Towards a Continuous Assessment of Cognitive Workload for Smartphone Multitasking Users

Angel Jimenez-Molina
University of Chile, ajimenez@dii.uchile.cl

Hernan Lira
University of Chile, hlira@ing.uchile.cl

Follow this and additional works at: https://arrow.tudublin.ie/hwork17

Part of the Artificial Intelligence and Robotics Commons, Graphics and Human Computer Interfaces Commons, and the Psychological Phenomena and Processes Commons

Recommended Citation
Towards a Continuous Assessment of Cognitive Workload for Smartphone Multitasking Users

Angel Jimenez-Molina and Hernan Lira

University of Chile, Faculty of Physical and Mathematical Sciences, Department of Industrial Engineering, Beauchef 851, Of. 603, Santiago, Chile
ajimenez@dii.uchile.cl, hlira@ing.uchile.cl

Abstract. The intermeshing of Smartphone interactions and daily activities depletes the availability of cognitive resources. This excessive demand may lead to several undesirable cognitive states, which can be avoided by continuously assessing the user cognitive workload. Recently, many attempts have emerged to assess this workload by using psychophysiological signals. This paper provides evidence that it is possible to train models that accurately identify in short time windows such cognitive workload by processing heart rate and blood oxygen saturation signals. This assessment could be applied in Smartphone notifications delivery, interface adaptations or cognitive capabilities evaluation.

1 Introduction

Multitasking is a usual behavior in people these days. Nowadays it is common to see people performing different tasks in the Smartphone while daily activities are being conducted, such as walking, talking or waiting the bus. Activities and Smartphone tasks demand the simultaneous use of several cognitive resources, such as perception, attention, short and long term memory, motor control, among others [11]. Excessive demand of cognitive resources can lead to distractions, increase errors, provoke stress and frustration, and reduce ability for mental planning, problem solving, and decision making [1, 6, 9, 10, 11].

In order to avoid these states, it is necessary to continuously assess the cognitive workload – the perceived level of mental effort that a task or multiple tasks induce in the person [15]. Typically, it has been evaluated using subjective rating methods applied after the person has already finished the task, making the evaluation dependent of her/his final perception [14]. The static nature of these methods makes them unfit for real-time evaluation [3]. In addition, they are also constrained by a reporting bias and lack ecological validity.

Recently, psychophysiological signals like cardiovascular or brain measures are being used to assess cognitive workload. This is based on the empirical evidence of the correlation between the physiological responses triggered by the nervous system and psychological stimuli [2]. The cardiovascular system is especially interesting for this purpose since it is highly sensitive to neurological and psychological processes [4].
In this paper, we leverage the capabilities of heart rate (HR) and blood oxygen saturation (SpO2) to research the possibility of assessing mental workload in real time during Smartphone interactions. In this regard, this paper attempts to answer the following research questions:

- **RQ1.** Are HR and SpO2 capable of discriminating between tasks with different levels of cognitive demand?
- **RQ2.** Is it possible to accurately classify on real time the users cognitive workload according to different levels of cognitive demand by combining HR and SpO2?

To answer these research questions, an experiment was conducted in which 50 users performed several tasks in different scenarios of cognitive resource demand.

This paper is organized as follows. Section 2 presents the related work. Methods are introduced in Section 3. Section 4 shows the experimental results. Discussion and conclusion are presented in Section 5.

## 2 Related work

The most comprehensive, quantitative existing method to assess cognitive workload is the multiple-resource model (MRM) of Wickens [13]. As shown in Table 1, it offers empirical evidence of three cognitive dimensions that can lead to competition and interference among cognitive resources if they are simultaneously used. The first dimension consists of three stages for processing task information, known as perceptual, central processing and response stages. The second dimension groups the cognitive resources demanded for each stage, e.g. selective attention, perception, working memory or motor control. The third dimension consists of a set of attributes that characterize cognitive resources. For instance, as shown in Table 1, the input modality - auditory, visual or tactile - and the processing code - spatial or verbal - jointly describe the selective attention.

