Using eye tracking to distinguish between different levels of cognitive workload

Karen Dijkstra - 0722049

Bachelor Thesis
Author: Karen Dijkstra
Student number: 0722049
Email: karendijkstra@student.ru.nl
Supervisor: Louis Vuurpijl
Radboud University Nijmegen

***Abstract*—Augmented cognition is a relatively new field in Human-Computer Interaction that aims for the development of systems that can detect the user's cognitive state in real-time and consequently adapt the system appropriately. Such an adaptation could improve the effectiveness of human-computer interfaces.**

In this thesis a method is developed to distinguish between multiple levels of cognitive workload using eye-behavior features. These features are based on features that are known to be distinctive when used to distinguish between other cognitive states [1]. The data are obtained using an eye tracker.

The results indicate that this method is task dependent to a degree, but these results are not conclusive. On the other hand, it is shown that this method *can* distinguish between different levels of cognitive workload within a task and that this measure is subject-independent.

I. Introduction

Human-Computer Interaction (HCI) focuses on improving the interaction between computers and humans by developing more effective modes of communication. An important component of user-interface design is to accommodate the needs of the user. Specific needs vary between users but can also vary for a user over time. There is a number of factors that affect perceptual and motor performance which include, but are not limited to:

- Arousal and vigilance
- Fatigue and sleep deprivation
- Perceptual (mental) load
- Monotony and boredom
- Fear, anxiety, mood and emotion

Ideally the interface would be adapted according to such changes in the user. Augmented Cognition is a relatively new sub field of HCI that aims to do this by accessing the internal state of the user almost directly:

> The goal of Augmented Cognition research is to create revolutionary human-computer interactions that capitalize on recent advances in the fields of Neuroscience, cognitive science and computer science. Augmented Cognition can be distinguished from its predecessors by the focus on the real-time cognitive state of the user, as assessed through modern neuroscientific tools. At its core, an Augmented Cognition system is a 'closed loop' in which the cognitive state of the operator is detected in real time with a resulting compensatory adaptation in the computational system, as appropriate. [2]

The factors mentioned above all affect the cognitive state of the user, and could possibly be measured by such neuroscientific tools. Such modern neuroscientific tools can consist of brain measures such as EEG or fNIR [3], [4], [5]. An alternative prominent approach uses data from eye trackers to distinguish between cognitive states [1]. Furthermore, in some cases a whole suite of physiological sensors is used for cognitive state evaluations [6].

A. Cognitive states

Every day people go through different cognitive states. At any given time, someone is happy or sad, relaxed or alert, distracted or focused or somewhere in between. People use these labels to describe themselves, as well as others: 'You look sad, are you ok?'

The cognitive states that carry these labels often manifest themselves in physical behavior. These are the cues that humans use to assign such states and these cues and others, can be used to allow computers to do the same. If such systems then adapt the interaction appropriately, on the basis of these states, this could increase the effectiveness of the interaction.

While there are many reasons why measuring someone's cognitive states can be useful and interesting, an obvious application is systems that can be used in the workplace to increase safety. People that work in aviation or other critical safety fields can carry jobs with high responsibility, where mistakes have large consequences. Measuring the cognitive states of such people could allow for the task of the individual to be adapted in a way that reduces the risk of mistakes occurring.

B. Research goals

The focus of this thesis will be on cognitive workload. Cognitive workload is especially relevant in the context of the workplace. A job consists of several tasks strung together; tasks impose a certain amount of cognitive workload on a user, but this amount of workload is not static and can change, even if the objective task load would be considered constant. This could, for example, be caused by a change in strategy by the user; a more effective strategy would probably require less cognitive workload.

When the amount of cognitive workload for performing a task could be determined automatically, a task could be adapted if deemed necessary. Tasks could be redistributed among co-workers to ensure that no one is overloaded, and someone that is underloaded could receive extra work. Such a system fits within the goal of augmented cognition.

So if techniques can be developed to determine different levels of cognitive workload, such knowledge could contribute to useful applications. There are various examples that pursue this idea. For example, Brain Computer Interfaces have successfully been used to distinguish between different levels of cognitive workload and even between different types of workload [4], [3], [5].

Eye tracking can also be used distinguish between different cognitive states. S. Marshall used seven features extracted from eye tracker data to classify between three different types of cognitive states (relaxed vs. engaged, focused vs. distracted and alert vs. fatigued) [1]. For this thesis I would like to investigate if these seven features from S. Marshall can be used to distinguish between low and higher levels of cognitive workload as well.

This prompts the following research questions:

- Can a method be developed that employs the same features (i.e. those used by S. Marshall) extracted from eye behavior data, to distinguish between different levels of cognitive workload [1]?
- If so: Is this method task dependent?
- And: Is this method subject dependent?

C. *Organization of this thesis*

The structure of this thesis is as follows: The next section describe relevant research findings on eye data features and subjective workload measures. In Section III I will describe the experiment that was performed in detail. In Section IV is described how the data obtained from the experiment was processed to generate the classification results. These results will then be shown in Section V. In the last section (Section VI) these results will be discussed and I will explain how they relate to the original research questions. Furthermore, any implications for future research will be discussed in this section as well.

II. Background

A. *Eye tracking*

Seven features, extracted from eye tracker data, were used by S. Marshall to distinguish between several mental states [1]. Eye tracker data from three experiments were analyzed; for each of the experiments two cognitive states were distinguished from each other. These were respectively: relaxed vs. engaged, distracted vs. focused and fatigued vs. alert.

In the first experiment, in which she distinguished between a relaxed vs. an engaged cognitive state, subjects either performed no task, or a mental arithmetic task. This mental arithmetic task was presented to them orally, so no visual stimuli were used in this experiment. During each of the conditions the subjects were asked to keep looking at the (empty) computer screen, so eye tracking could be performed.

In the second experiment, in which she distinguished between a focused vs. a distracted cognitive state, subjects drove a driving simulator in both conditions. In the distracted condition though, a separate task was presented orally as a distractor. For this distractor task different tasks were used such as solving simple arithmetic problem, or counting backwards.

In the third experiment, in which she distinguished between a fatigued vs. an alert cognitive state, subjects' eyes were tracked during three different tasks, over several sessions. After each session the fatigue level was measured. Based on these data the alert and fatigued sessions were separated, to define which sessions were marked as the alert condition and which as fatigued condition.

For these experiments a head mounted type of eye tracker was used. The data recorded by the eye tracker consisted of the diameter of the pupil of each eye and the x and y locations of the gaze of each eye. A short description of the seven features that were derived from these data can be found in Table I.

TABLE I
Short descriptions of the features described in S. Marshalls article (Paraphrased) [1]

Feature	Measure
Blinks (Left, Right)	The proportion of a second the eye is blinking
Movement (Left, Right)	The proportion of a second the eye is moving
Divergence	The distance between the horizontal gaze positions of both eyes averaged over a second
Index of Cognitive Activity (ICA) (Left, Right)	The number of changes in pupil diameter caused by cognitive activity per second

Since in this thesis I use these seven features to distinguish between different levels of cognitive load, I will discuss each of the features in more detail. On top of that is discussed what could be expected from each of the features in relation to cognitive workload.

1) The Blink feature: Blinks can be distinguished into three types: reflexive, voluntary and endogenous. Reflexive blinks occur in a response to external stimuli, designed to protect the eye and are more or less involuntary. Voluntary blinks are invoked by the person in question voluntarily. Blinks that occur in absence of any physical stimulus or intent are called endogenous blinks [7]. These blinks are influenced by perceptual and information processing, however when more visual attention is required by a task, the endogenous eye blinks are inhibited and delayed to a moment where the visual demand is reduced.

The blink feature measures the proportion of a second that the eye is blinking, once again two values are calculated: one for each eye.

The above mentioned effects give reason to expect the number of blinks to be lower in conditions with higher cognitive workload, in tasks that are visually demanding. At the same time the number of blinks could actually be larger in conditions where the cognitive workload is imposed by non visual sources, making this feature rather task dependent.

This latter effect can actually be observed, in the research of S. Marshall [1]. When distinguishing between a relaxed vs. an engaged cognitive state without any visual stimuli, more blinks occur during the engaged condition, in which mental arithmetic was performed (left: 0.080 and 0.140 respectively; right 0.090 and 0.150 respectively). In this condition more endogenous eye blinks are likely occurring, which are not being inhibited by visual attention.

On top of this, research indicates that the number of blinks increases with the Time on Task, a phenomenon already documented in the 1940's, see for example [8]. This could interfere with the already mentioned effects.

Because of these conflicting effects it is hard to predict the relationship between blinks and cognitive workload, which will depend on the type of task.

2) The Movement feature: Eye behavior can be divided into three types: The eye is blinking, the eye is fixating or the eye is moving. The movement feature is designed to measure the proportion of a second that the eye is moving.

This feature is influenced by the amount of blinks and fixations that are occurring and their respective durations, since when an eye is blinking, it cannot also be moving or fixating. Because of this it is hard to create a specific hypothesis of how this feature would correlate with different levels of cognitive workload: relationships between eye blinks and cognitive workload, and relationships between eye fixations and cognitive workload all play a role.

Some information can be gained by looking at the first experiment by S. Marshall, since the same feature was used. During the relaxed condition of this experiment, the movement feature average was lower than during the engaged condition (left: 0.621 and 0.550 respectively; right 0.604 and 0.522 respectively) [1].

This could indicate that there is a negative relationship between cognitive workload and the movement feature. Since these are only data from one task, it is hard to conclude anything about task dependency.

3) The Divergence feature: Divergence in this case refers to the distance between the horizontal gaze locations of each of the eyes. For the divergence feature the distances between the horizontal gaze locations on the screen are averaged over a second.

When focusing on something visually, the eyes adjust their vergence until the gaze locations of the eyes are very close or overlap. Therefore during conditions that require a lot of visual attention the divergence of the gaze locations would be expected to be small. This is made evident by the results of S. Marshall: during the experiment designed to distinguish between a focused vs. a distracted cognitive state, in which participants used a driving simulator, the mean divergence was much lower (0.049 and 0.041 respectively) than during the experiment that distinguished between a relaxed vs. an engaged cognitive state (0.146 and 0.192 respectively), in which participants were offered no visual stimuli [1].

