Classification of Cognitive Load from Task Performance & Multichannel Physiology during Affective Changes

Sazzad Hussain¹,², Siyuan Chen¹, Rafael A. Calvo², Fang Chen¹

1 National ICT Australia (NICTA)
Australian Technology Park
Eveleigh NSW 1430, Australia
+61(2)83060751
{first.last}@nicta.com.au

2 School of Electrical & Info. Eng.
The University of Sydney
NSW 2006, Australia
+61(2)93518171
Rafael.Calvo@sydney.edu.au

ABSTRACT

Detecting cognitive load (CL) is important for implementing adaptive systems aware of the user’s mental workload. The aim of such systems is to reduce error related risks during task-critical operations or to facilitate computer based (e.g. tutoring systems) learning. Self reports, task performance, behavioral and physiological changes are some of the indicators of CL. However, affective factors which occur naturally in everyday life can influence these variables, making CL detection challenging. In this paper, we investigate machine learning techniques for detecting CL levels from task performance and physiological channels during affective changes with the support of systematic experimental setup. Classification results are presented for detecting two levels (low & high) of CL from individual modalities and their combination using a vote classifier. We have achieved around 66% to 78% accuracy for detecting the two CL levels under affective changes (valence/ arousal).

Keywords
Cognitive load, affect, physiology, machine learning, data fusion.

1. INTRODUCTION

Cognitive load (CL) refers to the executive control of working memory imposed by tasks. It is the reflection of pressure people experience while completing the tasks [1]. CL is closely associated with learning, therefore, maintaining and experiencing its optimal range is important to achieve high productivity.

CL detection is an important step in the area of human machine/computer interaction, especially for applications that require intensive mental activity such as driving, air/train traffic control, learning sessions with intelligent tutoring systems (ITS) [2] etc. Detecting and monitoring CL is suitable for developing adaptive user interfaces and systems, whether the aim is to reduce error risks during critical operations or to improve learning experiences e.g. during ITS interactions. Various methods and modalities such as self reports, performance ranking (e.g. reaction time, accuracy), behavioral signatures (e.g. speech, pressure, dialog patterns, eye movement etc) and physiological patterns (heart activity, skin response, respiration, pupil dilation etc) can be indicators of cognitive load [1]. Pattern recognition and machine learning techniques applied to a combination of all these modalities could lead to the most effective CL detection.

However, affective factors such as valence and arousal can not only have effect on task performance [3, 4] but may also influence behavioral and physiological responses [5, 6]. As part of human nature, emotional dynamics may arise during task activity in everyday life, therefore, the automatic detection of CL can be challenging under the presence of affective changes.

In previous studies, a number of techniques have been investigated to measure CL levels (for detailed review see [1]). At the same time research in the area of affect detection have progressed separately (for detailed review see [7]). Self reports, behavioral and physiological modalities have been more or less used in both communities for finding evidence of CL and affective states separately. There has been some work investigating the impact of viewing emotional stimuli during task activity and how arousal or valence (especially negative valance) affects task performance [3, 4]. However, there has not been much work on automatic CL detection under affective changes, especially using physiological channels (some exceptions include [5, 6, 8]) or using multimodal information. Detecting CL and finding their correlations with specific physiological signal is very challenging, therefore using features from a variety of physiological signal with other modalities and variables is desirable for better detection accuracy.

This research investigates how workload and emotion affect physiological measures. The study aims to explore machine learning techniques for detecting cognitive load and affective states using multimodal approaches. As part of the research, in this paper, we have considered physiological signals such as electrocardiography (ECG), galvanic skin response (GSR), respiration (RESP), eye activity (e.g. blink, pupil dilation), and task performance related information (response time) for detecting CL load under affective (valence & arousal) changes. This research presents the automatic detection of CL with the support of a systematic experimental setup, feature selection techniques, and machine learning approaches. This study is challenging due to the fact that physiological channels are influenced by both affective and cognitive factors. Despite the challenges, we investigate techniques for robustly classifying CL levels and adaptively handle affective changes which are unrelated to the tasks. In this paper, we describe the experimental protocol for data collection and report classification results for detecting low and high CL levels using the reaction time (recTime) to complete tasks and physiological features (physio) for individual subjects and all subjects combined. The first aim of this paper is to investigate the performance of CL detection using recTime and physio separately. The second is to report the impact of fusing recTime and physio (fusionModel) for CL classification.
2. RESEARCH METHODOLOGY

2.1 Experiment and Data Collection

Data was collected from 20 healthy volunteers (11 male & 9 female) age ranged from 22 to 48 years ($M=27.65$, $SD=6.94$). They signed an informed consent prior to the experiment and were rewarded with movie vouchers for their participation. The experiment lasted approximately 50 minutes. This included 15 minutes of preparation (sensor setup and the explanation of experiment protocol) and 35 minutes of data collection, while participants completed arithmetic tasks with emotionally stimulated photos from the International Affective Picture System (IAPS) collection [9] in the back.