The interference among attributes of simultaneously demanded resources increases the task difficulty, which generates different levels of performance. For instance, Oulasvirta et al. [11] proposes a framework based on Wickens’s MRM, in which presents evidence of competition due to the use of attention by both Smartphone interaction tasks and daily activities in outdoor environments. There are three types of interferences: due to the input modality, the processing stage and the processing code. The first one is complete and occurs when two or more tasks use the same input modality. The second interference can be complete or partial and occurs when tasks use resources from the same stage. Finally, the third stage is complete and is verified when there exists a conflict in the processing code.

Regarding subjective methods to assess cognitive workload, the most widespread example is the NASA Task Load Index (TLX), which measures the mental and physical performance, as well as the effort and frustration of the user [8].

Haapalainen et al. [7] assesses tasks of visual perception and cognitive speed using several sensors, including HR. Through a Naive Bayes classifier, the median of the heat flux and the average absolute median of the electrocardiogram,
they achieve an average accuracy of 81.1%. Similarly, Fritz et al. [5] measure
tasks performed by professional software developers using eye tracker, electro-
dermal activity and electroencephalogram. They obtain 84.38% precision with
features extracted from the three signals training a Naive Bayes classifier. Ryu &
Myung [12] use electroencephalogram, electrooculogram and electrocardiogram
to evaluate the resolution of math sums with different levels of cognitive de-
mand. The three studies mentioned above use the NASA TLX as the standard
to evaluate their experimental designs [8].

3 Methods

Participants. An experiment was conducted with 50 participants, all engineer-
ing students of University of Chile (33 men and 17 women; mean age = 22.4
years, SD = 2.8 years), recruited through the institutional news Web applica-
tion. None of them suffered from cardiovascular diseases or was taking medica-
tions that could have affected their normal behavior, HR or SpO2 levels. All of
them were familiar with the use of smartphones. They received CL$5,000 (about
US$10) for their participation.

Task Design. Each participant was required to perform four treatments de-
digned to induce low and high cognitive workload. Treatments were designed
according to the Wickens’s MRM. The first two treatments consist of two sub-
tasks - reading a simple e-mail (ST1) and replying a simple e-mail (ST2) -, 
each of them performed in two scenarios: without verbal stimuli and with verbal
stimuli. The presence or not of verbal stimuli defines if the treatment induces
high or low cognitive workload respectively. For the third and fourth treatments
participants must perform five subtasks under the same scenarios: reading an e-
mail with search instructions (ST3), opening an application suitable to perform
the search instructions (ST4), entering the search parameters in the application
(ST5), reading and analyzing the search results (ST6), and replying the e-mail
with the results (ST7). The design of these treatments aims to manipulate the
cognitive workload through the cognitive interferences phenomenon described in
Wickens’s theories. Thus, a task will be more cognitively demanding as there
are more cognitive interferences in it.

At treatment one (T1) the participant must answer simple e-mails free of in-
terruptions. According to Wickens’s MRM she/he only needs to control her/his
personal space while performing the subtasks. Treatment two (T2) requires the
participant to perform ST1 and ST2, but at the same time to verbally answer
questions from the experimenter while listening to music (in her/his native lan-
guage) through a hearing aid. Thus, the participant should pay attention, exam-
ine, analyze, and make sense of the questions, while using auditory and visual
input modalities and working memory. At treatment three (T3) the participant
is free of interruptions, but she/he must read and respond an e-mail request-
ing to search for hotel prices, travel, cars and restaurant names using a specific
application. Finally, treatment four is similar to treatment three, but with the
Table 1. Cognitive Resources Demand and Interferences for each Treatment