Visual demand can certainly be a cause of cognitive workload, which would result in a lower divergence, but cognitive workload can also be caused by non visual sources. When looking at the results from the first experiment by S. Marshall again: in the engaged condition, while performing mental arithmetic, the divergence was actually higher than in the relaxed condition.

This indicates that there is no straightforward relationship between the divergence feature and cognitive workload: it depends mostly on the task that is executed.

4) The Index of Cognitive Activity feature: Behavior of the pupil size is controlled by several muscles in the eye and can be caused by different stimuli. Light is the most common cause of pupil dilation or constriction: more light and the pupil constricts, less light and the pupil dilates. Cognitive processing can also cause the pupil to dilate, this is referred to as the dilation reflex [9].

This phenomenon is very relevant for the current research. The issue that arises is how to differentiate between the light and the dilation reflex in data. S. Marshall has developed and patented a technique that employs wavelet theory to do precisely this: the Index of Cognitive Activity (ICA) [10].

Since the pupil diameter of both eyes is measured, the ICA feature results in two values, one for the left and one for the right eye.

The method to extract the ICA feature uses wavelets to extract the relevant changes in pupil diameter and to ignore the changes that are caused by fluctuations in light. This method will be explained in more detail in Section IV-B6. The feature values that are obtained consist of the number of times per second that an increase was measured by this method. These values usually range between 0-20 per second, where a higher index would be expected in conditions where the cognitive workload is higher.

This is also demonstrated by S. Marshalls research: in her first experiment where she distinguished between a relaxed vs. an engaged cognitive state, the mean ICA was higher in the engaged state (left: 0.144 and 0.280, respectively; right: 0.158 and 0.324, respectively) [1].

Both (linear) discriminant functions and (non-linear) neural networks were used to classify the data. The results across all three experiments and across both methods of classification ranged between 69% and 92%. Even though there is no distinction made between more than two levels of a given state, these results do at least show that a distinction can be made.

B. *Cognitive Workload*

Based on the results of S. Marshall's research, it is expected that different levels of cognitive workload can be distinguished from each other using this method: low levels vs. high levels at the least.

For her first experiment, subjects' eyes were tracked in two conditions: one where they relaxed and performed no task and a second one where they performed mental arithmetic. Her interpretation of the experiment classified the distinguished cognitive states as relaxed vs. engaged. The experiment could also be interpreted as containing conditions, in which in one no cognitive workload was imposed and in the other some cognitive workload was imposed in the form of mental arithmetic.

For this thesis I aim for an experiment where in all conditions at least some cognitive workload is imposed. To impose this workload, tasks will need to be defined in such a way that

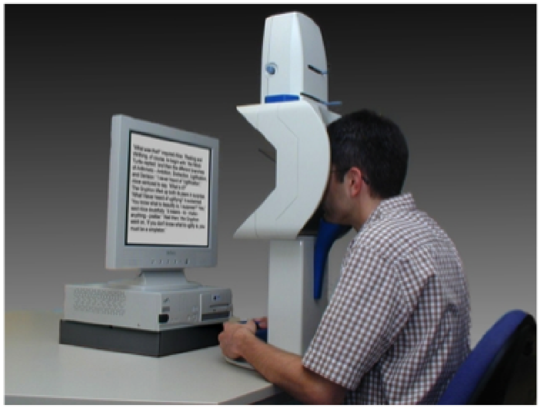


Fig. 1. The iView X'Hi-speed: a desk mounted eye tracker. During use the subject rests their chin on the support, to minimize head movements.

it can be verified that what is measured, is in fact cognitive workload.

To this end, several kinds of subjective workload questionnaires exist that have been developed in the past to measure a subject's cognitive workload. In an article by Rubio et al. three of such workload measures are compared to each other [11]. These are the NASA Task Load Index (NASA-TLX), Subjective Workload Assessment Technique (SWAT), and the Workload Profile [12], [13], [14].

To compare these subjective workload measures, Rubio et al. present two different tasks, at each two different difficulty levels, to the subjects, over 8 different conditions. Four of these conditions consist of single task conditions varied over the difficulty levels. The other four conditions are dual task conditions, which are varied over all possible combinations of the difficulties.

Their results show that each of the workload measures have their specific strengths. The Workload Profile performs best when it comes to differentiating between the eight conditions. This is reflected in the recommendations given in their conclusions:

> If the goal is a comparison between the mental workload of two or more tasks with different objective levels of difficulty, then the assessor should choose the Workload Profile. [11]

In this research, the tasks as described by Rubio et al. will be replicated and subjects will be asked to fill in the Workload Profile upon completion of the experiment. The results of the Workload Profile will be analyzed to assess if different levels of cognitive workload were indeed imposed on the subjects.

III. Methods

A. Equipment

The eye tracker used for this experiment is an iView X'Hi-Speed (see Fig. 1). It is a desk mounted system, on which users can rest their chin so that their head remains still. This is different from the head-mounted type of eye tracker, the EyeLink, as used by S. Marshall [1]. An advantage of the iView system is that it has a high sampling rate (500 Hz vs. 250 Hz for the EyeLink).

For real life scenarios, less obtrusive set ups are required, where users can move their head freely, as they can in normal working conditions. However, for the purpose of this research, the high temporal and spatial resolutions of the iView system are important. Furthermore, this set up allows for well-calibrated and uniform recording conditions, which is essential for the research sketched below.

The eye-tracker is connected to a pc that runs windows XP, this pc in turn is connected through a wireless dongle to another pc that runs the iView software. This other pc records the eye data. The recorded eye data consist of the pupil diameter of each eye and the screen coordinates of the gaze location. The raw data produced by the eye tracker will be discussed in more detail in section 4.

B. Experiment

As mentioned, the design of this experiment is inspired by the experiment from Rubio et al. [11]. It consists of two different types of tasks: Sternberg's Memory Searching task, from here on referred to as the Sternberg task, and a tracking task. These are also combined resulting in a third: a dual task. Each of these consist of two difficulty levels: easy and hard (6 conditions in total).

1) The Sternberg Task: The subjects are given a set of letters that they are told to remember. Subsequently the trial starts and they are prompted with random letters and asked whether or not this letter occurred in the original set. The subject uses the keyboard to answer 'Yes' or 'No'. As soon as the answer is given there is a pause, after which the next letter is shown. This repeats until the end of the trial. For the next trial the subject is first provided with a new set of letters to memorize, once this is done the new trial starts.

The difficulty of the task is modified through the number of letters that the subject has to remember: two or four.

2) The Tracking Task: A path is shown on the screen, together with a cursor. The subjects control the cursor and are told to follow the path with their cursor for as long as the trial lasts. The path curves left and right over the screen and the difficulty of the task is modified through steepness and size of the curves. The harder conditions require more and faster movements of the cursor.

The cursor is controlled by the mouse.

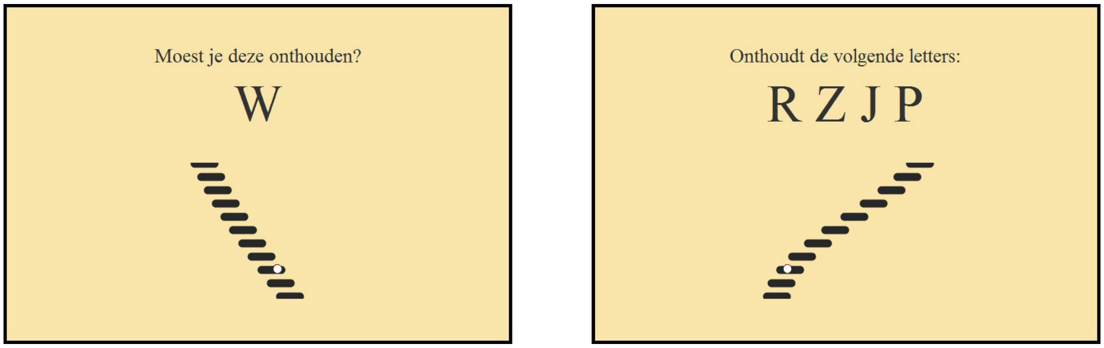
3) The Dual Task: For this task both the Tracking task and the Sternberg task are combined. The Sternberg task is displayed at the top of the screen and the Tracking task is positioned underneath. Since only one task requires the mouse and the other the keyboard, subjects can perform these tasks simultaneously, by using both hands.

Subjects were instructed to use the left hand to control the keyboard and the right hand for the mouse, for all conditions. None of the participants were left-handed.

For a visual of these tasks see Fig. 2.

C. Workload Questionnaire

The Workload Profile is recommended by Rubio et al. if the goal is to assess the relative workload of different types of tasks [11].



Moest je deze onthouden?

W

Onthoudt de volgende letters:

R Z J P

Fig. 2. Screen captures from the experiment. Before the trial started. the subject was prompted with a string of letters to remember for the Sternberg task. The path for the Tracking task remained frozen until trial start. This can be seen on the left. During a trial, the subject was prompted with a letter for the Sternberg task. The path would then be moving and the subject should follow it with the cursor. This can be seen on the right

TABLE II
WORKLOAD DIMENSION DISTINCTIONS USED IN THE WORKLOAD PROFILE.

Workload Dimensions	
Stages of Processing	Perceptual Response
Code of Processing	Spatial Verbal
Input	Visual ~~Auditory~~
Output	Manual ~~Speech~~

TABLE III
PARAMETERS OF THE EXPERIMENT, AS DEFINED AFTER THE PILOT.

Experiment details	
# conditions	6
# trials	42 (7*6)
trial duration	40 sec (1^{st} trial: 15 sec)
calibration time	+- 20 min
Workload Profile	+- 20 min
Experiment	+- 40 min

The subjects were asked to rate their experience on the Workload Profile, after they had completed the experiment.

The Workload Profile consists of four dimensions that distinguish between different types of cognitive workload. Each of these dimensions consist of two sub dimensions. These are shown in Table II. A subject is asked to rate each of the sub dimensions for a condition, relative to other conditions, which are subsequently added together to obtain the Workload profile rating of that condition. Two of the sub dimensions did not apply to the current experiment.

The instructions and examples as given by Tsang and Velazquez (the developers of the questionnaire) were provided to the subjects to help them fill it out [14].

D. Pilot

A small pilot (1 subject) was performed prior to the experiment to identify possible issues with the software and to detect any issues with the experiment itself.