A mental arithmetic task was designed where participants were asked to sum four numbers that were sequentially displayed on screen, and use their mouse to select the correct answer from 10 choices. Task difficulty levels were varied by the number of digits for addition and the number of carries produced by addition. Levels 1, 2 and 3 had one digit for each of the four adders but no carry, one carry and two carries produced during the addition process respectively, while levels 4 and 5 used two-digit addition with one carry produced at only the lower digit in level 4 and one carry generated at both digits in level 5 during each addition process. After a short training, each participant completed 7 sessions with a short break between each session. In every session, they were presented with 5 levels of tasks twice, and the sequence of the tasks was randomized. The first session used a plain grey background while the other 6 sessions used different images in the background. The images were selected from IAPS based on the arousal and valence normative ratings to induce emotions in a 3 arousal x 2 valence space. The light in the room was constant and the background image was not changed until participants had finished the task.

Eye activity was recorded using a FaceLAB 4\(^1\) (Seeing Machines) remote eye tracker system with a sampling rate of 60 Hz. Participants were free to move their head but instructed to keep their eyes within the screen display area. A webcam was placed on top of the screen, facing the participants. Due to the failures of eye calibration for 5 subjects, results are presented for 15 subjects in this paper. For collecting other physiological data, participants were equipped with ECG, GSR, and RESP sensors. BIOPAC MP150\(^2\) system with AcqKnowledge software was used to acquire the signals with a sampling rate of 1000 Hz (all channels). For collecting ECG, two electrodes were placed on the wrists. The GSR sensor was strapped around the index and middle finger (left hand) and a respiration band was strapped around the chest. Figure 1 shows the experimental setup with the sensors.

2.2 Computational Model

ECG, GSR, RESP signals were down sampled to 200Hz to reduce computation. Preprocessing of the signals and extraction of statistical features were done using the Augsburg Biosignal toolbox (AuBT) [10] in Matlab. Some features were related to the signal characteristics (e.g. heart rate variability, respiration pulse, frequency) and other features were common for all signals (e.g. mean, median, and standard deviation, range, ratio, minimum, and maximum). Matlab was also used for extracting eye features. Pupil and eye movement features (e.g. pupil size, fixation, saccade) were extracted from the data collected with the eye tracker and blink features (number of blinks, blink duration) were extracted from the video recorded with the webcam. A total of 181 features were extracted from the physiological channels. Figure 2, gives a snapshot of the visualization tool developed in Matlab for inspecting and annotating (optional) the offline raw physiological signals and face videos with corresponding affective states and task difficulty levels. This tool is also used for initiating feature extraction from the physiological channels.

Time stamps related to the task difficulty levels were synchronized with all the physiological channels and features were extracted using a 12 seconds window (same duration as the task activity). The reaction times/reaction time, which were measured from the instance of multi choice window onset to the first click on answers, were also extracted from the logs files. The feature vectors were then labeled with the corresponding task difficulty levels (level one to five). In this paper we investigate only low and high CL levels for classification. Levels one & two were merged and relabeled as lowCL and levels four & five were similarly relabeled as highCL. Level three was removed for this analysis; therefore, the computational model was trained and tested using a balanced class distribution, without applying any up or down sampling techniques.

\(^1\)FaceLab: [http://www.seeingmachines.com](http://www.seeingmachines.com)

\(^2\)BIOPAC: [http://www.biopac.com](http://www.biopac.com)
The Waikato Environment for Knowledge Analysis (Weka), a data mining package, was used for feature selection and classification. All physiological features were merged to obtain feature fusion. To reduce the dimensionality of the large number of features from physiology, CfsSubsetEval evaluator with BestFirst search was used for ranking and selecting the best features. The feature selection technique selected features that were best suited for separating the two classes (lowCL & highCL) and removed all other unnecessary features. We selected four machine learning algorithms; simple logistic, k-nearest neighbor (KNN), linear support vector machine (SVM), and decision trees for classification. Finally, a vote classifier for combining classifiers was applied with the majority vote rule. The training and testing for the dataset was performed separately with a 4-fold cross validation. Classification of lowCL & highCL was performed for individual subjects and all subjects combined with the following scenarios:

a. Classification using only reaction times (recTime).

b. Classification using only best physiological features (physio).

c. Classification after fusing reaction time and best physiological features (fusionModel).

3. RESULTS AND DISCUSSIONS

In this section we present the results for classifying the two levels of CL (lowCL & highCL). Firstly, we discuss about recTime and the physiological features. Secondly, we present the classification results for recTime, physio and fusionModel.

