Don’t Die Horribly!
A pupil size based workload management system

Marco Wirthlin (S2876507)
Dept. Artificial Intelligence
m.wirthlin@student.rug.nl

ABSTRACT

The aim of this paper is to explore the viability of interactive programs which dynamically adapt to its users’ workload, measured on-line and non-invasively via pupil dilation. To this end, we constructed a simple videogame where users had to avoid collision with falling blocks as long as possible and tested its potential experimentally. The games’ difficulty (speed of falling blocks) reactively adjusted depending on the users’ workload. For the program to correctly interpret user pupil dilation, a short baselining phase calibrates the optimal dilaton threshold. Our game would increase in difficulty unless this threshold gets crossed. Furthrmore, the more threshold violations occur, the lower gets the difficulty set.
Results show that the system successfully adjusts its difficulty to all users that took part of our study. 5 of 7 users perform the task better (survive longer) when difficulty adjustment is enabled. This opens up exciting possibilities as everydays integration of eyetrackers will soon be a reality.

Author Keywords

pupil dilation; eye tracking; user models; human factors; serious games; workload; adaptive interfaces;

INTRODUCTION

Wouldn’t it be wonderful if your computer *knew* when you are stressed or tired, and unclutter your screen, so you can just focus on what you were working on without all those other open windows or toolbars? A car that succsessfully *recognizes* that its driver is falling asleep, and thus takes control and slows down, could save many lives. Flight systems which *notice* that the pilot is unable to react or unable to perceive false bearings have already prooven successfull [6].

All this may become possible if “the machine” was designed to have a model of its user. In other words, for artificial systems to be able to adapt to the humans’ current state, the system needs to *understand* “how the human is doing” (the *operative performance*). This requires the systems’ designers to implement a model which can “characterize or predict” human performance [2]. Those models, like the one described in this paper, adress very domain specific human factor issues in particular tasks and may have very different goals. For instance, a model that predicts a pilots’ maximal response time for collision avoidance will differ from a model that estimates fatigue not only in its implementation, but also in the psychological processes it relies on. A reaction time estimation model most likely will rely on theories about attention [7] and the pilots’ reaction time measurments, while fatigue predictor models will rely on EEG or EOG [1].

For the present study, our goal was to build a simple proof-of-concept video game that would adapt its difficulty to the user. To this end, we implemented a user model based on pupil dilation measurements, inferring the users’ workload during the game. In other words, we build a workload management system based on pupillary dilation measuremtens. We hypothesized that the higher the games’ difficulty (game speed being measured in pixels per second), the higher the subjects’ workload (inferred with pupil dilation, pupil diameter in pixels) and thus, the lower task performance (lower in-game survival rate) and vice-versa. We tested this hypothesis experimentally with a within-subject design, where we compared the subjects’ task performance and relationship between difficulty levels and pupil dilation among two conditions: The experimental condition with reactive difficulty adjustment enabled and the control condition, where random difficulty levels were imposed. But, why is workload interesting in this context and how does it relate to pupillary dilation?

“Mental Workload” has been defined as “the poriton of an individual’s limited mental capacity that is actually required by task demands” [1]. It conceptualized as the result of the interaction between the task properties, the context in which the task is being performed and “the skills, behaviours and perceptions” of a user [3]. This construct integrates and characterizes the interplay of the different aspects of human machine interactions well. In our study, we controlled task context (keeping it constant across participants) and task demands with game difficulty. While this is a very useful construct, its effective assesment has been difficult as subjects usually have to disengage from the task and fill out self reports. This makes non-invasive, on-line indicators of workload very attractive. The connection between workload and pupil dilation has been discussed in the literature already for decades [5] and has been firmly established in a wide range of tasks [4] [8]. Despite of these firm claims, pupillometry never has played a mayor role in applied research because of the many factors that influence the pupil size such as light density, distance, breathing frequency etc. [8]. Luckily, if light is controlled for, pupil dilation seems to be quite a reliable measurement [8] and changes in pupil size correlate with changes in mental workload [5].