<table>
<thead>
<tr>
<th>Processing Stage</th>
<th>Perceptual Encoding</th>
<th>Central Processing</th>
<th>Responsive Processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive Resource</td>
<td>Selective Attention</td>
<td>Perception</td>
<td>Working Memory</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Processing Code</td>
<td>Processing Code</td>
</tr>
<tr>
<td>Treatment Code</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>Auditory Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>T2</td>
<td>Visual Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>T3</td>
<td>Verbal Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>T4</td>
<td>Spatial High</td>
<td>Low</td>
<td>Low</td>
</tr>
</tbody>
</table>

cognitive workload induced by the experimenter questions and the music listening. One interference takes place in both, treatment one and three. This is caused by the simultaneous use of motor control resources and verbal code attributes. In contrast, at treatments two and four several partial interferences occur. This since both, the subtasks and the verbal stimulus require motor control and working memory resources simultaneously. We summarize in table 1 the level of cognitive interference associated to each treatment given the demanded cognitive resources, the input modality and processing code utilized for the task.

**Apparatus.** HR and SpO₂ were captured with a pulseoximeter that measure the pulse in beats per minute (bpm) and the blood oxygen saturation through the relation of the light irradiated from the blood with hemoglobin (Hb) and hemoglobin with oxygen (Hb + O₂). Both signals were captured at a rate of 50 Hz. The pulseoximeter is connected to an e-health shield from the Coocking Hacks company that sends data to a Linux server on the cloud.

The Subtasks were performed on a Samsung Galaxy S5 Smartphone. The interactions with the smartphone were video taped with an eye tracker in the shape of glasses. Recording the interaction is required to get the starting and ending timestamps of each subtask, which allows to measure duration of the subtasks in milliseconds. The eye tracker is connected to a Microsoft Surface Tablet that stores the video streams locally. Both the data in the cloud and the streams in the Tablet were latter synchronized using the Unix timestamp.

**Experimental Procedure and Task Validation.** The experiments were conducted individually through the following procedure: 1) The participant is asked to read and complete an informed consent (authorized by the ethical committee of the Faculty of Physical and Mathematical Sciences of University of Chile); 2) the participant is asked to close her/his eyes during 90 seconds in order to relax; 3) the experimenter randomizes the order in which the treatments will be performed and applies them; 4) after each treatment the participant must complete the NASA TLX test and is asked to close eyes for 30 seconds in order to relax again.

A pretest was conducted with five participants, who are not included in the experimental group, in order to validate each treatment design. A repeated mea-
Table 2. Repeated Measures ANOVA for Heart Rate and Oxygen Saturation

<table>
<thead>
<tr>
<th>Signal</th>
<th>Measure</th>
<th>ST1</th>
<th>ST2</th>
<th>ST3</th>
<th>ST4</th>
<th>ST5</th>
<th>ST6</th>
<th>ST7</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR</td>
<td>p-value</td>
<td>&lt;2e-16</td>
<td>&lt;2e-16</td>
<td>&lt;2e-16</td>
<td>&lt;2e-16</td>
<td>&lt;2e-16</td>
<td>&lt;2e-16</td>
<td>&lt;2e-16</td>
</tr>
<tr>
<td>SpO2</td>
<td>p-value</td>
<td>&lt;2e-16</td>
<td>&lt;2e-16</td>
<td>&lt;2e-16</td>
<td>&lt;2e-16</td>
<td>&lt;2e-16</td>
<td>&lt;2e-16</td>
<td>&lt;2e-16</td>
</tr>
</tbody>
</table>

Data Processing. Data records of each subtask were labeled with the cognitive workload (low or high) induced by each treatment (previously validated through the NASA TLX score). Each record consists of timestamp, subtask, HR value, SpO2 value and label. These records are divided in a set of one-second time window. For each window, the mean, median and standard deviation (SD) are computed as features. In this way, new records are created following the structure \{timestamp, subtask, mean_{HR}, mean_{SpO2}, median_{HR}, median_{SpO2}, SD_{HR}, SD_{SpO2}, label\}. By applying correlation and principal component analysis, it is determined that the mean is the best feature for both signals. Due to the high variance in the biological measures between humans, the measure within subjects is standardized by subtracting the mean from each value and dividing the difference by the SD.

