Improving Cognitive Load Level Measurement through Preprocessing of Psychophysical Data by Random Subspace Time-Series Method

Nada Attar, Paul Fomenky, Wei Ding, Marc Pomplun
Department of Computer Science
University of Massachusetts
Boston, USA
{nattar, pfomenky, ding, marc}@cs.umb.edu

***Abstract*— It will significantly facilitate cognitive load measurements to design a filtering method that reduces the error in the preprocessing of eye data while making a maximum of data available for analysis. We discuss a new approach that uses a Random Subspace (RS) ensemble method with sequential window frames of different sizes over the time course of the experiment to preprocess data for predicting the level of cognitive load. To investigate the suitability of the RS method, we carefully administrated two different visual search tasks imposing different levels of cognitive load to thoroughly evaluate the proposed method. Using our new filtering method, samples were evenly removed from both conditions (Nada: It is not clear to me what those two conditions are?), producing balanced datasets. In our experiments, the RS preprocessing method kept 85% of the original dataset, compared to 68% for the conventional baseline method method. Furthermore, the RS filter method is able to produce higher classification accuracy with regard to cognitive load level than the conventional approach. The results suggest that using machine learning in designing preprocessing techniques, instead of rigidly using a given hard threshold, to filter human psychophysical data can effectively improve cognitive load measurements.**

Index Terms*-- *Cognitive load; data preprocessing; machine learning; psychophysical experimentation; pupil dilation

I. Introduction

Being aware of a user's mental status is an important step in making interacting systems that are more usable, i.e., impose lower mental workload and stress level on their users. A suitable indicator for such demands is cognitive load, which reflects the user's effort while completing particular tasks [7]. Obtaining accurate measurements of cognitive load level in preprocessing the data before analysis is crucial for the success of such efforts. Most studies of human behavior reject trials that have incorrect responses, for example, when participants failed to find the target in a search task. Removal of faulty data is performed at the data preparation stage, where trials, blocks of trials, or even entire subject's data sets are excluded from analysis if the subject's performance falls below a given threshold, based on the assumption that these data contain more noise than informative data as reference [22]. However, this assumption using a given hard threshold may mistakenly discard valuable informative data along with noise.

Using an algorithm for selecting relevant data from psychophysical experiments such as pupil or EEG data can be a useful solution to avoid human error and to reduce the number of features that need to be analyzed. A large number of algorithms have been proposed for feature selection from human data [2]. In reference [17], Qian et al. studied visual target detection events, developing a pupillary response feature extraction method that helped select useful pupil data to improve the analysis based only on EEG signals. To assure the robustness of the method for selecting pupil data and EEG channels, they applied a two level linear classifier to obtain cognitive task related analysis of EEG and pupil responses. Their decision to use a finite pupil data sample and EEG features selected by those classifiers improved the classification performance during the analysis stage.

The true characteristics of psychophysical data are difficult to define, and some tasks induce higher cognitive load or stress. For example, we cannot be sure what the participants are doing or what kind of response the pupil size is indicating. Hence, it is useful to consider a method to classify the data rather than interpreting raw psychophysical data. Thus, defining such quantitative exclusion criteria is important as using human judgment may bias the analysis and possibly leave significantly more samples in some experimental conditions than in others. In the following, we will refer to this type of analysis as the conventional method.

To help filtering out data that only adds noise and does not reflect cognitive processing, it is important to measure relevant features accurately and in real time. Many studies have used pupil size as a useful, quick feature and reliable indicator of a person's cognitive load [16]. Measuring this variable allows the study of moment-to-moment deployment strategies during a given task and the monitoring of cognitive load during the course of that task [6], [15], [21], [22]. In typical laboratory tasks, many factors can easily influence the observer's pupil size. This includes ambient luminance, arousal, or any emotional stimulus content. Such interferences need to be

Commented [WD1]: Nada, Could you write a stronger sentence here? I believe the RS method does not simply evenly remove samples from both conditions. Could you discuss it more technically in principle why the RS method can produce a balanced dataset?