As a result of this pilot, several small bugs were fixed. Furthermore both the number of trials and the length of the trials were reduced due to the subject's reports. This reduced the amount of data gathered per subject, but fatigue in subjects can affect the results and longer experiment times make subjects less willing to participate.

The final parameters that were defined for the experiment can be seen in Table III.

E. Procedure

There were six participants, all male students: ages 20-25. Two subjects wore glasses, which pose no problems for the eye tracking system used. The subjects each sat through one session. At the start of the session the eye-tracker was adjusted to them and subsequently calibrated. They were asked to stay seated for the rest of the experiment, barring a break halfway through the experiment in which they were permitted to get up and stretch their legs. After this break the eye-tracker was re-calibrated.

The experiment lasted on average between an hour and an hour and a half, including the calibration time and the time required to fill out the Workload Profile. About 20 minutes of usable data were recorded per subject. To avoid measuring fatigue rather than workload, the order of the conditions was counterbalanced, using a Latin Square design. Since there were both six participants and six conditions, each of the participants was subjected to a unique order of conditions.

Furthermore, The first trial of each condition only lasted for 15 seconds, instead of 40 seconds, so subjects could get used to the tasks. The data from these trials did not significantly differ from the rest of the data and were therefore included in data analysis.

IV. DATA ANALYSIS

The raw data acquired from the eye tracker are processed in several different steps, until eventually the results of the classification are available. In Fig. 3 the set up of the processing pipeline is depicted. Each of these steps will be discussed in detail, starting with the preprocessing of the raw data. All of this processing was performed in Matlab.

A. Preprocessing

The eye tracking software records the data at a sampling rate of 500hz. The resulting data file consists of the header lines,

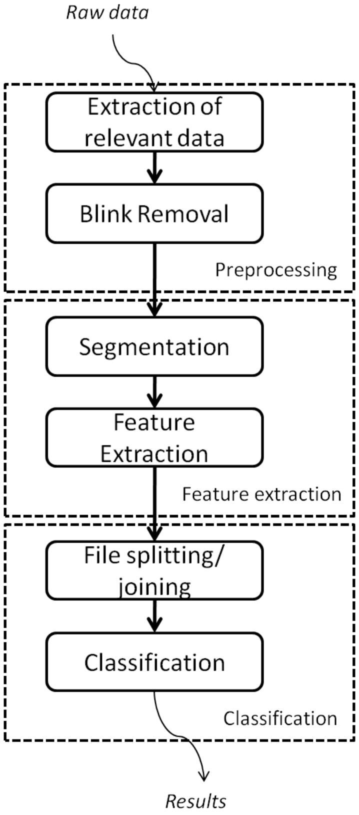


Fig. 3. The data processing pipeline. The analysis of the data is divided in several stages: the preprocessing of the data, the extraction of features of the data and the classification of the feature data. These stages are explained in detail in Chapter IV.

TABLE IV
DESCRIPTIONS OF THE RELEVANT TYPES OF DATA PRODUCED BY THE EYE TRACKING SYSTEM. AN EXAMPLE OF THE DATA OUTPUT FROM THE EYE TRACKER CAN BE FOUND IN APPENDIX B.

Data type	Description
Pupil location	The location of the pupil on the eye-tracker camera image, for each eye the x and y location
Pupil diameter	The area and diameter of the pupil on the camera for both eyes.
Corneal reflection location	The location of the corneal reflection on the eye tracker camera image, for each eye the x and y location
Gaze location	For each eye the x and y location on the computer screen (requires calibration)

describing the settings used for the eye tracking; the column headers and the recorded data. It also contains any messages sent to the recording pc. These messages mark events, such as trial starts and trial ends.

Any line after the column headers contains both a timestamp of when the line was recorded and a type that reflects whether the line contains data or a message (‘SMP’ or ‘MSG’). The data lines contain 16 columns of data. An example of output produced by the eye tracking system can be found in Appendix B.

These columns contain 4 different types of data as shown in Table IV.

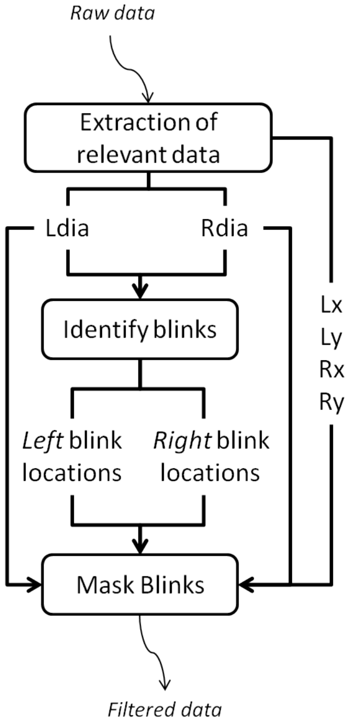


Fig. 4. Preprocessing schema. From the raw data obtained from the eye tracker the relevant data vectors are extracted: Ldia, Rdia, Lx, Ly, Rx, and Ry. The Ldia and Rdia are used to identify the blinks. These blinks locations are saved and used to mask the blinks in these data vectors, i.e., the blink locations are replaced by NaNs (Not a Number) in each of the vectors. These stages are explained in detail in Chapter IV-A.

TABLE V
RAW DATA COLUMNS RELEVANT FOR FEATURE EXTRACTION AND WHAT THEY WILL BE REFERRED TO AS.

Column	Name
Left pupil, diameter	Ldia
Right pupil, diameter	Rdia
Left eye, X location	Lx
Left eye, Y location	Ly
Right eye, X location	Rx
Right eye, Y location	Ry

These data require preprocessing before the features can be extracted. In Fig. 4 the several steps taken to preprocess the data are depicted.

1) Extraction of relevant data: Of these four types, only the gaze location and the pupil diameter are required for feature extraction. All columns that are not needed can be removed; this also goes for any lines of data that do not belong to a specific trial (i.e., the breaks between trials). These lines are identified by the messages that denote the trial start and end. These messages also contain the information needed to label the data lines that do belong to a trial, with a condition and trial label. After labeling these messages are also removed and only the data relevant for feature extraction remain. These data consist of 8 columns(see Table V).

2) Blink removal: Once the relevant data have been extracted, the blinks need to be removed. When the eye closes for a blink the eye tracker becomes unable to track both the pupil and the corneal reflex until the eye opens again. Therefore all

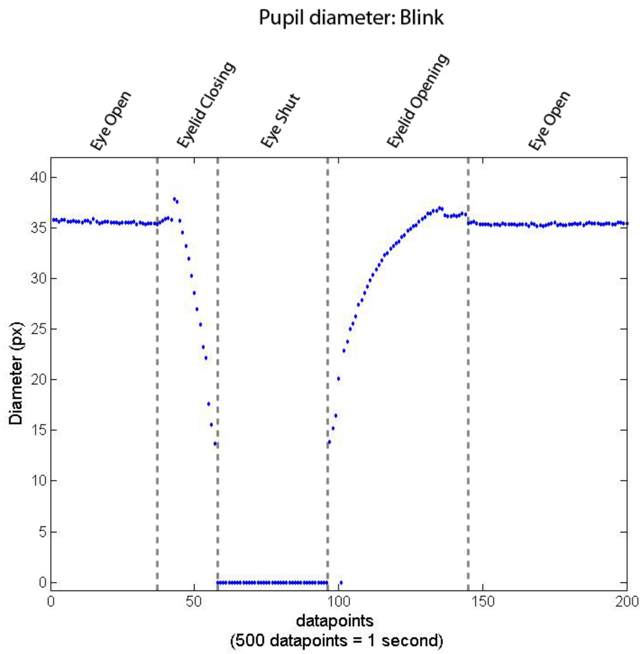


Fig. 5. Pupil diameter during a blink (500 data points equal one second)

the data columns for that eye will report zeros for that period of time. For the short period before and after this, when the eyelid still covers the eye partially, the data the eye tracker records is unreliable.

The moment that the eyelid starts to occlude the eye, up until the moment that the eyelid has cleared out of the way again, is considered a blink. A visual representation of how these blinks appear in pupil diameter data can be observed in Fig. 5.

You can see clearly here how the diameter starts to drop once the eyelid starts occluding the pupil. Once the eyelid opens again, the pupil diameter increases until it is back to normal.

These blinks need to be removed from the data entirely so that each of the features can be calculated accurately. This is achieved by locating the blinks in the data and replacing or ‘masking’ them for each data column with NaNs (Not a Number).

These blinks are located in the data by setting a threshold for the pupil diameter; every data point that falls underneath the threshold is marked as a blink. The goal of setting a threshold is to catch as many points that are part of the blink as possible.

The problem is that the blink does not show up in the data points only as periods of zero pupil diameter, but also as periods of non-zero diameter: the moments in which the eyes are closing and opening again. The threshold creation involves a trade-off; the goal is catching as many blinking data points as possible, while still making sure that no other data points are clipped.

Because the average pupil diameter varies over time and not just per subject, one threshold per subject is not sufficient. Instead the threshold is recalculated for each trial.

The blink threshold is determined by making a histogram of the diameter data. It can be best explained by looking at the histogram and pupil diameter graph, which are depicted in Fig. 6

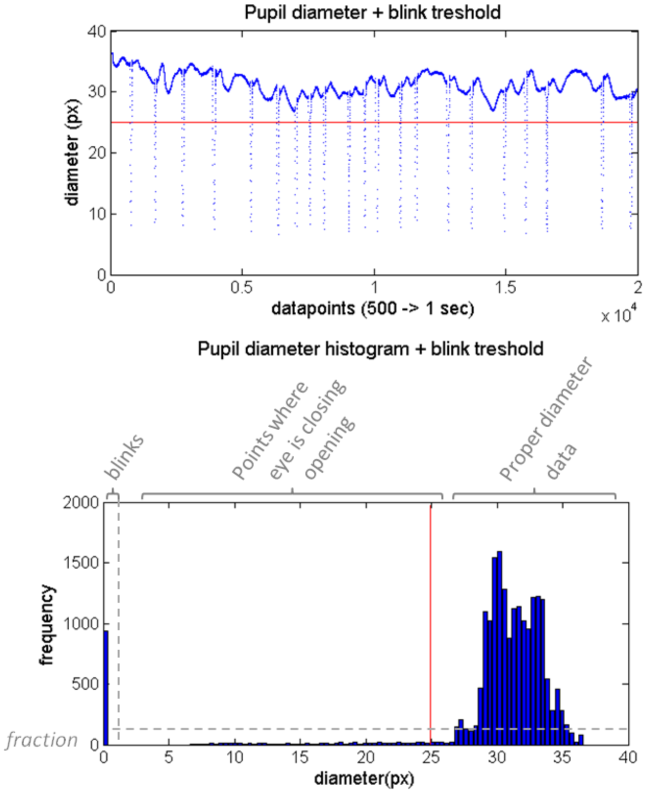


Fig. 6. The blink threshold in relation to the pupil diameter graph and histogram. The threshold is depicted as a red line and is set at 25 in this example. Note how none of the ‘proper’ data points are clipped.