4 Experimental Results

In average, the HR in treatments one and three is 76.90 bpm ± 3.7 bpm versus 82.70 bpm ± 4.8 bpm in treatments two and four. The values of SpO2 are 96.06 bpm ± 1.8 bpm and 98.51 bpm ± 1.2 bpm according to the same treatments. Clearly, both HR and SpO2 increase with the difficulty of the subtask.

Statistical Analysis. In order to answer the first research question it is necessary to apply a RM-ANOVA, since all the participants are exposed to the four treatments. The null hypothesis indicates that the mean for each group are equal. However, as shown in Table 2 it is possible to reject the null hypothesis both for HR and SpO2. Therefore there is evidence that the signals used in this paper allow to discriminate different levels of cognitive workload (RQ1).

Machine Learning. To perform the classification of cognitive workload level, a 10-fold cross validation is performed in which the classes are evenly distributed within each set and randomly selected in order to avoid bias. The independent variable is the cognitive level of the task. Table 3 shows the classification results.
Table 3. Classification Results using Support Vector Machine and $HR + SpO_2$

<table>
<thead>
<tr>
<th>ST</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>ST</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ST_1$</td>
<td>82.78%</td>
<td>81.80%</td>
<td>76.68%</td>
<td>$ST_4$</td>
<td>81.23%</td>
<td>83.53%</td>
<td>78.03%</td>
</tr>
<tr>
<td>$ST_2$</td>
<td>81.94%</td>
<td>88.12%</td>
<td>84.01%</td>
<td>$ST_6$</td>
<td>84.40%</td>
<td>88.68%</td>
<td>82.89%</td>
</tr>
<tr>
<td>$ST_3$</td>
<td>80.19%</td>
<td>72.85%</td>
<td>71.19%</td>
<td>$ST_7$</td>
<td>85.71%</td>
<td>83.98%</td>
<td>81.65%</td>
</tr>
</tbody>
</table>

Table 3 provides the classification results using Support Vector Machine and features of $HR$ and $SpO_2$ combined ($RQ_2$). It is noticeable that the classification models have measures of accuracy and performance in range within previous studies in the field. The combination of both signals has the best predictive power (greater than 80%). Separately, $HR$ dominates $SpO_2$, in terms of accuracy and precision, on each subtask. For instance, in $ST_1$, $HR$ and $SpO_2$ have an accuracy of 78.49% and 68.12% respectively. It is worth to remark that $ST_6$ and $ST_7$ have the best results, being $ST_7$ the subtask with the higher accuracy (85.71%).

5 Discussion and Conclusion

Psychophysiological signals are an objective mean to evaluate cognitive workload of Smartphone multitasking users. The current availability of non-invasive and inexpensive sensors, opens plenty of room to improve this approach. Particularly, in the area of Ubiquitous Computing, by providing a new type of information as an input to better understand user context.

This paper provides evidence to support that $HR$ and $SpO_2$ signals are capable of discriminating between tasks of different difficulty ($RQ_1$), and that it is possible to classify between two levels of cognitive difficulty by training machine learning models ($RQ_2$). Given that the treatments were designed based on real scenarios, these findings could be applied to everyday activities, outside controlled environments. In this way, this work contributes towards a continuous, non-invasive detection of the users cognitive workload while she/he interacts with Smartphones in their daily activities. This line of work has great potential to impact areas such as design and dynamic adaptation of mobile applications, identification and assessment of individual cognitive differences, determination of appropriate moments to deliver content or interrupt the user, among others.

The extension of this work runs through different lines of action. Firstly, to gather more relevant information by increasing the number of acquired signals. Secondly, to apply this methodology to other user contexts such as the Web. Finally, to improve the machine learning models.

Acknowledgments. This work was financed by Conicyt Fondecyt 11130252 and had the continuous support of “Instituto Sistemas Complejos de Ingeniería” (Conicyt: Proyecto Basal FBO16). Also, the authors would like to thank to Jorge Gaete-Villegas and all the experiment participants.
References