Commented [WD2]: Please be specific. What is the "conventional method"?

reduced in order for them not to impede cognitive load measurement. Pomplun and Sunkara [16] conducted an experiment to compare the effects of cognitive workload and display brightness on pupil size during a visual monitoring task. Other studies measured cognitive load under luminance changes during an arithmetic task [11]. These experimental results proposed a fine grained approach for cognitive load measurement under laboratory conditions that involve changes in the visual properties of the stimuli.

During data preprocessing, in order to reduce the influence of pupil tracking data that is unrelated to the cognitive load on the classification results, we used a Random Subspace (RS) ensemble classifier, employing the error rate of small size intervals from pupil data time series during an ongoing experiment. RS has been found to work well for problems with large dimensionality and excessive feature-to-instance ratio [12], [13], [18]. To enable a theoretical analysis of RS time series, we assumed that only a small (known) proportion of the features are important to the classification (pupil size and type of two conditions). The remaining features, such as reaction time (RT), response key, and blinks were not considered substantial.

II. Pupil Data Measuring during Visual Search Task

Visual search is arguably the most common task for measuring cognitive workload [14]. It has therefore received substantial research interest to understand the link between cognitive load and search performance [9], [10], [19]. When individuals perform a visual task, pupil size appears to be a function of the cognitive effort and attention required [8], and it specifically reflects the cognitive effort required to perform complex visual tasks [3], [5], [20], [23]. In reference [8], Porter, Troscianko, and Gilchrist, used pupillometry to study visual working memory load during visual search. Their visual search and counting task experiments manipulated search difficulty by varying the number of distractors as well as the heterogeneity of the distractors, and the dilatory patterns were compared between the two tasks. The results indicated an almost constant, large pupil size during the counting task. In contrast, during search, pupil size increased from the start of the trial onward, which the authors interpreted as showing increasing cognitive load as the search progressed.

Our study collected pupil data in the same procedure of the study by Attar et al. [19]. In their experiment, the authors used mean pupil size as the main variable to test how cognitive load varies during search based on the auditory feedback on task performance that subjects receive. That technique was able to improve the cognitive engagement as suggested by the greater effort being devoted to the search task, even though the fixation time on the items in both conditions was similar. The result of their study showed that providing online instant feedback during the sequential search task improved search performance and helped subjects find the next target faster. Interestingly, this increase in search efficiency was not due to the longer pause duration associated with target responses as suggested by other studies on visual selection, in which the longer planning time could result in a better target selection [1], [4]. Instead, the better performance was likely due to the greater cognitive effort. That is, when the feedback was provided, the pupil size increased, indicating that more attentional resources may have been devoted than when only neutral sound was provided in the control condition.

III. Objectives of the Study

The current study investigated the suitability of using the random subspace machine learning method in preprocessing the raw data and compared it to the conventional method. We collected pupil data from a multiple-target visual search task with two conditions that require different levels of mental effort. Since the two conditions induce different levels of cognitive load [19], time series defined by windows from different consecutive trials were expected to contain upward or downward steps in cognitive load. Intuitively, classification to filter the trial time-series could be viewed as the detection of these steps. The resulting data was then used to measure mean pupil size for each of the experimental conditions to see if filtering the data could increase accuracy in measuring different levels of cognitive load than filtering manually.

IV. Experiment

A. Data Collection

The data was collected from 17 healthy 18 to 35 years old volunteers who signed consent forms before the experiment. All had normal or corrected to normal vision. Each subject received a $10 honorarium. Eye movements were tracked and recorded using an SR Research EyeLink-2k system. Its sampling frequency was set to 1000 Hz. Stimuli were presented on a 22-inch ViewSonic LCD monitor at a viewing distance of 50 cm. Its refresh rate was set to 75 Hz and its resolution was set to 1024 x 768 pixels. Participant responses were entered using a keyboard.

A total of 120 displays (1024 x 768 pixels) were generated by a MATLAB script. Each display was composed of 32 Gabor patches (27 distractors and 5 targets) pasted on a gray background, each with a radius of 1 degree. The targets were oriented either vertically or horizontally. The distractors were oriented randomly with a minimum angular difference of 12° from both the vertical and the horizontal orientation. The orientations of targets in the same trial were identical and randomly selected. To make sure that the objects did not overlap, we set a minimum distance of 3 degrees between the centers of any two Gabor patches. A sample stimulus is shown in Figure. 1.