In the next section, I will first explain the experimental setup and methodological considerations, then give a succint step by step description of the followed procedure. Finally, I will explain how we implemented the user model.

METHOD

Participants and Design:
8 Participants (6 male, 2 female, ages between 20 and 35, all healthy adults with normal or corrected to normal vision) were randomly selected among the students and staff of the Faculty of Science and Engineering (University of Groningen). None had previous experience with the game. Note that generalizabilty was not of interest, but to determine if pupillometry is a viable option for adaptable software. Nontheless, it is reasonable to assume that similar results as evidenced in this study can be generalized to similar participants. The order of application of experimental and control condition was counterbalanced by gender and subject number. The dependend variables were the survival rate (measured in how many times the participants lost during the game) and the relationship between pupil dilation (pupil diameter, in pixels) and "gravity", which was determined through visual inspection of the plotted time-series.

Materials and Procedure:
The experiment was conducted in the eyetracking laboratory in the Faculty of Science and Engineering. During the experiment, we opted to turn off the lights, so the only light source would be the computers' LCD screen (brand unknown, screen resolution 1600x1200 pixles). The computer was a Apple Inc. product. The game source code was executed in the 'Psychopy' environment. We employed a headrest, to minimize head movements, typical for EEG and eyetracking experiments. We furthermore instructed the participants on the task, after which we proceeded with calibration. The employed eyetracker was a "EyeLink 1000" manufactured by SR Research Ltd. Relevant configurations were the following: Thresholds were set to "auto threshold", illumination power to 75%, pupil size data to "diameter" and sample rate to 250 Hz. We proceeded with gaze calibration as described in the eyetracker manual, pages 26-28. After the calibration, the games' main menu appeared on the screen. The participants had to click on "Start Game" and enter their subject ID (supplied by us), their gender and age. Once the entries were completed, a final screen appeard, reapeating the previous instructions. Once the participants were ready, any key-stroke would start the "baselining" phase. Once complete, experimental and control phases were applied as described in the "Design" section. The experiment concluded after the successfull completion of all conditions.

Task Characteristics
General Characteristics: We programmed the video game, used in our task, in the python programming language, making use of the "pygame" library [9]. The game source code is avialable on-line [11]. In addition, we made use of the "pylink" library (part of the "Psychopy" IDE) in order to make API calls (requesting information about the pupil size) to the eyetracker computer. As visible in an screenshot of the game in figure 1, at the bottom is the player, which has to dodge the falling blocks. Should the player fail to do so, the game freezes for 1 second, displaying the message: "You died horribily".

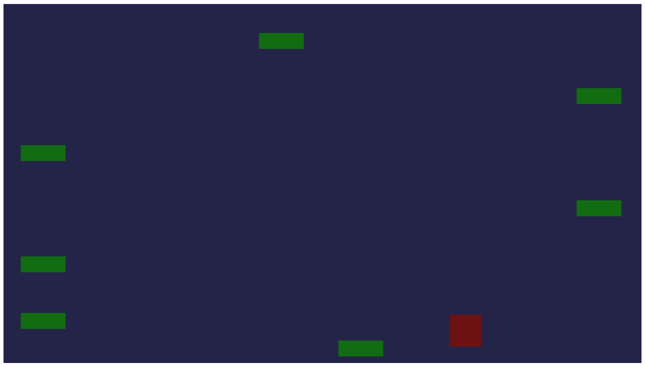

Figure 1. A screenshot from the game during any phase of the experiment. The player is represented by the reddish square. The goal is to avoid the falling green rectangles.

The game was run at 30 frames per second. The amount of blocks, the participants had to dodge was held constant (maximum 12 blocks on the screen at all times). We divided the upper half of the screen in 8 columns in order to build a coordinate system in which the obstacles spawn randomly with equal probability. The only limitation is that they can not spawn on top of each another (occupy the same space). As faster the blocks fall (as higher the difficulty is), the bigger the vertical distance between the blocks. We included this so the game would be cognitively more demanding at higher difficulties, but still solvable. Special care was taken with controlling the luminance of the screen. The RGB values of each elements add up to 144 points, and as there is always the same amount of objects on-screen, constant luminance was achieved.

