% accuracy.

The last classification problem is related to the classification of three levels of CW; high, medium and low. The GNB was proved to be the most efficient in terms of accuracy reaching up to 59% correct predictions. Overall, GNB and DT models seemed to be able to identify correctly the three levels of CW, high, medium and low, however the insertion of the medium class decreased significantly the accuracy percentage.

6. Discussion and conclusions

In this work, we presented an eye-tracking dataset to be used for the analysis of cognitive workload levels. The dataset comprises of eye and gaze recordings signals from 47 participants, where each participant engaged in visual search related tasks. Each task differs from the others in terms of the existence of time constraint and/or a secondary task and is rated from the participants based on the NASA-TLX workload index.

Our statistical analysis revealed that the Tasks induced different levels of cognitive workload. Both subjective and performance measures reveal that multi tasking and time pressure have induced a higher level of CW than the one induced by single tasking and absence of time pressure. Two-way repeated measures ANOVA showed that multi tasking had a significant effect on 17 eye features while time pressure had a significant effect on 7 eye features. Fixation frequency and pupil diameter seem to be the most sensitive features as they exhibit a significant effect of both multi tasking and time pressure. Fixation frequency decreases in multi tasking and increases with time pressure, while pupil diameter increases both with multi tasking and time pressure (Table 3).

Overall, the highest success rate was observed during the binary classification between T2 and T4, achieving 98% accuracy, while T2 is effectively distinguished also from T4. However, the classification attempts between tasks which both included or not included the secondary task, resulted in considerable loss in model performance, particularly when distinguishing between T3 and T4. The strong effect of backwards counting in model accuracy is confirmed with the effective identification of T1 and T2 from T3 and T4. Additionally, it was challenging for the models to identify separately the four tasks at a satisfactory level.

In terms of estimating cognitive workload levels, both binary and multi-class identification tests produced encouraging results, with up to 88% correct predictions between low and high CW with the GNB classifier. Furthermore, the C2 class had a substantial effect on the models' performance resulting in accuracy decrease. Finally, results indicated that the GNB model emboldens the further investigation of classification between three or more CW levels by the addition of extra number of samples.

We evaluated a range of eye parameters including fixations, saccades, blinks, and pupil size, as well as the capabilities of

numerous machine learning models in a variety of categorization scenarios. Our findings corroborate earlier research and reveal that cognitive workload has an influence on eye movements and pupillary responses. Specifically and in line with the ideas of (7; 8; 9; 10), cognitive workload levels may be recognized successfully using only eye-tracking characteristics. Furthermore, our findings extend beyond prior studies such as (11), revealing considerable advances in terms of accuracy and highlighting the importance of continuing research in this field. Additionally, we established a substantial correlation between ocular characteristics and the four experimental tasks, demonstrating the possibility of developing a cognitive workload detection system with a high degree of discretization capability.

In this work, we presented a dataset comprising of eye movement features gathered as each of the subjects solved visual search puzzles and conducted supplementary tasks, later translated in terms of cognitive workload levels. Despite the considerable contribution of analogous databases to the research community, our proposed dataset includes a much higher sample size and a wider spectrum of eye and gaze metrics, allowing for the examination of their relationships with various cognitive states. Additionally, the dataset is annotated using not only the individuals' NASA RTLX scores, but also the tasks in which they participated.

Although the aforementioned advantages of our work make the dataset an important contribution to the scientific community, the empirical results reported herein should be considered in the light of some limitations. First of all, an analysis of

Table 3: Results of two-way repeated measures ANOVA.