To investigate the effect of instant auditory feedback on search performance, two feedback conditions were tested: One was an auditory feedback condition in which two types of sounds were used to indicate whether the response subjects made (about finding an individual target and fixating on a target) was correct or not. The other was a control condition in which subjects always received an identical sound whenever they made a response. Subjects performed 10 trials per block and 6 blocks in each condition (1 for practice and 5 for the actual test), and both conditions were presented in an alternating order. Every trial was followed by a 3-second blank screen as a baseline so that pupil size had time to go back to its "resting" state and measurement of pupil size was not influenced by the preceding trial.

Commented [WD3]: This assumption is not justifiable. A reviewer may attack your assumption here. Could you re-phrase this sentence?

Commented [WD4]: If a number is larger than 9, we can just use its number format.

An instruction screen was shown before each block to inform subjects about the experimental condition's type. Each image display was only presented once to each subject, either in the auditory feedback condition or in the neutral sound condition.

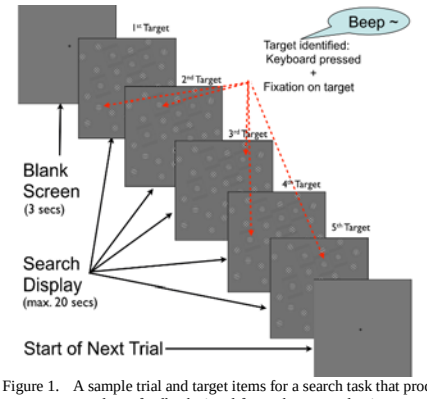

Figure 1. A sample trial and target items for a search task that produces sound as a feedback signal for each target selection.

During a trial, subjects were instructed to search for the five targets and press the 'x' key on the keyboard whenever they fixated on one of the targets. After they had heard a sound (either a feedback or neutral sound), they were to continue searching for the next target in the display. The next trial would begin once subjects pressed the space bar to indicate that they had found all targets or the stimulus had been shown for 15 seconds (timeout).

Subjects were informed that every trial had exactly five targets. They were trained to fixate on the target while they hit the response key. Some fixations landed on the blank area rather than on any search item. When this happened, we assumed this fixation to be aimed at the nearest item, i.e., the one whose center had the shortest Euclidean distance to the current fixation location. If the Euclidean distance was greater than a threshold which was set to 3.8 degrees, the fixation was not assigned to any of the items. A target response was considered a hit if subjects fixated on the target while making a response on the keyboard. A miss was counted if a selected item had already been previously selected during the same trial (a revisit), or the fixation during the response was on a distractor. If a subject found all five targets, the trial was considered as a passing trial; otherwise it was marked as a failing trial.

B. *Labeling the Data*

The class labels were of two control types (1 is used for the auditory feedback condition class; -1 is used for the neutral sound condition class). We also extracted the response which was either 1 for passing a trial with all targets being found or 0 for failing if any of the five targets were missed. MATLAB was used for extracting the average size of the pupil from the eye tracker data.

V. Data Preprocessing

A. *Using the Conventional Method*

The data was analyzed with the same technique that previous studies had used, in that only the correct trials in which participants found the five targets and hit the correct key were included. This was decided as trials with responses to less than five targets might add noise to the analysis. The responses were identified as either a hit or a miss based on the button press.

This technique of preprocessing the data collection described earlier left us with an unequal proportion of correct trials between conditions. Overall correct trials for participants in the auditory feedback condition was 79.31%, while they only achieved 67.54% correct trials in the neutral sound condition, which was a significant difference of more samples remaining in the in auditory feedback condition than in the neutral sound condition, $t(16) = 3.83$, $p < .05$.

B. *Using the RS Time-Series Method*

Although the data set had several features (pupil size, user response, blinks, reaction time, and class label), we only used pupil size to classify the cognitive load by predicting the condition (auditory feedback condition vs. neutral sound condition). For each trial, the percentage of average pupil dilation relative to the baseline at the beginning of each trial was computed. Subsequently, all the data was aligned in the same temporal order in which they were measured during the experiment. There were 120 trials in 12 alternating blocks. Every block consisted of 10 trials of one condition, leading to a total of 120 pupil data samples per subject (see Figure. 2).