It is assumed that for each blink the number of points belonging to the eye opening and closing is is relatively constant and that they are distributed evenly over the area between zero and the proper diameter data. For each trial, the number of data points where the diameter is zero is counted and this number is divided by a constant. At any point, in the area between zero and the proper diameter data, the frequency of data should be lower than this fraction. This constant is subject-dependent and derived by observation of the subject’s data.

When a bin in the histogram matches this frequency (moving right from zero), the value for this bin is taken as the threshold. This threshold is subsequently lowered by 2 pixels, to make sure not too many data are clipped.

Once the threshold is found, and the data that fall underneath are marked as a blink location, a margin is applied at the front and end of this location. This margin is required to catch any floating blinking points still left in the data, since the use of a threshold will not cut all desired points.

In the last step, the data belonging to the specific eye are masked using the blink location information that was previously obtained. The following pseudo-code illustrates in detail how this algorithm works:

```
$bin[0], frequency[0] =$ histogram$(pupilDiameter)$
$binNumber = 0$
```

```
while $frequency[binNumber] < frequency[0]/4$ do
    increment $binNumber$
end while

$blinkTreshold = bin[binNumber] - 2$

$blinkLocations = zeroes(length(pupilDiameter))$

for $i = 0$ to length($pupilDiameter$) do
    if $pupilDiameter_i < blinkTreshold$ then
        $blinkLocations_i = 1$
    end if
end for

apply_margin($blinkLocations, marginLeft, marginRight$)

for $i = 0$ to length($blinkLocations$) do
    if $blinkLocations_i == 0$ then
        $pupilDiameter_i = NaN$
        $eyeLocationX_i = NaN$
        $eyeLocationY_i = NaN$
    end if
end for
```

An example of the position of the threshold, resulting from the use of this algorithm, in relation to the pupil diameter can be found in Fig. 6.

B. Feature Extraction

Now that the blinks are filtered from the data, it can be segmented and from each of these segments a feature vector can be calculated. This process is illustrated in Fig. 7.

1) Data Segmentation: Because the features are defined over a period of one second, the segment size should also cover one second. In this case the data were recorded at 500Hz and therefore each segment consists of 500 data points.

To calculate the features, the data from each trial will be split in segments of 500 data points; the data points left over at the end are ignored. Technically these data points could be used to create an additional segment, using either a mirroring technique or by extrapolating from the available data points, but this will only result in 42 extra data points at best (6 conditions * 7 trials). Since this would result in a marginal performance increase at best and require relatively large amount effort to do so, I opted to ignore these data points instead.

2) The Features: From each segment, seven features are calculated. These features are mostly the same as the ones used by S. Marshall, with the exception of the Index of Cognitive Activity Index (ICA) feature(s) [1].

The ICA feature is an invention by S. Marshall and while I have used the technique as inspiration for one of my features, I did not manage to duplicate it entirely [10]. Therefore I will not refer to it as the ICA feature, but simply the Cognitive Activity feature. I will explain more about the differences later.

In Table VI you can see which filtered data vectors are required to calculate each of the features.

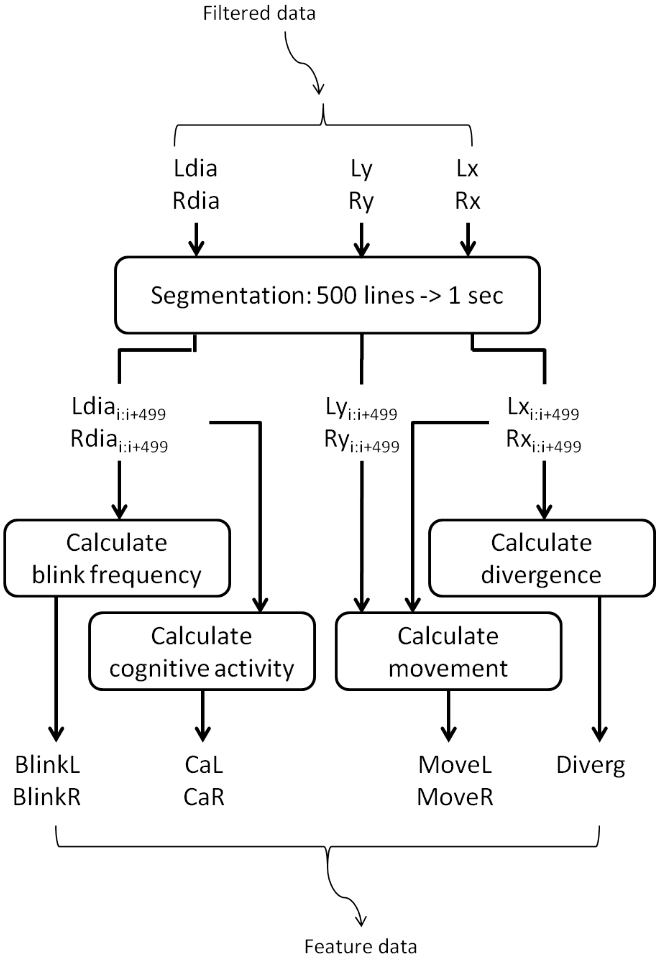


Fig. 7. Feature extraction schema. The filtered data vectors are segmented into chunks of 500 data points. These chunks are then used to calculate the different feature values: e.g., to calculate the movement feature value for the left eye, the x and y locations of the left eye are required. Each of these steps are explained in detail in Chapter IV-B.

TABLE VI
THE SEVEN FEATURES AND THE REQUIRED FILTERED DATA VECTORS REQUIRED TO CALCULATE THEM

Feature	Data required
Blinks, Left	Ldia
Blinks, Right	Rdia
Movement, Left	Lx, Ly
Movement, Right	Rx, Ry
Divergence	Lx, Rx
Cognitive Activity, Left	Ldia
Cognitive Activity, Right	Rdia

3) The Blink feature: The blink feature is defined as the proportion of a second in which the eye is blinking, i.e., the eyelid is moving across the eye [1].

This is precisely what has been removed from the data during blink removal, so in order to calculate this feature, the number of NaN's per segment of 500 data points are counted:

$$Blinks = \frac{\sum_{i=1}^{500} isNaN(pupilDiameter_i)}{500}$$

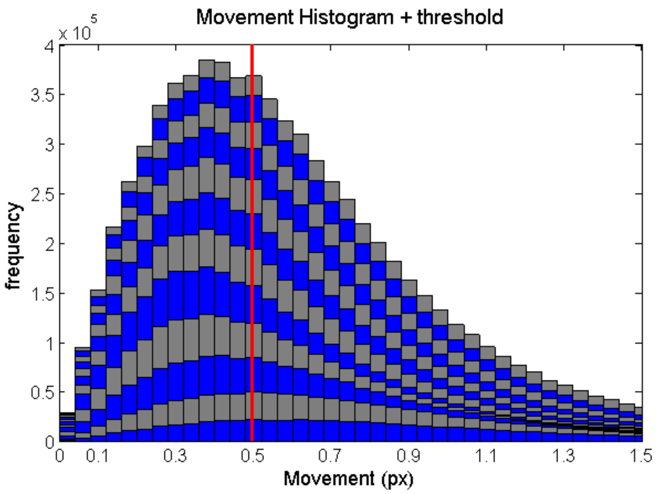


Fig. 8. Histogram of distance between consecutive eye gaze locations, for each eye, for all participants. The threshold used to differentiate between fast and slow eye movements divides the data roughly in two.

where

$$isNaN(x) = \begin{cases} 1 & \text{if } x = NaN \\ 0 & \text{otherwise} \end{cases}$$

The formula described is used twice to calculate a feature value for each eye.

4) The Movement Feature: The eye movement feature is defined as the proportion of a second in which the eye is moving (i.e. neither stationary nor blinking) [1].

To calculate this, two consecutive gaze locations for an eye are taken and the distance between them is calculated. If the distance is longer than a certain threshold, the eye is considered to be moving.

S. Marshall used a threshold of one pixel to divide the data into fixations or blinks, and eye movement. While the data used by Marshall had a frequency of 250Hz, the data here were recorded at a frequency of 500Hz. To account for this difference either half the data points would have to be ignored and instead only the distance between every other data point calculated, or the threshold should be made twice as strict.

It does seem counter-intuitive to take a threshold of half a pixel, but the data seem to support the idea. In Fig. 8 the histogram of the movement is plotted. As you can see the mean of each of those graphs seems to lie close to the 0.5 pixel mark.

This illustrates that the 0.5 mark is not completely arbitrary, but actually provides a good division of the data. It might imply that the feature I extracted is not necessarily the proportion of a second in which the gaze is moving, but rather the proportion of a second where the gaze is moving fast, with the inverse being the proportion of a second where the gaze is moving slowly or not at all.

The feature is calculated as follows:

$$Movement = \frac{\sum_{i=1}^{500} geq(\delta_i, \theta)}{500}$$

where

$$geq(x, y) = \begin{cases} 1 & \text{if } x \geq y \\ 0 & \text{otherwise} \end{cases}$$

and

$$\delta_i = \sqrt{(X_i - X_{i-1})^2 + (Y_i - Y_{i-1})^2}$$

and

$$\theta \text{ is } 0.5$$

and

X and Y are the x and y coordinates of the eye on the screen

The formula described is used twice to calculate a feature value for each eye.

5) The Horizontal Divergence Feature: The last feature calculated is the divergence between the eyes. It is defined as horizontal distance between the gaze points of each eye [1].