	Tasking		**Time**		**Interaction**
Feature	p	Dif. (multi-single)	p	Dif. (with-without)	p
Fixation frequency	**0.000**	-0.355	**0.000**	0.222	**0.026**
Fixation duration	0.638		**0.000**	-23,365	0.175
Fixation duration Variation	**0.003**	0,045	**0.022**	-0,039	0.208
Saccade frequency	**0.000**	1,624	0.752		0.579
Saccade amplitude Variation	**0.000**	0,118	0.146		0.392
Saccade velocity	**0.000**	125,456	0.648		0.214
Saccade velocity Variation	0.244		**0.024**	-0,072	**0.019**
Saccade velocity Skewness	**0.000**	-1,155	0.561		0.879
Saccade velocity Kurtosis	**0.000**	-4,093	0.885		0.721
Peak saccade velocity	**0.000**	144,184	0.714		0.288
Peak saccade velocity Variation	0.458		**0.016**	-0,059	**0.035**
Peak saccade velocity Skewness	**0.000**	-1,167	0.646		0.792
Peak saccade velocity Kurtosis	**0.000**	-3,648	0.503		0.692
Saccade duration	**0.000**	3,758	0.841		0.938
Saccade duration Variation	**0.000**	0,373	0.448		0.143
Blink frequency	**0.000**	0,201	0.473		0.067
Blink duration	**0.011**	21,901	0.652		0.232
Pupil diameter	**0.000**	0,254	**0.012**	0,106	0.149
Pupil diameter Variation	**0.001**	0,015	0.443		0.930
Pupil diameter Skewness	0.616		**0.042**	-0,196	0.190
Pupil diameter Kurtosis	**0.004**	1,994	0.246		0.100

Two-way repeated measures ANOVA with Bonferroni adjustment. p-value and mean difference of all features between two levels of two factors: Tasking (multi or single) and time pressure (with or without). Statistically significant differences ($p<0.05$) are in bold.

CW levels based not on different tasks but on trials themselves, would result in a much greater sample size, thus providing the opportunity to exploit deep learning methods for CW identification purposes. This was not applied in the current study as the duration of each trial was not long enough to have robust gaze pattern alteration among trials. More complex problem solving could address this issue. Furthermore, it seems that the CW induced by time pressure could have been higher. This could have been achieved using a visible countdown instead of a simple guideline to finish the assignment "as quickly as possible". Finally, as mentioned in Section 4.3, the experimental setup was set in a way so that the image luminance effect on pupil size would be minimized, and this attempt was finally evaluated statistically.

The dataset is made available to the academic community and we firmly encourage other researchers and academics to test their methods and algorithmic approaches on this highly challenging database.

Acknowledgment

This project has received funding from the European Unionâs Horizon 2020 research and innovation program under grant agreement No 826429 (Project: SeeFar). This paper reflects only the authorâs view and the Commission is not responsible for any use that may be made of the information it contains.

References

[1] M. S. Young, K. A. Brookhuis, C. D. Wickens, P. A. Hancock, State of science: mental workload in ergonomics (1 2015). doi:10.1080/00140139.2014.956151.

[2] B. Xie, G. Salvendy, Review and reappraisal of modelling and predicting mental workload in single- and multi-task environments, Work and Stress 14 (1) (2000) 74–99. doi:10.1080/026783700417249.

[3] R. L. Charles, J. Nixon, Measuring mental workload using physiological measures: A systematic review, Applied Ergonomics 74 (2019) 221–232. doi:10.1016/j.apergo.2018.08.028.

[4] B. Cain, A review of the mental workload literature, 2007.

[5] E. Debie, R. Fernandez Rojas, J. Fidock, M. Barlow, K. Kasmarik, S. Anavatti, M. Garratt, H. A. Abbass, Multimodal fusion for objective assessment of cognitive workload: A review, IEEE Transactions on Cybernetics 51 (3) (2021) 1542–1555. doi:10.1109/TCYB.2019.2939399.

[6] Y. Zhou, S. Huang, Z. Xu, P. Wang, X. Wu, D. Zhang, Cognitive workload recognition using eeg signals and machine learning: A review, IEEE Transactions on Cognitive and Developmental Systems (2021) 1–1doi:10.1109/TCDS.2021.3090217.

[7] V. Skaramagkas, E. Ktistakis, D. Manousos, N. S. Tachos, E. Kazantzaki, E. E. Tripoliti, D. I. Fotiadis, M. Tsiknakis, Cognitive workload level estimation based on eye tracking: A machine learning approach, in: 2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE), 2021, pp. 1–5. doi:10.1109/BIBE52308.2021.9635166.