The baselinging phase started immediately after the instruction screen. In figure 2, the red line shows how we lowered the games' difficulty each 2 seconds by 2.5 pixels/second and the blue line how pupil dilation reacted to the difficulty (note that in order to make game difficuly and pupil dilation comparable, their values had to be z-transformed). A slight downdrift of the pupil dilation is already noticeable.

During this phase, we determined the "threshold", a central component of our model. According to the literature, human pupils require more that a second to adjust for certain conditions [10]. This is why each "step" lasts two seconds, so the pupil could adjust to the difficulty level. Furthermore, we discarded blinking artifacts and took only the last 10 pupil dilation measurments for each difficulty "step" into account. Our aim was to ensure to work only with "adjusted" pupil sizes, wich we denote as B. The threshold is denoted as T, and was determined as in equation 1:

$$T = \frac{1}{n} \cdot \sum_{i=1}^{n=210} B_i + \sigma_B \cdot \alpha \qquad (1)$$

Where α is the critical value from the standard normal distribution, such that $P(z_{1.28}) \approx 0.1$ and σ_B the standard distribution of B. We chose this α such that the expected probabilty of pupil dilations higher than T is 10%. Note that

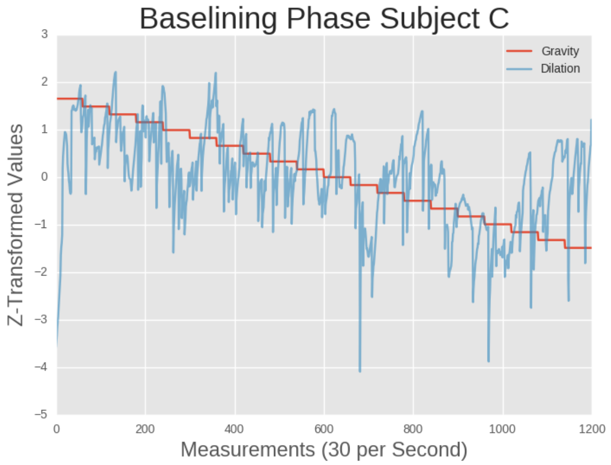

Figure 2. In this phase, difficulty gets gradually lowered each 2 seconds by 2.5 pixels per second, starting at 62.5 and stopping at 12.5 . Total duration of the phase: 42 seconds. This graph is repersentative for all subjects.

we assume that B is normally distributed and homoscedastic, which is doubtful due the random nature of our task generation and individual differences among participants. Nevertheless, we found this approach to work well. Once the threshold was determined, it could be used in the experimental phase.

The experimental phase is initiated with Difficulty (gravity, denoted as G) equals 15. Our model would rise the difficulty of the game each timestep of 2 seconds (denoted as κ) by 10% ($G \cdot \tau$ where $\tau = 0.1$), which is the base difficulty increment $\Delta_{G_t+1} = G_t \cdot \tau$. While this operation is performed for each timestep κ, the amount of violations of the threshold T will affect Δ_{G_t+1}. This modifier is calculated as depicted in equation 2.

$$\Delta_T = G_t \cdot \frac{1}{\kappa \cdot F \cdot \tau} \cdot \gamma \tag{2}$$

Where F is the current framerate (30 frames per second), κ the update time step (2 seconds) and τ the learning rate of 0.1 and γ is the amount of threshold transgressions. The final amount the game difficulty is modified is depicted in equation 2 and 3.

$$G_{t+1} = G_t \cdot (G_t + \Delta_{G_t+1} - \Delta_T) \tag{3}$$

When taking a nearer look at equation 2, we can see that when 6 threshold transgressions occurr, $\Delta_T = G_t$. In other words, we expected to occurr at least 6 transgressions each κ in general. The higher γ (past the expected 6), the lower the gravity will be set. The result of this approach can be seen in figure 3.

The control phase was necessary in order to be able to compare to our experimental condition. This way, we are able to determine if our user model actually had an impact on the subjects task performance. During this phase, pupil dilation measurements were disregarded and G was set at random each κ. G-values ranged from 15 to 50. As with the experimental condition, the task performance, measured in how many times the subjects lost the game, was logged. The total duration of the phase was 182 seconds.