It is calculated by taking the horizontal locations of the gaze of each eye on the screen and calculating the distance between them. Data instances where the eye is blinking are excluded and have to be subtracted from the total segment length for the division.

$$Divergence = \frac{\sum_{i=1}^{500} isNaN(Lx_i, Rx_i)\sqrt{(Lx_i - Rx_i)^2}}{\sum_{i=1}^{500} \neg isNaN(Lx_i, Rx_i)}$$

where

$$isNaN(x, y) = \begin{cases} 1 & \text{if } x = NaN \vee y = NaN \\ 0 & \text{otherwise} \end{cases}$$

6) The Cognitive Activity Feature: The Index of Cognitive activity (ICA) feature is designed to differentiate between the light and the dilation reflex in pupil diameter data [10].

I diverged slightly from the method to calculate the ICA feature. To avoid any misunderstandings I have named this feature the Cognitive Activity feature. To explain where the method differs, it is necessary to first explain how the Index of Cognitive Activity is calculated.

The first step is to apply a wavelet transform to the pupil diameter data.

Wavelet analysis is especially suited to analyze non-stationary signals. A wavelet has a compact support, i.e., it has a finite start and finish. Furthermore, wavelets are designed to retain both time and frequency information, in contrast to, for example, conventional Fourier Analysis, which only retains frequency information.

A family of wavelets is derived from a mother wavelet ψ by dilation and translation:

$$\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$$

where

$$j \text{ is an index of dilation}$$

$$k \text{ is an index of translation}$$

This family of wavelets encodes x. In the case of the ICA, the pupil diameter data are substituted for x.

This encoding consists of a smoothed approximation of the signal, plus all the wavelet coefficients. Together they can be used to reverse the transformation fully.

Each wavelet coefficient c_{jk} expresses how well the wavelet ψ_{jk} fits to the signal and so the wavelet coefficients represent the signal locally. To obtain the ICA, the wavelet coefficients are used to find the changes in pupil diameter caused by cognitive processing.

The mother wavelet used in this case, is the Daubechies(32) [15]. The number 32 refers to the number of wavelet coefficients. This number was inferred from the directions given to obtain the ICA, by S. Marshall [10].

According to these directions, to be able to isolate the dilation reflex from the light reflex, the Daubechies(8) should be used for eye tracking data with a frequency of 60Hz and the Daubechies(16) for a frequency of 250Hz. Since the data obtained from this experiment have a frequency of 500Hz, scaling according to those suggestions results in the use of a Daubechies(32). Choosing this wavelet will ensure that per time unit I would obtain the same amount of wavelet coefficients as S. Marshall.

The signal is denoised by removing the parts of the signal with wavelet coefficients below or above certain thresholds using a minimax threshold estimation algorithm as given in [16].

The application of the wavelet and this denoising technique are combined together in a function in the Matlab Wavelet Toolbox: $wden$. This function was used to perform the wavelet decomposition.

In order for this function to be applied to the data the blinks need to be interpolated. An interpolation function already implemented in Matlab was used for this (*interp1*).

After both these functions have been applied, the next step is to find the places in the pupil diameter signal where there was a specific increase, and check whether the corresponding wavelet detail coefficient was high. If this is the case this position is marked, otherwise it is ignored. The ICA feature is then the number of marked places per second. As a result of this last step noise is decreased.

This last step was not described in sufficient detail for me to replicate it adequately. Instead I used a threshold to filter the wavelet coefficient vector and counted the resulting number of data points above the threshold:

$$CognitiveActivity = \sum_{i=1}^{500} geq(detailCoef_i, \theta)$$

where

$$\theta = 0.1$$

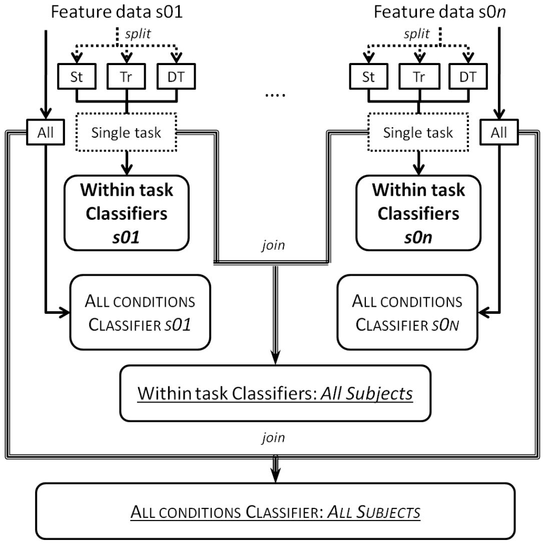

Feature data s01
split
St
Tr
DT
....
Feature data s0*n*
split
St
Tr
DT
All
Single task
Single task
All
Within task Classifiers *s01*
Within task Classifiers *s0n*
join
All conditions Classifier *s01*
All conditions Classifier *s0n*
Within task Classifiers: *All Subjects*
join
All conditions Classifier: *All Subjects*

Fig. 9. Classification schema. The feature files for each participant are split based on the task, resulting in three files: St, Tr and DT, with the data from the Sternberg task, the Tracking task and the Dual task, respectively. Furthermore data files are joined to create new data files that contain the data from all participants combined. These files are then fed to the relevant type of classifier. A distinction is made between different types of classifiers, because the results of each type aim to answer a specific research question.

and

$$detailCoef = \text{ the wavelet detail coefficient}$$

The method described above is used twice to calculate a feature value for each eye.

A note on the difference between the ICA feature and the Cognitive Activity feature: S. Marshall specifies that the ICA feature should typically fall between 0-20 counts per second, where 0 counts indicate conditions with low cognitive workload and 20 counts indicate high cognitive workload [1]. With this threshold function applied to the wavelet coefficients the data from this experiment range also generally between these values. I therefore assume that this CA feature is at least comparable to that from S. Marshall.

C. Classification

When the features are extracted, the data can be classified. The goal of the classification is to answer the research questions posed in the introduction. To answer these questions the data need to be classified in different ways. The different type of classifiers that are used can be seen in Fig. 9. Prior to classification the feature files are split and/or joined to create the required input.

Typesetting has been used to clarify which classifiers are designed to answer which specific question:

- Can the same features as used by S. Marshall, extracted from eye data, be used to distinguish between different levels of Cognitive Workload[1]?

  The classifiers designed to answer this question are **bolded** in Fig. 9
- Is this method task dependent?

The classifiers designed to answer this question are in SMALLCAPS in Fig. 9

- Is this method subject dependent?
  The classifiers designed to answer this question are underlined in Fig. 9

In Fig. 9 can also be seen how the files are split and joined. The split files contain data from each of the specific tasks, that each consist of data from two difficulty levels. Therefore the classifiers that they serve as input for, are called within-task classifiers. These classifiers each solve two-class problems.

The all-conditions datasets consist of all the feature data of a participant. These serve as input for classifiers that solve a six-class problem, since they are designed to distinguish between all six different conditions.

Lastly, the data for all participants are joined and serve as input for the classifiers that will solve either the within task problems or the all-condition problem, but this time for all participants at the same time.

1) Classifiers: WEKA, software that holds a collection of machine learning algorithms, was used to classify the data [17]. Each dataset is classified twice, once using a Multilayer Perceptron (MLP) and once using a Support Vector Machine (SVM). For both of these an implementation exists in WEKA.

The MLP was chosen on the basis that it was also used by S. Marshall [1]. She used a MLP implementation in Matlab, with a hidden layer of 5 nodes that employed *early stopping* using a validation set of 25%. In *early stopping*, the validation set is used to check between each iteration through the training set, whether or not the performance is still increasing. If not, training is stopped.

Furthermore, her results were obtained through a hundred runs per dataset. A specific test set was separated prior to classification to test the generalization of the networks afterwards. The main reported results were the classification rates of the network on all data.

As you can see in Table VII MLP's with a hidden layer of 5 nodes are used together with *early stopping*, also with a validation set of 25%. Instead of running each dataset a hundred times, I have chosen to use k-fold cross validation of 10 for each classifier, which will be repeated 10 times. This also results in a hundred classification rates per dataset, but at the same time makes sure that, for each dataset, each instance has been used for training once per run out of the 10. Furthermore I will be reporting the results on the test set, not performance on the entire dataset.

The other method of classification used by S. Marshall was linear discriminant function analysis. Since WEKA does not offer this, an SVM algorithm has been used instead. The implementation of the SVM in WEKA uses the Sequential Minimal Optimization algorithm by J. Platt [18]. The parameters used for this algorithm are shown in Table VIII.

The kernel used for classification is the Pearson IIV function-based universal kernel [19]. This kernel has been designed to be more universally applicable than, for example, the Polynomial or Radial Basis Function kernels, to circumvent the problem of being forced to choose the kernel best suited for the data through extensive testing.

TABLE VII
PARAMETERS SET FOR THE MULTILAYER PERCEPTRON CLASSIFIER IN WEKA

Multilayer Perceptron	
Back propagation	
Parameter	Value
decay	false
hiddenNodes	5 (one layer)
learningRate	0.3
normalizeAttributes	true
momentum	0.2
validationSetSize	25%
validationThreshold	20

TABLE VIII
PARAMETERS SET FOR THE SUPPORT VECTOR MACHINE CLASSIFIER IN WEKA

Support vector machine	
Sequential Minimal Optimization (SMO)	
Parameter	Value
buildLogisticModels	false
filterType	normalize training data
kernel	Puk (Pearson IIV function-based universal kernel)

For this classifier, 10-fold cross validation was also applied and repeated 10 times.

V. RESULTS

A. Workload Profile

During classification of the data, the assumption is that there is an actual difference in cognitive workload that the subjects are under in different conditions.

In order to check this assumption, the Workload Profile ratings for all subjects need to be analyzed.

In Table X, the results of each participant can be found.

There are differences noticeable between subjects, both in the scaling of the ratings and in the relative difference between conditions. On the other hand, some trends in relative difference between conditions are visible. Statistical analysis can show whether these trends are also significant.