[8] X. Liu, T. Chen, G. Xie, G. Liu, Contact-free cognitive load recognition based on eye movement, Journal of Electrical and Computer Engineering 2016 (2016) 1–8. doi:10.1155/2016/1601879.

[9] G. Prabhakar, A. Mukhopadhyay, L. Murthy, M. Modiksha, D. Sachin, P. Biswas, Cognitive load estimation using ocular parameters in automotive, Transportation Engineering 2 (2020) 100008. doi:10.1016/j.treng.2020.100008.

[10] C. Wu, J. Cha, J. Sulek, T. Zhou, C. Sundaram, J. Wachs, D. Yu, Eye-tracking metrics predict perceived workload in robotic surgical skills training, Human Factors: The Journal of the Human Factors and Ergonomics Society 62 (2019) 001872081987454. doi:10.1177/0018720819874544.

[11] J. Chen, Q. Zhang, L. Cheng, X. Gao, L. Ding, A Cognitive Load Assessment Method Considering Individual Differences in Eye Movement Data, in: IEEE International Conference on Control and Automation, ICCA, Vol. 2019-July, IEEE Computer Society, 2019, pp. 295–300. doi:10.1109/ICCA.2019.8899595.

[12] M. Plechawska, M. Tokovarov, M. Kaczorowska, D. Zapala, A three-class classification of cognitive workload based on eeg spectral data, Applied Sciences 2019, Vol. 9, Page 5340 9 (2019) 5340. doi:10.3390/APP9245340.

[13] V. Skaramagkas, G. Giannakakis, E. Ktistakis, D. Manousos, I. Karatzanis, N. Tachos, E. E. Tripoliti, K. Marias, D. I. Fotiadis, M. Tsiknakis, Review of eye tracking metrics involved in emotional and cognitive processes, IEEE Reviews in Biomedical Engineering (2021) 1–1doi:10.1109/RBME.2021.3066072.

[14] S. G. Hart, L. E. Staveland, Development of nasa-tlx (task load index): Results of empirical and theoretical research, Advances in Psychology 52 (1988) 139–183. doi:10.1016/S0166-4115(08)62386-9.
URL /record/1988-98278-006

[15] E. Ktistakis, V. Skaramagkas, D. Manousos, N. S. Tachos, E. Tripoliti, D. I. Fotiadis, M. Tsiknakis, COLET: A Dataset for Cognitive workLoad estimation based on Eye-Tracking, type: dataset (Jan. 2022). doi:10.5281/ZENODO.5913227.

[16] D. Tao, H. Tan, H. Wang, X. Zhang, X. Qu, T. Zhang, A systematic review of physiological measures of mental workload, International Journal of Environmental Research and Public Health 16 (8 2019). doi:10.3390/ijerph16152716.

[17] G. G. Menekse Dalveren, N. E. Cagiltay, Insights from surgeonsâ eye-movement data in a virtual simulation surgical training environment: effect of experience level and hand conditions, Behaviour and Information Technology 37 (5) (2018) 517–537. doi:10.1080/0144929X.2018.1460399.

[18] X. He, L. Wang, X. Gao, Y. Chen, The eye activity measurement of mental workload based on basic flight task, in: IEEE International Conference on Industrial Informatics (INDIN), 2012, pp. 502–507. doi:10.1109/INDIN.2012.6301203.

[19] H. Sheridan, E. M. Reingold, Chess players' eye movements reveal rapid recognition of complex visual patterns: Evidence from a chess-related visual search task, Journal of Vision 17 (3) (2017) 4–4. doi:10.1167/17.3.4.

[20] R. Mallick, D. Slayback, J. Touryan, A. J. Ries, B. J. Lance, The use of eye metrics to index cognitive workload in video games, in: Proceedings of the 2nd Workshop on Eye Tracking and Visualization, ETVIS 2016, Institute of Electrical and Electronics Engineers Inc., 2017, pp. 60–64. doi:10.1109/ETVIS.2016.7851168.