Paired two-tailed t-tests indicate that recTime was significantly higher for highCL compared to lowCL for all individual subjects (t(27)>2, p<0.05). This is obvious because a higher reaction time is expected for tasks with more difficulty.

As for physio, the feature selection algorithm (explain in previous section) is suitable for finding the best features that can separate the two CL levels. The feature selection algorithm selected different number of features from different subjects. There were more features from particular channels and few features (or none) from other channels. For the combined subjects, 10 features were selected and there was at least one feature from each channel.

This study gives preliminary evidence how the different channels contributed for classifying the CL levels. Table 1 gives a brief summary of the distribution of features selected from the individual channels. From the table, ‘Total features’ is the total number of features extracted from each subject. ‘Max features selected’ gives the maximum number of features that were selected from the individual subjects. ‘No. of distinctive features’ gives the number of unique features that were selected over all the individual subjects. ‘Features selected’ gives the name of the features that were selected from the combined subjects dataset. In the table, ecgPQ (range) refers to the range value of the p and q peaks of ECG. 1Diff and 2Diff stands for the first and second derivation of the GSR/RESP signal. Pulse stands for the respiration pulse rate from the signal. As for EYE, the pupil and blink represents the pupil diameter and eye blink respectively.

Table 1. Distribution of physiological features selected for CL classification

<table>
<thead>
<tr>
<th>Feature</th>
<th>ECG</th>
<th>GSR</th>
<th>RESP</th>
<th>EYE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total features</td>
<td>84</td>
<td>21</td>
<td>67</td>
<td>9</td>
</tr>
<tr>
<td>Max features selected (individual sub)</td>
<td>8</td>
<td>5</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>No. of distinctive features (all individual sub)</td>
<td>25</td>
<td>12</td>
<td>26</td>
<td>6</td>
</tr>
<tr>
<td>Features selected (combine)</td>
<td>ecgPQ (range)</td>
<td>gsr (min)</td>
<td>gsr1Diff (mean)</td>
<td>gsr2Diff (min)</td>
</tr>
</tbody>
</table>

The ZeroR classifier in Weka was used for calculating the baseline accuracy which was 50% for the two classes. Figure 3 gives the results for classifying lowCL & highCL using recTime, physio and fusionModel. The results are presented for datasets from the individual subjects and the combined subjects. For individual subjects, the performance of detecting lowCL and highCL from recTime (M = 72.38, SD = 9.91) is slightly higher than physio (M = 70.47%, SD = 12.78). The fusionModel (M = 78.33%, SD = 8.15) however, can improve the accuracy over both recTime and physio. The improvement for fusionModel is not extremely high but significant (table 2). For combined subjects the improvement of the fusionModel (75.47%) is notably higher compared to physio (66.19%) but marginally higher than recTime (74.64%).

![Figure 3. Classification results for detecting lowCL & highCL from reaction time, physiology and combine features.](image)

Table 2 shows how the integration (fusionModel) of reaction time with physiological features significantly improved the accuracy of detecting the two CL levels over the individuals. The t-test for reaction time and physiology indicate that the difference between recTime and physio is not significant. The recTime is a suitable reference for performance since higher reaction time is predicted by the CL models from higher CL levels (also supported by the statistical tests earlier). Therefore, the performance of CL level detection using physio is similar to recTime.

---


---
Despite the differences in experimental conditions and modalities, it is still useful to discuss our study in the context of others. Giroud et al. [11] achieved an average accuracy of 61.1% to classify easy and hard game-playing tasks using functional near-infrared spectroscopy (fNIRS). Haishfeld et al. [12] similarly used fNIRS and obtained 68% classification accuracy for low and high spatial workloads. Haapalainen et al. [13] in their study presented classification results for distinguishing low and high CL using physiological features during cognitive tasks. They have reported around 80% classification accuracy using physiological features. However, these studies have not considered affective changes during CL measurement. Even though the accuracy for physiology in our paper is slightly lower than Haapalainen detecting the CL level under affective changes is challenging and the accuracy we have achieved is quite satisfactory.

4. CONCLUSION
This research investigates the classification of CL levels using task performance related information and physiological signals under the influence of affective changes. Reaction time is a good indicator of CL levels however; the performance of physiology had similar accuracy. The fusion of reaction time and physiological features can significantly improve the accuracy of classifying low and high CL levels. Moreover, the accuracy of classification using combined subjects was similarly good, which can be a good motivation for developing subject independent models for CL detection. As an extension of this work, we can further investigate the following:

a. Accuracy of detecting more variety of CL levels (e.g. one, two, three, four, and five).

b. Identify features that are influenced by affect.

c. Accuracy of detecting CL levels under different affective (arousal/valence) levels or states.

d. Impact of affective changes on task performance.

5. ACKNOWLEDGMENTS
Sazzad Hussain was supported by Endeavour Award and National ICT Australia (NICTA). NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

6. REFERENCES