For this analysis the first null hypothesis is:

$$H_0 : \text{ No conditions differ in workload score}$$

With the alternative hypothesis:

$$H_a : \text{ At least one condition differs in workload score}$$

The hypothesis is tested by performing a GLM Repeated Measures with the within-subject factor: task difficulty(1-6), and the dependent variable: the Workload Profile rating. The results were significant(Huyn-Feldt:F = 8.582, p = 0.04). Therefore the alternative hypothesis is true: at least one condition differs in the workload score from the others.

The follow-up question is if all pair wise comparisons are also significant. The results can be seen in Table XI.

It shows that two of the pair wise comparisons within-task are significant (3vs4, p = 0.01; 5vs6, p =0.005). The pair wise comparison between 1vs2 is not significant, and neither are most of the other pairs.

TABLE IX
MEANS AND STANDARD DEVIATIONS ACROSS DATA FROM ALL PARTICIPANTS OF THE FEATURES REPORTED PER CONDITION

Means + (standard deviations)							
		Sternberg		Tracking		Dual Task	
		Easy	Hard	Easy	Hard	Easy	Hard
Blinks	Left	0.091(0.143)	0.118(0.170)	0.042(0.113)	0.028(0.090)	0.034(0.097)	0.079(0.171)
	Right	0.098(0.155)	0.119(0.169)	0.044(0.114)	0.033(0.100)	0.035(0.096)	0.082(0.173)
Movement	Left	0.498(0.148)	0.504(0.155)	0.541(0.113)	0.534(0.083)	0.527(0.154)	0.503(0.155)
	Right	0.520(0.175)	0.527(0.182)	0.569(0.158)	0.574(0.130)	0.558(0.184)	0.542(0.178)
Divergence		10.780(7.433)	17.272(11.533)	13.090(10.248)	10.425(7.778)	16.862(11.644)	17.558(13.721)
Cognitive Activity	Left	3.478(3.073)	4.517(4.770)	3.390(3.981)	2.673(2.968)	3.229(3.979)	4.834(6.092)
	Right	4.255(4.198)	4.950(5.274)	3.373(4.350)	2.848(3.753)	2.815(3.315)	4.562(5.883)

TABLE X
WORKLOAD PROFILE RESULTS. THE RESULTS ARE OBTAINED BY SUMMING THE SCORES ON THE DIFFERENT SUB DIMENSIONS OF THE WORKLOAD PROFILE. BOTH THE INDIVIDUAL SUBJECT SCORES AND THE AVERAGE AND STANDARD DEVIATION ARE DISPLAYED FOR EACH CONDITION. 'E' AND 'H' REFER TO THE DIFFICULTY OF THE TASK, EASY AND HARD RESPECTIVELY.

Task	s01	s02	s03	s04	s05	s06	μ	σ
Sternberg(e)	3.1	0.75	0.75	1.7	0.9	1.7	1.48	0.91
Sternberg(h)	3.3	0.85	1.25	2.3	1.5	1.9	1.85	0.87
Tracking(e)	2.9	0.25	0.65	0.6	2.9	0.6	1.32	1.23
Tracking(h)	3.3	0.55	0.9	1.15	3.2	0.9	1.67	1.24
Sternberg(e) Tracking(e)	4.7	1.05	1.25	3.15	4.2	1.35	2.62	1.62
Sternberg(h) Tracking(h)	4.9	1.5	2.05	3.9	4.85	2.6	3.3	1.46

TABLE XI
SIGNIFICANCE LEVELS OF PAIR WISE COMPARISONS OF WORKLOAD PROFILE RESULTS. LEVELS SIGNIFICANT AT P $<$ 0.05 ARE STARRED: '*'.

Comparison		Sig. (p-values)
Condition 1	vs. 2	0.10
	vs. 3	0.736
	vs. 4	0.702
	vs. 5	0.084
	vs. 6	0.013*
Condition 2	vs. 3	0.276
	vs. 4	0.679
	vs. 5	0.168
	vs. 6	0.018*
Condition 3	vs. 4	0.01*
	vs. 5	0.08
	vs. 6	0.01*
Condition 4	vs. 5	0.016*
	vs. 6	0.01*
Condition 5	vs. 6	0.005*

These results could be explained by the low number of subjects used for this research. Still this will have some implications for any results I might find, these will be discussed in the Discussion section.

B. *Feature distribution*

To find out what relations exist between the features and cognitive workload, the means and standard deviations of each of the features are found in Table IX.

When looking at the blink feature for the Sternberg task, the mean is lower in the easy condition than it is in the hard condition (left:0.090 vs. 0.118 respectively; right: 0.098 vs. 0.119 respectively). For the Tracking task this relationship is inversed (left: 0.042 vs. 0.028, easy vs. hard; right: 0.044 vs. 0.033, easy vs. hard). Finally, for the Dual Task, the mean for the easy condition compared to the hard condition is once again lower (left: 0.034 vs. 0.079 respectively; right: 0.035 vs. 0.082). This indicates that this feature is likely task dependent, and that a more general, task independent, relationship between cognitive workload and the proportion of blinks per second cannot be specified.

Looking at the movement feature in the same vain: For the Sternberg task, the means increase with cognitive workload; For the Tracking task the differences between the means are not significant for both eyes and for the Dual Task the means decrease with cognitive workload.

Next, looking at the divergence feature: For the Sternberg and Dual Task task, the means increase with cognitive workload; For the Tracking task the means decrease with cognitive workload.

Lastly, looking at the Cognitive Activity feature: For the Sternberg and Dual Task task, the means increase with cognitive workload; For the Tracking task the means decrease with cognitive workload.

Looking at the standard deviations for each of the features, they appear to be rather large in comparison to the mean of the respective features, especially since all features range from zero up. This can be explained by looking at the distributions of the features. Not all of them follow the pattern of a normal distribution, which becomes obvious when looking at the box plots of the features, for each of the conditions. This plot can be found in Appendix A.

C. *Classification Results*

In SectionIV-C was explained how the different types of classifiers were set up in order to allow the results obtained from these classifiers to answer the research questions posed in the introduction.

The results from these different types will also be reported separately, to keep this distinction clear.

1) Within-task classifiers using individual models: The within-task classifiers are those that classify between the two difficulty levels of each task (i.e., Sternberg(easy) vs.

TABLE XII
RESULTS FROM THE EXPERIMENTS DISTINGUISHING BETWEEN COGNITIVE STATES BY S. MARSHALL‘[1]. A DISTINCTION IS MADE BETWEEN RESULTS OF SINGLE MODEL CLASSIFIERS AND INDIVIDUAL MODEL CLASSIFIERS. THE RESULTS FROM LINEAR DISCRIMINANT FUNCTION (LDF) AND MULTILAYER PERCEPTRON (MLP)ARE SEPARATED.[1]

Individual Model		
		μ
LDF	Relaxed vs. Engaged	73%
	Focused vs. Distracted	70%
	Alert vs. Fatigued	73%
MLP	Relaxed vs. Engaged	79%
	Focused vs. Distracted	69%
	Alert vs. Fatigued	79%

Single Model		
		All subjects
LDF	Relaxed vs. Engaged	62%
	Focused vs. Distracted	59%
	Alert vs. Fatigued	n/a
MLP	Relaxed vs. Engaged	65%
	Focused vs. Distracted	61%
	Alert vs. Fatigued	n/a

TABLE XIII
CLASSIFICATION RESULTS FROM THE WITHIN-TASK CLASSIFIERS FOR THE INDIVIDUAL MODELS.

Individual model			
		μ	σ
MLP	Dual Task	82.68%	12.79%
	Sternberg	79.40%	13.68%
	Tracking	75.60%	13.61%
SVM	Dual Task	83.00%	12.40%
	Sternberg	80.44%	12.47%
	Tracking	76.66%	12.53%

TABLE XIV
CLASSIFICATION RESULTS FOR THE ALL-TASK CLASSIFIERS FOR THE INDIVIDUAL MODELS.

Individual model		
	μ	σ
MLP - All Data	55.16%	9.97%
SVM - All Data	56.52%	8.50%

Sternberg(hard), Tracking(easy) vs. Tracking(hard) and Dual Task(easy) vs. Dual Task(hard)). The classifiers were trained and tested on the individual models which means that for each participant a new model was trained.

The results of this classification can be found in Table XIII.

The classification rates shown are generally well above the chance level for a two class problem, which is 50%.

For comparison, the results from S. Marshall can be found in Table XII. Since they are being compared to the results from the individual models from the experiment, the results in the individual model table should be consulted.

2) All-task classifiers using individual models: The all-task classifiers are those that classify all six conditions. Once again, for each of the participants a new model was trained.

The results obtained can be found in Table XIV

These classification rates are noticeably lower than those from the within task classifiers, but are still well above chance level: the chance level for a six class problem would be 16.7%.

3) Within-task + all-task classifiers using a single model: The within-task and all-task datasets were also combined for all participants and used to create single model classifiers.

These results can be found in Table XV.

Compared to the performance of the individual model classifiers these classification rates are distinctly lower.

The within-task results on the single model can be compared to the single model results by S. Marshall. For these results see Table XII

TABLE XV
RESULTS FROM SINGLE MODEL CLASSIFICATION FROM BOTH WITHIN-TASK AND ALL-TASK CLASSIFIERS.

Single Model			
			All Subjects
MLP	Task data	Dual Task	70.24%
		Sternberg	70.75%
		Tracking	70.79%
	All data		35.16%
SVM	Task data	Dual Task	75.85%
		Sternberg	73.72%
		Tracking	73.88%
	All data		44.58%

VI. DISCUSSION

Based on the results, conclusions can be drawn about the research questions posed in the introduction.

I will reiterate these questions for clarity:

- Can a method be developed that employs the same features (i.e. those used by S. Marshall) extracted from eye behavior data, to distinguish between different levels of cognitive workload[1]?
- Is this method task dependent?
- Is this method subject dependent?

A. Conclusions

1) Main research question: Can a method be developed that employs the same features (i.e. those used by S. Marshall) extracted from eye behavior data, to distinguish between different levels of cognitive workload [1]?

The results from the within-task classifiers using the individual models, averaged over all participants, range from 76% to 83% for the MLP, and from 77% to 83% using the SVM. These results are at a comparable level and even somewhat higher than the results found by S. Marshall [1], which range from 69% to 79% for the individual models. From this can be concluded that this method definitely can distinguish between different levels of cognitive workload.