We applied the RS method using a window that takes an interval of trials from every two successive (and alternating) conditions. We chose windows spanning 4, 6, 8, and 10 trials. Every window size was tested with RS on the time series of the experiment. The center of the window was placed at the middle between every two blocks as it is shown in Figure. 2, i.e., covering half of the trials from the auditory condition and half of the trials from the neutral condition. For example, a window of size 8 covered the 4 last trials from block *i* and the 4 first trials of block (*i+1*).

All the trials inside the window were classified using RS. Then, the window slid to the middle of the next two alternating blocks (if a window covers the last trials of block *i* and the first trials of block (*i +1*), then after the shift it will cover the last trials of block (*i +1*) and the first trials of block (*i +2*)). The classification was performed over the course of the entire experiment. A high accuracy is reported when the RS was able to predict the class label (auditory feedback condition or neutral sound condition) based on the pupil dilation data within the window. The pupil measurement is expected to fluctuate between each two blocks, indicating different cognitive loads for different types of conditions. If the accuracy was above 50% for the trials inside a window, we marked those trials as passing trials. Otherwise, the entire window content was marked as failing trials. We repeated this method for each subject and used all four window sizes. The data from failed trials is shown in Table I, including the proportion of rejected trials for RS using window interval time-series and the conventional method using user response in the auditory condition and the neutral

Commented [WD6]: What are those "previous studies"?

Commented [WD7]: Again, it is not clear what is the "data collection described earlier"?

Commented [WD8]: I think that you want to say that more informative samples are used for the trials, the higher accuracy could be achieved.
Could you make this paragraph clearer because it is a very important empirical observation?

Commented [WD9]: Be consistent with size 8

Commented [WD5]: It is a bit confusing why you used 1 and -1 for the two control types, but 1 and 0 for failed and passing trials?

condition. RS was able to provide equal numbers of rejected trials for both conditions while the conventional method provided imbalanced numbers.

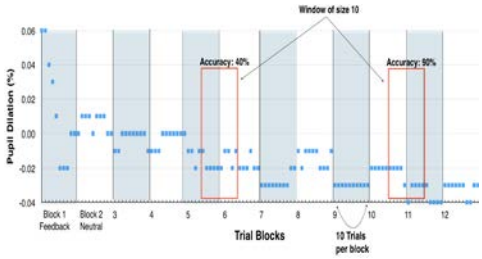

Figure 2. An example of one subject's time series data.

TABLE I. REJECTED TRIALS FOR BOTH CONDITIONS

Filtering Method	Rejected Trials (%)	
	Auditory Condition	*Neutral Condition*
Conventional method	20.69	32.46
RS window size 4	36.33	36.33
RS window size 6	55	55
RS window size 8	15.3	15.3
RS window size 10	20	20

VI. COGNITIVE LOAD MEASUREMENT

To further investigate how the cognitive resources were distributed throughout the search process, we examined the participants' average pupil size when they performed the search task after we had preprocessed the data using the RS and the conventional method.

A. Using the Conventional Method

We compared the pupil size to investigate whether different cognitive effort was devoted depending on the two conditions during the continuous visual search. For this purpose, the mean pupil size (in pixels in the camera image) was computed during each correct trial only. The incorrect trials that we defined by the user response in Table 1 were rejected from the measurement. Pupil dilation was significantly greater in the auditory feedback condition than in the control condition, $t(16) = 2.98$, $p < .05$.

B. Using the RS Time-Series Method

We rejected the trials inside any window that did not have an accuracy higher than 50% for each subject as shown in Table 1. For each subject we measured the average pupil size during the correct trials. Paired sample t-tests were conducted to analyze the difference in mean pupil size for each subject's data generated by the new methods. This difference in pupil size between the auditory feedback condition and neutral sound condition for the RS methods with window intervals was significant, all $ts(16) > 2.56$, $ps < .005$. Figure. 3 shows mean pupil size (in pixels in the camera image) for each experimental condition. The error bars indicate standard error of the mean. The conventional method and RS time-series with different window sizes were used to preprocess data. The pairwise differences between the results of the RS methods were not significant, preventing any conclusions about which window size is best to use for filtering data, all $ts(16) < 1.38$, $ps > .1$. However, there was a greater difference in pupil dilation between the auditory feedback and neutral sound conditions when data was removed according to the RS method than when using the conventional method. The difference in pupil size between two conditions using the conventional method was 23.5, while for the RS method it was 43, 45, 37.7, and 47.5 pixels for window sizes of 4, 6, 8, and 10, respectively.