[21] M. D. Rivecourt, M. N. Kuperus, W. J. Post, L. J. Mulder, Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight, Ergonomics 51 (2008) 1295–1319. doi:10.1080/00140130802120267.
URL https://pubmed.ncbi.nlm.nih.gov/18802817/

[22] H. J. Foy, P. Chapman, Mental workload is reflected in driver behaviour, physiology, eye movements and prefrontal cortex activation, Applied Ergonomics 73 (2018) 90–99. doi:10.1016/j.apergo.2018.06.006.

[23] M. Borys, M. Tokovarov, M. Wawrzyk, K. WesoÅowska, M. Plechawska, R. Dmytruk, M. Kaczorowska, An analysis of eye-tracking and electroencephalography data for cognitive load measurement during arithmetic tasks, Institute of Electrical and Electronics Engineers Inc., 2017, pp. 287–292. doi:10.1109/ATEE.2017.7905130.

[24] T. Kosch, M. Hassib, P. W. WoÂ´zniakwoÂ´zniak, D. Buschek, F. Alt, Your eyes tell: Leveraging smooth pursuit for assessing cognitive workload (2018). doi:10.1145/3173574.3174010.
URL https://doi.org/10.1145/3173574.3174010

[25] I. P. Bodala, Y. Ke, H. Mir, N. V. Thakor, H. Al-Nashash, Cognitive workload estimation due to vague visual stimuli using saccadic eye movements, in: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2014, Institute of Electrical and Electronics Engineers Inc., 2014, pp. 2993–2996. doi:10.1109/EMBC.2014.6944252.

[26] Y. Yang, M. McDonald, P. Zheng, Can drivers' eye movements be used to monitor their performance? A case study, IET Intelligent Transport Systems 6 (4) (2012) 444–452. doi:10.1049/iet-its.2012.0008.

[27] E. Isbilir, M. P. Cakir, C. Acarturk, A. S. Tekerek, Towards a multimodal model of cognitive workload through synchronous optical brain imaging

and eye tracking measures, Frontiers in Human Neuroscience 13 (2019) 375. doi:10.3389/fnhum.2019.00375.

[28] R. Contreras, J. Ghajar, S. Bahar, M. Suh, Effect of cognitive load on eye-target synchronization during smooth pursuit eye movement, Brain Research 1398 (2011) 55–63. doi:10.1016/j.brainres.2011.05.004.

[29] E. Siegenthaler, F. M. Costela, M. B. Mccamy, L. L. D. Stasi, J. Otero-Millan, A. Sonderegger, R. Groner, S. Macknik, S. Martinez-Conde, Task difficulty in mental arithmetic affects microsaccadic rates and magnitudes, European Journal of Neuroscience 39 (2014) 287–294. doi:10.1111/ejn.12395.

[30] X. Gao, H. Yan, H.-j. Sun, Modulation of microsaccade rate by task difficulty revealed through between- and within-trial comparisons, Journal of Vision 15 (3) (2015) 3–3. doi:10.1167/15.3.3.

[31] K. Krejtz, A. T. Duchowski, A. Niedzielska, C. Biele, I. Krejtz, Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze, PLoS ONE 13 (9 2018). doi:10.1371/journal.pone.0203629.

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

[33] M. I. Ahmad, I. Keller, D. A. Robb, K. S. Lohan, A framework to estimate cognitive load using physiological data, Personal and Ubiquitous Computing (2020) 1–15doi:10.1007/s00779-020-01455-7.

[34] R. Bednarik, J. Koskinen, H. Vrzakova, P. Bartczak, A. P. Elomaa, Blink-Based Estimation of Suturing Task Workload and Expertise in Microsurgery, in: Proceedings - IEEE Symposium on Computer-Based Medical Systems, Vol. 2018-June, Institute of Electrical and Electronics Engineers Inc., 2018, pp. 233–238. doi:10.1109/CBMS.2018.00048.

[35] T. Čegovnik, K. Stojmenova, G. Jakus, J. Sodnik, An analysis of the suitability of a low-cost eye tracker for assessing the cognitive load of drivers, Applied Ergonomics 68 (2018) 1–11. doi:10.1016/j.apergo.2017.10.011.