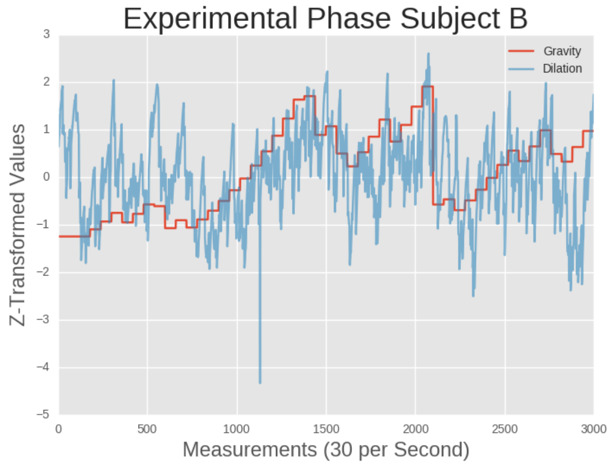

Figure 3. In this phase, difficulty gets determined by the pupil dilation. The user model could approximate the difficulty (gravity) quite closely to the pupil dilation. Total duration of the phase: 182 seconds. This graph is repersentative for all subjects but subject 'F', which disengaged from the task and had to be disregarded from analysis.

Data Analysis

Survival rates (task performance) both from experimental condition and control condition were averaged and compared. A linear mixed effects model was used in order to determine significance of the difference between the two conditions. For reproductibilities sake, the complete data set and code for data analyis is available at [11].

RESULTS

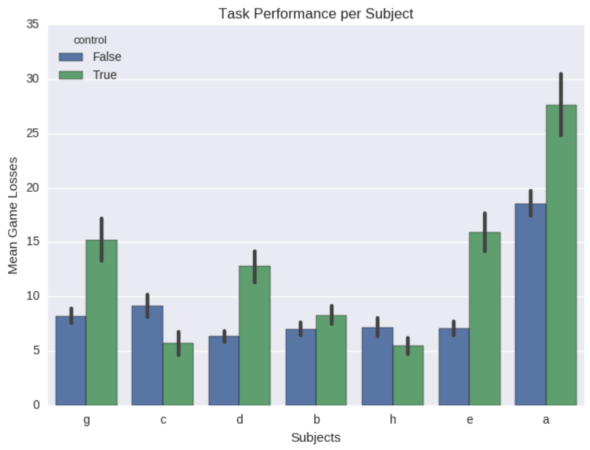

Figure 4. The blue bars represent the experimental, the green the control condition. For 5 out of 7 subjects, task peformance is clearly better (lower mean game losses). For a prototype, these are very promising results.

Figure 4 depicts how the subjects compare in task performance between the experimental and control condition.

Subjects ‘C’ and “H” performed worse in the experimental condition than in the control condition. The rest of the subjects performed better. The standard deviation remains low for all means. The linear mixed effects model $task\_performance \sim conditions * (1|Subjects)$ yielded $P(t_{2.404}) \approx 0.002$, which means that the factor “condition” is highly significant (we reject the null hypothesis of equality of groups). In figure 5, it is apparent that the task performance during the experimental condition (the “packed” lines at the bottom) is higher and less variant than in the control condition. It is interesting to note how similar these lines behave, not only in variance, but also in slope. This is not the case of the lines from the experimental condition, where each line increases at a different rate.

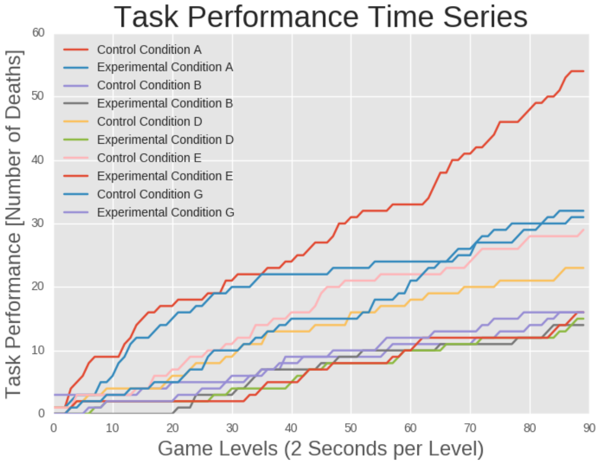

Figure 5. Time series of Task Performance. Subjects “C” and “H” have been excluded from this plot in order to visulize better the effects of the experimental condition.