It should be noted that the performance on the Sternberg task is 75% while there was no significant effect found between the respective Workload Profile ratings. There are multiple explanations possible for this: the lack of a significant effect is caused by the low number of subjects (n=6), or classifiers are not actually distinguishing between different levels of cognitive workload but between something else.

The individual workload profile ratings of each of the participants (see Table X) indicate that the former explanation is quite likely, since each of the participants has rated the harder Sternberg condition as imposing more cognitive workload than the easy Sternberg condition.

2) Is this method task dependent?: Different tasks can elicit different eye behavior in a subject. The question is whether this method can still distinguish between different levels of cognitive workload if that cognitive workload is imposed by different types of tasks.

The classifiers on all-tasks using individual models performed at 55% for the MLP and at 57% for the SVM. These classification rates are in the low range, but definitely above chance level, which is 16.7% in this case.

On the other hand, the results from the means and standard deviations of each of the features seemed to indicate that most of the features on their own are very task dependent. With such differences in the direction of the relationship with cognitive workload for each of the features, classification rates ranging from 55% to 57% could indicate that this specific combination of features removes some of this task dependence.

Furthermore, the lack of significant effects found during statistical analysis, could explain a proportion of the low classification rates as well (only a few were found: for reference see XI).

Without clearer evidence it is impossible to draw a definite conclusion, but the results do indicate that this method is task-dependent, if at least to a degree.

3) Is this method subject dependent?: To answer this question, the data from all participants were combined and a single model was trained for all the within-task datasets.

The results range from 70% to 71% for the MLP and from 74% to 76% for the SVM. Comparing this to the results of the individual models these results are only about 5% to 10% lower. It is to be expected that a single model performs somewhat worse than models that are trained for a specific individual, since variance in eye behavior can be expected between different subjects. Such variance already shows in the standard deviation of the classification rates for the different tasks (around 12% - 13%).

Furthermore there is also variance between the individual workload profile ratings for each of the conditions.

All in all, this method definitely seems to generalize well over different subjects. The conclusion is that this method is not subject dependent.

B. Future research

These findings indicate that a system to distinguish between different levels sufficiently distinct of cognitive workload in real time could be developed. The system would have to be restricted to a single task that could be composed of smaller subtasks that are executed simultaneously (as in the Dual Task condition). Furthermore the system would require training on the expected levels of cognitive workload before use. This limits real world applications for the moment, but with continued research perhaps advancements could be made.

Whether this method is task dependent or not requires more compelling evidence. To solve the problem of a lack of significant effects between Workload Profile ratings, the difference in cognitive workload imposed by the different tasks should be increased, or a larger amount of subjects should participate.

Currently seven features were used to classify the data; in future research more features could be investigated. Established research shows that the addition of features that describe the mean or standard deviation of an existing feature can increase performance [20]. While this was specifically applied to multi stroke gesture recognition, it is conceivable that this could work for these eye-behavior features as well.

For example the divergence feature currently describes the average of the divergence between the eyes over a segment. Since the calibration can affect the baseline divergence for a subject, extra information could be gained from the standard deviation of the divergence over a given segment.

Furthermore, the classification performed in this research was offline. A new line of research could investigate whether or not an online implementation could be achieved. While the classification rates from the within-task classifiers were perhaps too low for success with an online system, they could be increased by increasing the interval size of the features. Currently the data were classified with instances covering one second of eye tracker data. This could be increased to 5 or even 10 seconds per interval.

Moving toward such an online system is the next step in achieving the goal of a true Augmented Cognition: To create a real-time system that detects the user's cognitive state and adapts accordingly.

REFERENCES

[1] S. Marshall, "Identifying cognitive state from eye metrics," *Aviation, Space, and Environmental Medicine*, vol. 78, no. 5, Section II, 2007.

[2] A. Kruse and D. Schmorrow, "Session overview: Foundations of augmented cognition," in *Foundations of Augmented Cognition*, D. Schmorrow, Ed. Lawrence ErlBaum Associates, 2005.

[3] A. Gevins, M. Smith, and H. Leong, "Monitoring working memory load during computer-based tasks with eeg pattern recognition methods," *Human Factors*, vol. 40, pp. 79–91, 1998.

[4] A. Girouard, E. Solovey, L. Hirshfield, K. Chauncey, A. Sassaroli, S. Fantini, and R. Jacob, "Distinguishing difficulty levels with non-invasive brain activity measurements," in *Proceedings of the 12th IFIP TC 13 International Conference on Human-Computer Interaction: Part I*, ser. INTERACT '09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 440–452.

[5] L. Hirshfield, E. Solovey, A. Girouard, J. Kebinger, A. Sassaroli, Y. Tong, S. Fantini, and R. Jacob, "Brain measurement for usability testing and adaptive interfaces: An example of uncovering syntactic workload with functional near infrared spectroscopy." in *Human Factors in Computing Systems*, 2009.

[6] C. Ikehara, M. Crosby, and D. Chin, "A suite of physiological sensors for assessing cognitive states," in *Foundations of Augmented Cognition*, D. Schmorrow, Ed. Lawrence ErlBaum Associates, 2005, pp. 273–282.

[7] J. Stern, L. Walrath, and R. Goldstein, "The endogenous eyeblink," *Psychophysiology*, vol. 21, pp. 22–33, 1984.

[8] A. Carpenter, "The rate of blinking during prolonged visual search," *Journal of Experimental Psychology*, vol. 38, pp. 587–591, 1948.

[9] I. Lowenfeld, *The pupil: Anatomy, physiology, and clinical applications.* Iowa State University Press (Ames and Detroit), 1993.

[10] S. Marshall, "Method and apparatus for eye tracking and monitoring pupil dilation to evaluate cognitive activity," Patent, 07 2000, uS 6090051.

[11] S. Rubio, E. Diaz, J. Martin, and J. Puente, "Evaluation of subjective mental workload: A comparison of swat, nasa-tlx, and workload profile methods," *Applied Psychology: an International Review*, vol. 53, no. 1, pp. 61–86, 2004.

[12] S. Hart and L. Staveland, “Development of nasa-tlx (task load index): Results of empirical and theoretical research.” *Human mental workload*, pp. 139–183, 1988.

[13] G. Reid and T. Nygren, “The subjective workload assessment technique: A scaling procedure for measuring mental workload,” *Human mental workload*, pp. 185–218, 1988.

[14] P. Tsang and V. Velazquez, “Diagnosticity and multidimensional subjective workload ratings,” *Ergonomics*, vol. 39, no. 3, pp. 358–381, 1996.

[15] I. Daubechies, “Ten lectures on wavelets,” *SIAM review*, vol. 35, no. 4, pp. 666–669, 1988.

[16] D. Donoho and J. Johnstone, “Ideal spatial adaptation by wavelet shrinkage,” *Biometrika*, vol. 81, pp. 425–455, 1994.

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reuteman, and I. Witten, “The weka data mining software: An update,” *SIGKDD Explorations*, vol. 11, 2009.

[18] J. Platt, “Fast training of support vector machines using sequential minimal optimization,” in *Advances in Kernel Methods - Support Vector Learning*, B. Schoelkopf, C. Burges, and A. Smola, Eds. MIT Press, 1999.

[19] B. stn, W. Melssen, and L. Buyden, “Facilitating the application of support vector regression by using a universal pearson vii function based kernel,” *Chemometrics and Intelligent Laboratory Systems*, vol. 81, pp. 29–40, 2006.

[20] D. Willems, R. Niels, M. van Gerven, and L. Vuurpijl, “Iconic and multi-stroke gesture recognition,” *Pattern Recognition*, vol. 42, pp. 3303–3312, 2009.