[36] M. Pomplun, S. Sunkara, Pupil Dilation as an Indicator of Cognitive Workload in Human-Computer Interaction, 2003.

[37] E. H. Hess, J. M. Polt, Pupil size as related to interest value of visual stimuli, Science (1960) 349–350doi:10.1126/science.132.3423.349.

[38] O. Palinko, A. L. Kun, A. Shyrokov, P. Heeman, Estimating cognitive load using remote eye tracking in a driving simulator, in: Eye Tracking Research and Applications Symposium (ETRA), 2010, pp. 141–144. doi:10.1145/1743666.1743701.

[39] W. Soussou, M. Rooksby, C. Forty, J. Weatherhead, S. Marshall, EEG and eye-tracking based measures for enhanced training, in: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 2012, pp. 1623–1626. doi:10.1109/EMBC.2012.6346256.

[40] M. Nakayama, Y. Hayakawa, Relationships between Oculo-Motor Mesures as Task-evoked Mental Workloads during an Manipulation Task, in: Proceedings of the International Conference on Information Visualisation, Vol. 2019-July, Institute of Electrical and Electronics Engineers Inc., 2019, pp. 170–174. doi:10.1109/IV.2019.00037.

[41] A. Bayat, M. Pomplun, The influence of text difficulty level and topic on eye-movement behavior and pupil size during reading, in: Proceedings - 2016 2nd International Conference of Signal Processing and Intelligent Systems, ICSPIS 2016, Institute of Electrical and Electronics Engineers Inc., 2017. doi:10.1109/ICSPIS.2016.7869898.

[42] S. Nikolopoulos, K. Georgiadis, F. Kalaganis, G. Liaros, I. Lazarou, K. Adam, A. Papazoglou-Chalikias, E. Chatzilari, V. Oikonomou, P. Petrantonakis, I. Kompatsiaris, C. Kumar, R. Menges, S. Staab, D. Müller, K. Sengupta, S. Bostantjopoulou, Z. Katsarou, G. Zeilig, M. Plotnik, A. Gotlieb, S. Fountoukidou, J. Ham, D. Athanasiou, A. Mariakaki, D. Comandicci, E. Sabatini, W. Nistico, M. Plank, A Multimodal dataset for authoring and editing multimedia content: The MAMEM project, Data in Brief (2017). doi:10.5281/zenodo.834154.

[43] Y. Li, M. Liu, J. M. Rehg, In the Eye of Beholder: Joint Learning of Gaze and Actions in First Person Video, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 11209 LNCS, Springer Verlag, 2018, pp. 639–655. doi:10.1007/978-3-030-01228-1_38.

[44] A. Fathi, Y. Li, J. M. Rehg, Learning to recognize daily actions using gaze, in: Lecture Notes in Computer Science, Vol. 7572, Springer, Berlin, Heidelberg, 2012, pp. 314–327. doi:10.1007/978-3-642-33718-5_23.

[45] R. Carmi, L. Itti, The role of memory in guiding attention during natural vision, Journal of Vision 6 (9) (2006) 4–4. doi:10.1167/6.9.4.
URL `https://doi.org/10.1167/6.9.4`

[46] K. A. Ehinger, B. Hidalgo-Sotelo, A. Torralba, A. Oliva, Modelling search for people in 900 scenes: A combined source model of eye guidance, Visual Cognition 17 (6-7) (2009) 945–978, pMID: 20011676. doi:10.1080/13506280902834720.

[47] A. Quattoni, A. Torralba, Recognizing indoor scenes (2010) 413–420doi:10.1109/CVPR.2009.5206537.

[48] S. Plainis, Y. Orphanos, M. K. Tsilimbaris, A modified etdrs visual acuity chart for european-wide use, Optom Vis Sci. 84 (2007) 647–653.

[49] B. Xie, G. Salvendy, Review and reappraisal of modelling and predicting mental workload in single- and multi-task environments, Work and Stress 14 (2000) 74–99. doi:10.1080/026783700417249.

[50] J. C. Byers, A. Bittner, S. Hill, Traditional and raw task load index (tlx) correlations: Are paired comparisons necessary?, Taylor Francis, 1989, pp. 481–485.