DISCUSSION AND CONCLUSION

The results show that the workload management system adapted the difficulty of our video game, based on pupil size measurements, *reasonably* well. Not all participants performed better in the experimental condition, but *most* did, which is already a promising result which speaks in favor of our system. Now of course the question remeains: Why is that? To find out we would have had to log not only how many times a participant “dies horribly”, but also when (the exact time). This would allow to parse game losses with pupil dilation and to gain further insight in task resolution. Another factor that makes analysis complicated is the fact that for each participant, the task is different due to the random generation of the obstacles (green rectangles, see figure 1). Each participant faced another distribution of obstacles, this is also the reason why we could not plot averaged pupil dilation time series or just mean pupil dilations. While the *coupling* between pupil dilation and game difficulty, like in figure 3, was the same for all participants, the exact task properties was not. This might also account for the worse performance of subjects “C” and “H”. Maybe their task was just harder by chance or they disengaged from the task too much. Interesting is that with such a random task, the task performance would develop like in figure 5. Experimental task performance was much more invariant, which means that our user model not only reduced the *amount* of game losses, but also their *dynamics*. This pattern strongly speaks in favor of our system, as in many domains not necessarily outcomes are interesting, but the process of task execution (better drive controlled and save, as one “failure” could lead you to “die horribly”). So, not necessarily cuantity, but quality is important, and our system seems to cause more invariant and controlled task executions. We think that the present work shows the potential of adaptive systems not only in video games, but for any continuous, attention hungry tasks in many domains. We just need the machines make understand us a little better.

REFERENCES

1. Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., and Babiloni, F. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews 44* (2014), 58–75.

2. Byrne, M. D., and Pew, R. W. A history and primer of human performance modeling. *Reviews of human factors and ergonomics 5*, 1 (2009), 225–263.

3. Hart, S. G., and Staveland, L. E. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in psychology 52* (1988), 139–183.

4. Just, M. A., Carpenter, P. A., and Miyake, A. Neuroindices of cognitive workload: Neuroimaging, pupillometric and event-related potential studies of brain work. *Theoretical Issues in Ergonomics Science 4*, 1-2 (2003), 56–88.

5. Kahneman, D., and Beatty, J. Pupil diameter and load on memory. *Science 154*, 3756 (1966), 1583–1585.

6. Marstall, J., Miller, M. E., and Poisson III, R. J. Collaboration in the cockpit: Human–system interaction beyond the autopilot. *ergonomics in design 24*, 1 (2016), 4–8.

7. North, R. A., and Gopher, D. Measures of attention as predictors of flight performance. *Human Factors 18*, 1 (1976), 1–14.

8. Schwalm, M., Keinath, A., and Zimmer, H. D. Pupillometry as a method for measuring mental workload within a simulated driving task. *Human Factors for assistance and automation* (2008), 1–13.

9. Shinners, P. Pygame - python game development. Retrieved from `http://pygame.org/`, (2011).

10. Wang, J. T.-y. Pupil dilation and eye tracking. *A handbook of process tracing methods for decision research: A critical review and users guide* (2011), 185–204.

11. Wirthlin, M. Endlessrunner repository. Available on `https://github.com/Seneketh/repo_user_models`, (2016).

User Models Work Division

Teun:

- Game Base Code (Setup)
- User Model Logic and Implementation
- Eyelink to model
- Debugging code
- (check the github)

Marco:

- Experimental tools like datalogging
- Obstacle generation, OOP polish
- Text on screens, menus
- Eyelink to model
- Debugging
- (check the github)

Daniel:

- Literature search
- Experiment development
- EL calendar checks

We were all present during the experiments.