This finding suggests that using our machine-learning based method could better filter the data thus led to greater accuracy in distinguishing different levels of cognitive load than filtering manually, depending on the correctness of user responses.

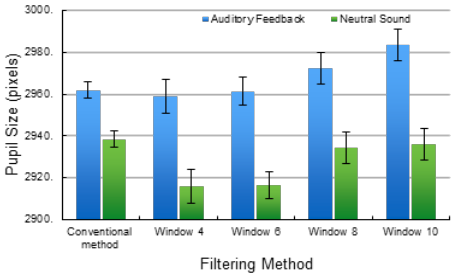

Figure 3. Mean pupil size according to each preprocessing method

VII. CATION APPROACH

In this section, we used different classification approaches to evaluate detecting different cognitive load levels using the dataset filtered by the RS method and the dataset filtered by the conventional method. This section also gives more insight into the effect of window size on the results of the RS method. Ten-fold cross-validation was used for evaluating four classifier approaches.

We report the results in terms of the accuracy achieved from each of the filtered datasets. Using classifier approaches to detect different levels of cognitive load in the resulting dataset using each of the window sizes resulted in better accuracies than using the conventional method. Table II shows the classification results for detecting levels of cognitive load from mean pupil size for each dataset using the new different RS methods and the conventional method. These results indicate that the new method leads to better detection of different levels of cognitive load, regardless of the window size used. The results show that we can achieve reasonable classification accuracy, if we use an automatic learning technique to filter the data instead of using the conventional method.

TABLE II. CLASSIFICATION ACCURACIES USING THE FILTERED DATASETS

Data Resulted from Classifier	Accuracy of Detecting Cognitive Level of Four Classifier Approaches (%)			
	Random Subspace	***Decision Tree***	***Logistic Regression***	***Multilayer Perceptron***
Conventional method	65	80.7	65.25	55.98
RSM window size 4	77	84.9	84.01	56.86
RSM window size 6	75.6	83.68	82	61.17
RSM window size 8	84	85.98	83.92	58.52
RSM window size 10	85	86.71	84.50	60.73

VIII. CONCLUSIONS

Our study provided an effective method to reduce errors related to preprocessing of pupil data before analysis, where a machine learning implementation using the random subspace method showed better outcomes than the conventional method. This work was motivated by the need for additional methods for evaluating implicit physiological features, such as pupil dilation, to measure cognitive load and to include as much data as possible in the analysis. The study showed that choosing different sizes of the window in the data preprocessing can effectively remove samples of pupil data that are unrelated to workload. Furthermore, this method removes the same amount of samples from the two tasks and balances the data that is used in the subsequent analysis.

Our implementation could be very useful for experiments where we cannot be sure what the participants are doing or what kind of response the pupil dilation is indicating, allowing for the study of the behavioral result of these experiments. Prior research efforts depended on the user response using the keyboard to consider the data as valid to reflect the cognitive load. In our study, using the pupil size feature to predict the cognitive load succeeded and led to a sample size that was larger than for the conventional method with window sizes 8 and 10. Further study is required to find an effective way for choosing the optimal window size for the classifier.

We have proved that machine learning techniques can be applied to significantly facilitate the measurement of cognitive load in two different visual search tasks. As an extension of this work, the method we described could be applied to classifying workload in other contexts such as arithmetic problems or reading tasks. Moreover, it may be possible to detect multiple, finer-grained cognitive load levels. We believe that these findings have potential applications in designing interfaces that use the state of a user's cognitive load rather than a user's manual response, or for smart interfaces that adapt to a user's levels of cognitive load.

REFERENCES

[1] A. Cohen & R. B. Ivry, (). "Density effects in conjunction search: evidence for coarse location mechanism of feature integration," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 17, pp. 891-901, 1991.

[2] A. K. Jain and D. Zongher, "Feature selection: evaluation, application and small sample performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19 (2), pp. 153–158, 1997.

[3] A. Nuthmann and E. van der Meer, "Time's arrow and pupillary response," *Psychophysiology*, vol. 42(3), pp. 306-317, 2005.