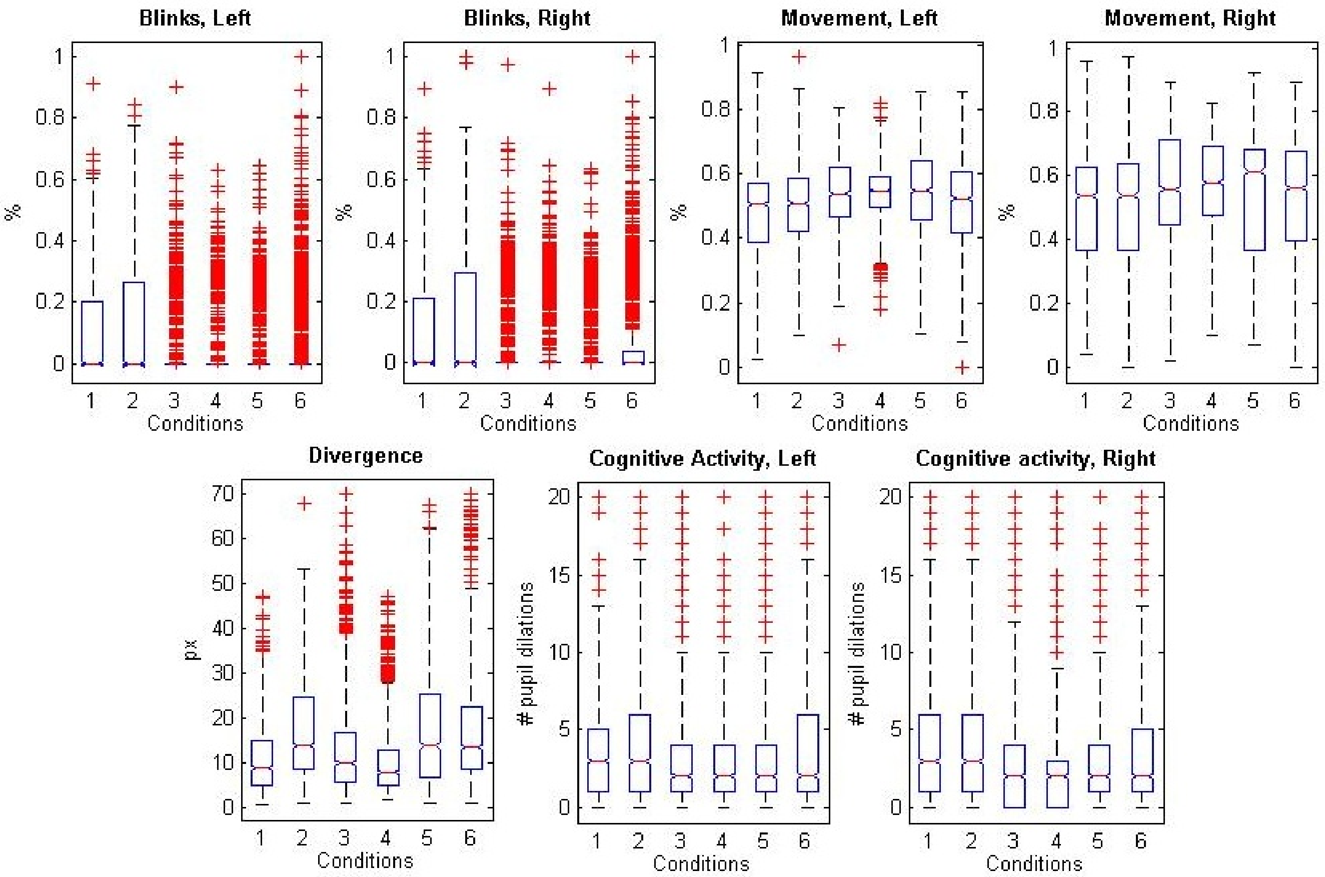
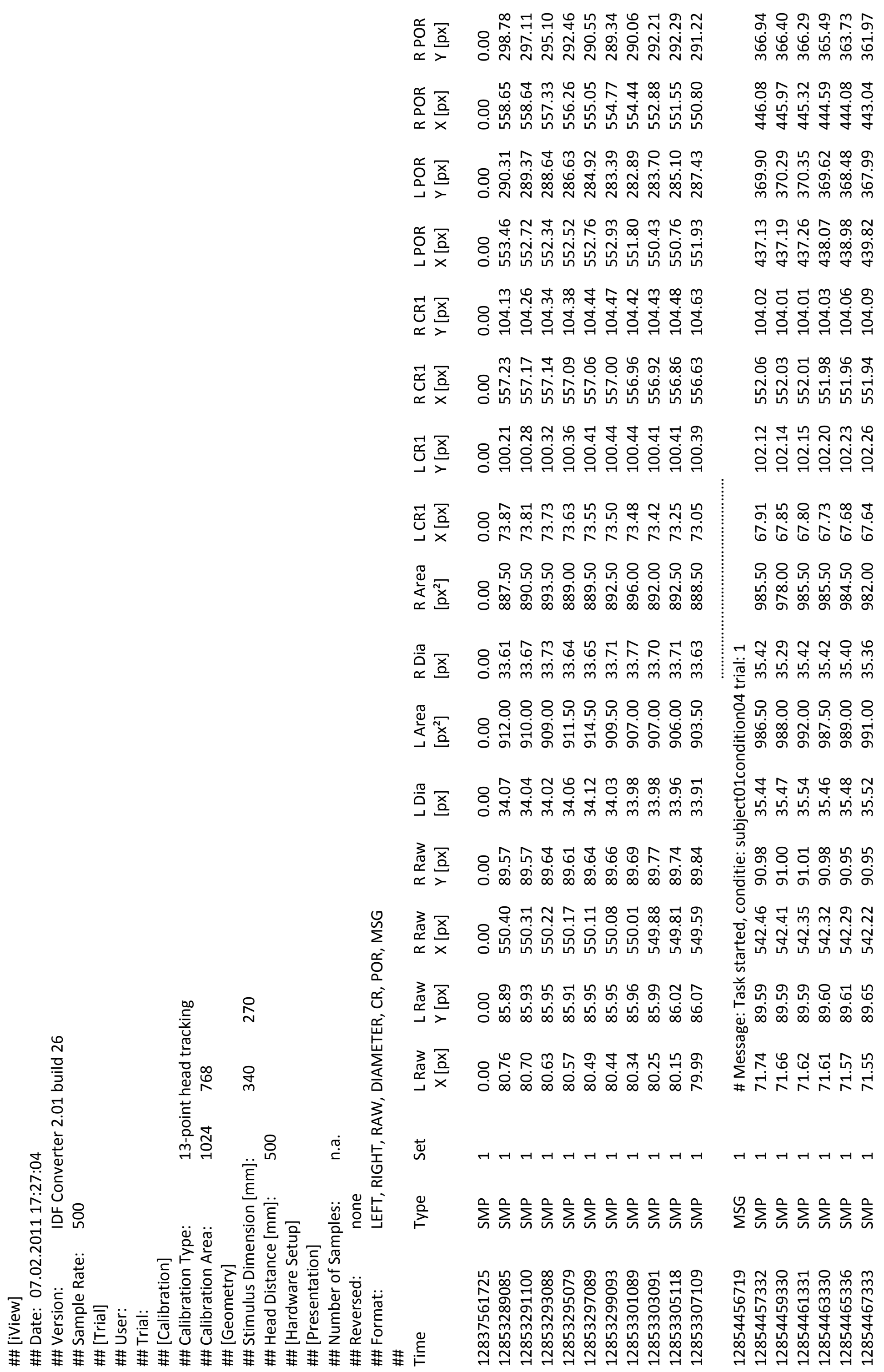


Fig. 10. Box plots for each of the features used in classification of data from all the participants. Each plot represents one feature, in which each of the box plots represents one condition. Condition numbers mean the following: 1 = Sternberg, easy; 2 = Sternberg, hard; 3 = Tracking, easy; 4 = Tracking, hard; 5 = Dual Task easy, 6 = Dual Task, hard. For each of the box plots the top and bottom represent the 75th and the 25th percentile, respectively. The line near the middle represents the 50th percentile which is also the median. The ‘whiskers‘ of the box plot can extend to twice the length of the box plot in both upward and downward direction, and extend up to the last data point that can be reached in that distance. Data points beyond these whiskers are marked with a red + and are normally considered outliers. (They are only necessarily outliers if the data plotted generally follow a normal distribution).

APPENDIX A
FEATURE DISTRIBUTIONS



```
## [iView]
## Date:  07.02.2011 17:27:04
## Version:        IDF Converter 2.01 build 26
## Sample Rate:    500
## [Trial]
## User:
## Trial:
## [Calibration]
## Calibration Type:         13-point head tracking
## Calibration Area:         1024     768
## [Geometry]
## Stimulus Dimension [mm]:           340      270
## Head Distance [mm]:       500
## [Hardware Setup]
## [Presentation]
## Number of Samples:        n.a.
## Reversed:       none
## Format:         LEFT, RIGHT, RAW, DIAMETER, CR, POR, MSG
##
Time          Type   Set   L Raw   L Raw   R Raw   R Raw   L Dia   L Area   R Dia   R Area   L CR1   L CR1    R CR1   R CR1    L POR   L POR   R POR   R POR
                           X [px]  Y [px]  X [px]  Y [px]  [px]    [px²]    [px]    [px²]    X [px]  Y [px]   X [px]  Y [px]   X [px]  Y [px]  X [px]  Y [px]

12837561725   SMP    1     0.00    0.00    0.00    0.00    0.00    0.00     0.00    0.00     0.00    0.00     0.00    0.00     0.00    0.00    0.00    0.00
12853289085   SMP    1     80.76   85.89   550.40  89.57   34.07   912.00   33.61   887.50   73.87   100.21   557.23  104.13   553.46  290.31  558.65  298.78
12853291100   SMP    1     80.70   85.93   550.31  89.57   34.04   910.00   33.67   890.50   73.81   100.28   557.17  104.26   552.72  289.37  558.64  297.11
12853293088   SMP    1     80.63   85.95   550.22  89.64   34.02   909.00   33.73   893.50   73.73   100.32   557.14  104.34   552.34  288.64  557.33  295.10
12853295079   SMP    1     80.57   85.91   550.17  89.61   34.06   911.50   33.64   889.00   73.63   100.36   557.09  104.38   552.52  286.63  556.26  292.46
12853297089   SMP    1     80.49   85.95   550.11  89.64   34.12   914.50   33.65   889.50   73.55   100.41   557.06  104.44   552.76  284.92  555.05  290.55
12853299093   SMP    1     80.44   85.95   550.08  89.66   34.03   909.50   33.71   892.50   73.50   100.44   557.00  104.47   552.93  283.39  554.77  289.34
12853301089   SMP    1     80.34   85.96   550.01  89.69   33.98   907.00   33.77   896.00   73.48   100.44   556.96  104.42   551.80  282.89  554.44  290.06
12853303091   SMP    1     80.25   85.99   549.88  89.77   33.98   907.00   33.70   892.00   73.42   100.41   556.92  104.43   550.43  283.70  552.88  292.21
12853305118   SMP    1     80.15   86.02   549.81  89.74   33.96   906.00   33.71   892.50   73.25   100.41   556.86  104.48   550.76  285.10  551.55  292.29
12853307109   SMP    1     79.99   86.07   549.59  89.84   33.91   903.50   33.63   888.50   73.05   100.39   556.63  104.63   551.93  287.43  550.80  291.22
                                                                                .............................................
12854456719   MSG    1     # Message: Task started, conditie: subject01condition04 trial: 1
12854457332   SMP    1     71.74   89.59   542.46  90.98   35.44   986.50   35.42   985.50   67.91   102.12   552.06  104.02   437.13  369.90  446.08  366.94
12854459330   SMP    1     71.66   89.59   542.41  91.00   35.47   988.00   35.29   978.00   67.85   102.14   552.03  104.01   437.19  370.29  445.97  366.40
12854461331   SMP    1     71.62   89.59   542.35  91.01   35.54   992.00   35.42   985.50   67.80   102.15   552.01  104.01   437.26  370.35  445.32  366.29
12854463330   SMP    1     71.61   89.60   542.32  90.98   35.46   987.50   35.42   985.50   67.73   102.20   551.98  104.03   438.07  369.62  444.59  365.49
12854465336   SMP    1     71.57   89.61   542.29  90.95   35.48   989.00   35.40   984.50   67.68   102.23   551.96  104.06   438.98  368.48  444.08  363.73
12854467333   SMP    1     71.55   89.65   542.22  90.95   35.52   991.00   35.36   982.00   67.64   102.26   551.94  104.09   439.82  367.99  443.04  361.97
```

Fig. 11. An example of output generated by the eye tracking system. —- indicate a jump in data, to show how messages show up in data.

APPENDIX B
OUTPUT FROM THE EYE TRACKER