[51] K. L. Young, A. N. Stephens, K. L. Stephan, G. Stuart, An examination of the effect of google glass on simulated lane keeping performance, undefined 3 (2015) 3184–3191. doi:10.1016/J.PROMFG.2015.07.868.

[52] L. D. J. Shiber, D. N. Ginn, A. Jan, J. T. Gaskins, S. M. Biscette, R. Pasic, Comparison of industry-leading energy devices for use in gynecologic laparoscopy: Articulating enseal versus ligasure energy devices, Journal of Minimally Invasive Gynecology 25 (2018) 467–473.e1. doi:10.1016/J.JMIG.2017.10.006.

[53] M. Georgsson, Nasa rtlx as a novel assessment tool for determining cognitive load and user acceptance of expert and user-based usability evaluation methods, European Journal of Biomedical Informatics 16 (2020). doi:10.24105/EJBI.2020.16.2.14.

[54] J. Townsend, G. Ashby, Methods of modeling capacity in simple processing systems, Cognitive theory 3 (1978).

[55] R. Bruyer, M. Brysbaert, Combining speed and accuracy in cognitive psychology: Is the inverse efficiency score (ies) a better dependent variable than the mean reaction time (rt) and the percentage of errors (pe)?, Psychologica Belgica 51 (2011) 5–13. doi:10.5334/PB-51-1-5.

[56] S. Chartier, P. Renaud, An online noise filter for eye-tracker data recorded in a virtual environment, 2008, pp. 153–156. doi:10.1145/1344471.1344511.

[57] A. T. Duchowski, Eye Movement Analysis, Springer London, London, 2003, pp. 111–128. doi:10.1007/978-1-4471-3750-4$_9$.
URL

ELSEVIER

Computer Methods and Programs in Biomedicine 00 (2022) 1–13

cmpb_logo.PNG

www.elsevier.com/cmpb

COLET: A Dataset for COgnitive workLoad estimation based on Eye-Tracking

Emmanouil Ktistakis[a,b,1,*], Vasileios Skaramagkas[a,1], Dimitris Manousos[a], Nikolaos S. Tachos[c], Evanthia Tripoliti[d], Dimitrios I. Fotiadis[c], Manolis Tsiknakis[a,e]

[a] *Institute of Computer Science, Foundation for Research and Technology Hellas (FORTH), GR-700 13 Heraklion, Greece*
[b] *Laboratory of Optics and Vision, School of Medicine, University of Crete, GR-710 03 Heraklion, Greece*
[c] *Biomedical Research Institute, FORTH, GR-451 10, Ioannina, Greece and the Dept. of Materials Science and Engineering, Unit of Medical Technology and Intelligent Information Systems, University of Ioannina, GR-451 10, Ioannina, Greece*
[d] *Dept. of Materials Science and Engineering, Unit of Medical Technology and Intelligent Information Systems, University of Ioannina, GR-451 10, Ioannina, Greece*
[e] *Dept. of Electrical and Computer Engineering, Hellenic Mediterranean University, GR-710 04 Heraklion, Crete, Greece*

Abstract

The cognitive workload is an important component in performance psychology, ergonomics, and human factors. Unfortunately, publicly available datasets are scarce, making it difficult to establish new approaches and comparative studies. In this work, COLET-COgnitive workLoad estimation based on Eye-Tracking dataset is presented. Forty-seven (47) individuals' eye movements were monitored as they solved puzzles involving visual search tasks of varying complexity and duration. The authors give an in-depth study of the participants' performance during the experiments while eye and gaze features were derived from low-level eye recorded metrics, and their relationships with the experiment tasks were investigated. The results from the classification of cognitive workload levels solely based on eye data, by employing and testing a set of machine learning algorithms are also provided. The dataset is available to the academic community.

Keywords:
Cognitive workload, Workload classification, Eye movements, Machine Learning, Eye-tracking, Affective computing

*Corresponding author
URL: mankti2@gmail.com (Emmanouil Ktistakis)
[1] Authors contributed equally