[4] C.C. Wu and E. Kowler, "Timing of saccadic eye movements during visual search for multiple targets," *Journal of Vision*, vol. 13(11), pp. 1–21, 2013.

[5] C. Karatekin, J. W. Couperus, and D. J. Marcus, "Attention allocation on the dual task paradigm as measured through behavioral and psychophysiological responses," *Psychophysiology*, vol. 41, pp. 175-185, 2004.

[6] E. H. Hess and J. M. Polt, "Pupil size in relation to mental activity during simple problem solving," *Science*, vol. 143, pp. 1190-1192, 1964.

[7] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. M. Van Gerven, "Cognitive load measurement as a means to advance cognitive load theory," *Journal of Educational Psychology*, vol. 38(1), pp. 63-71, 2003.

[8] G. Porter, T. Troscianko, and I. Gilchrist, "Effort during visual search and counting: insights from pupillometry," *Quarterly Journal of Experimental Psychology*, vol. 60, pp. 211–229, 2007.

[9] J. Palmer, C. T. Ames and D. T. Lindsey, "Measuring the effect of attention on simple visual search," *Journal of Experimental Psychology*: Human Perception and Performance, vol. 19(1), pp. 108–130, 1993.

[10] J. Schwark, J. Sandry, J. MacDonald and I. Dolgov, "False feedback increases detection of low prevalence targets in visual search," Attention, *Perception, and Psychophysics*, vol. 74(8), pp. 1583-1589, 2012.

[11] J. Xu, Y. Wang, F. Chen, and E. Choi, "Pupillary response based cognitive workload measurement under luminance changes," *International Conference on Human-Computer Interaction (INTERACT'11)*, pp. 178-185, 2011.

[12] L. I., Kuncheva and C. O. Plumpton, "Choosing parameters for random subspace ensembles for fMRI classification. MCS," Lecture Notes in CS, vol. 5997, pp. 54-63, 2010,

[13] L. I. Kuncheva, J. J. Rodriguez, C. O. Plumpton, D. E. J. Linden and S. J. Johnston, "Random subspace ensembles for fMRI classification," *IEEE Transaction on Medical Imaging*, vol. 29(2), pp. 531-542, 2010.

[14] M. Bravo & K. Nakayama, "The Role of Attention in Different Visual Search Tasks," Perception and Psychophysics, 51, pp. 465-472, 1992.

[15] M. Carrasco, Visual attention: the past 25 years," *Vision Research*, vol. 51, pp. 1484–1525, 2011.

[16] M. Pomplun and S. Sunkara, "Pupil dilation as an indicator of cognitive workload in human-computer interaction," *International Conference on HCI*, pp. 542-546, 2003.

[17] M. Qian, M. Aguilar, K. N. Zachery, C. Privitera, S. Klein, T. Carney and L.W. Nolte, "Decision level fusion of EEG and pupil features for singletrial visual detection analysis," *IEEE Transactions on biomedical Engineering*, vol. 56(7), pp. 1929-1937, 2009.

[18] M. Skurichina and R.P.W. Duin, "Bagging, boosting and the random subspace method for linear classifiers," *Pattern Analysis and Applications*, vol. 5, pp. 121–135, 2002.

[19] N. Attar, C. Wu and M. Pomplun, 'The effect of immediate accuracy feedback in a multiple-target visual search task', in *In Proceedings of the 36th annual meeting of the cognitive science society*, pp. 1868-1873, 2014.

[20] R. F. Stanners, M. Coulter, A. W. Sweet, and P. Murphy, "The pupillary response as an indicator of arousal and cognition," *Motivation and Emotion*, vol. 3, pp. 319–340, 1979.

[21] S. Ahern and J. Beatty, 'Pupillary responses during information processing vary with scholastic aptitude test scores', *Science*, vol. 205, pp. 1289–1292, 1979.

[22] S. Jynge, "Effort and pupil behavior in visual search task," Master's thesis, University of Oslo, Department of Psychology, 2010.

[23] S. P. Verney, E. Granholm and S. P. Marshall, "Pupillary responses on the visual backward masking task reflect general cognitive ability," *International Journal of Psychophysiology*, vol. 52, pp. 23–36, 2004.