My Watch Says I’m Busy: Inferring Cognitive Load with Low-Cost Wearables

Martin Gjoreski
Mitja Luštrek
Department of Intelligent Systems
Jozef Stefan Institute
Jozef Stefan International Postgraduate School
Ljubljana, Slovenia
martin.gjoreski@ijs.si

Veljko Pejović
Faculty of Computer and Information Science
University of Ljubljana, Slovenia
veljko.pejovic@fri.uni-lj.si

Abstract

To prevent undesirable effects of attention grabbing at times when a user is occupied with a difficult task, ubiquitous computing devices should be aware of the user’s cognitive load. However, inferring cognitive load is extremely challenging, especially when performed without obtrusive, expensive, and purpose-built equipment. In this study we examine the potential for inferring one’s cognitive load using merely cheap wearable sensing devices. We subject 25 volunteers to varying cognitive load using six different Primary tasks. In parallel, we collect physiological data with a cheap device, extract features, and then construct machine learning models for cognitive load prediction. As metrics for the load we use one subjective measure, the NASA Task Load Index (NASA-TLX), and two objective measures: task difficulty and reaction time. The leave-one-subject-out evaluation shows a significant influence of the task type and the chosen cognitive load metric on the prediction accuracy.

Author Keywords

Mobile sensing; Cognitive load inference; Wearable sensing

ACM Classification Keywords

H.5.2. [Information Interfaces and Presentation (e.g. HCI)]: User Interfaces; H.1.2. [Models and Principles]: User/Machine Systems

ACM.
UbiComp/ISWC’18 Adjunct,, October 8–12, 2018, Singapore, Singapore
ACM 978-1-4503-5966-5/18/10.
https://doi.org/10.1145/3267305.3274113

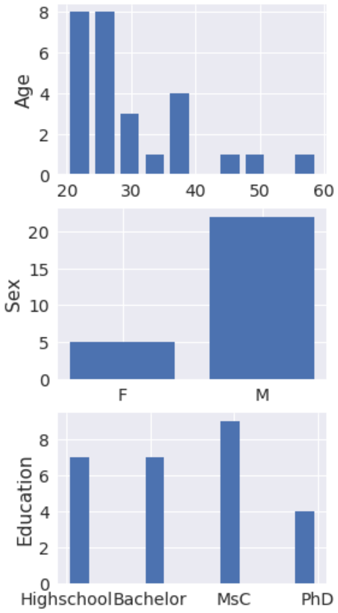


Figure 1: Demographic information histograms.

Introduction

We are interacting with computing devices in an ever expanding range of contexts: we work at our PCs, rely on smartphones for location-based services, we use in-car entertainment systems, manage home heating and lighting through a system of IoT devices, to name a few. However, as long as ubiquitous computing devices remain unaware of the user's internal state, this interaction is crude.

Human attention is one of the most precious resources, as evidenced by how mobile apps and online services compete for it. To prevent the above undesirable effects of grabbing attention at times when a user is occupied with a difficult task, ubiquitous computing devices should be aware of the current cognitive load the user is experiencing. While the human body does elicit a physiological response to (mentally) difficult situations, inferring cognitive load is nevertheless extremely challenging. To date, efforts have been geared towards studies in which sophisticated equipment was used to demonstrate the correlation between certain physiological signals, such as heart rate variability and skin conductance, with the cognitive load. For cognitive load inference to be of practical use in improving the way computing devices interact with their user, it has to be done via cheap unobtrusive means available to a large population.

Interruptibility and Cognitive Load Inference

Despite limited cognitive resources, the human brain is fascinating in its ability to multitask. Threaded cognition theory explains this ability in an analogy similar to the way multithreading is implemented in single-CPU/core computers – a cognitive resource is assigned exclusively to a single thread, while multiple threads may contend for the same resource [10]. Starting from these postulates and through experimental verification Borst et al. describe disruption to a primary task caused by an interrupting task – if a cognitively demanding task is deprived of sufficient resources, its problem state has to be stored in declarative memory and retrieved once the resources are again available [2]. The process of storage/retrieval may incur errors, may slow down task processing, and consequently may result in increased user frustration [3]. Thus, identifying and avoiding to interrupt at moments when cognitively demanding tasks are active is crucial for efficient attention management in ubiquitous computing.

The most common means of measuring cognitive load is NASA Task Load Index (NASA-TLX), a set of questions that, if administered immediately after the task, allow post hoc analysis of the cognitive load [6]. For an active attention management system, however, we require momentary estimation of a user's cognitive load. Physiological signals, such as heart rate variability, skin conductance, pupil dilation, and others have been shown to correlate with cognitive load [8, 9]. Several studies in particular have confirmed the link between physiological signals and office-based cognitive tasks of varying difficulty [14, 5]. Yet, these signals were measured with specialized equipment constraining the user to a lab setting, restricting a user's movement, and preventing general applicability of the system. Alternative means of inferring cognitive load that bypass physiological signal measurements and rely on smartphone interactions and mobile sensed data have not yielded results [13]. Therefore, in this work we concentrate on non-intrusive cognitive load inference using physiological data captured by cheap off-the-shelf wearable devices.

Data Collection

In order to collect physiological signals characteristic of situations where a user is more or less cognitively engaged, we conducted an experiment in which the participants were solving cognitive tasks of varying difficulty. The experiment

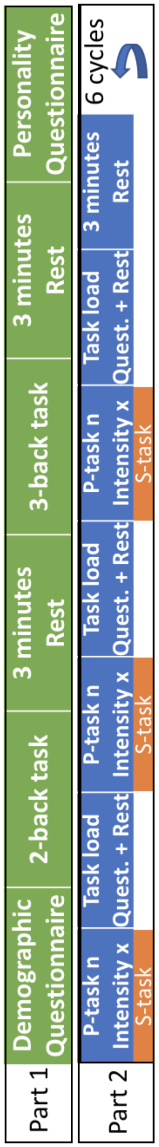


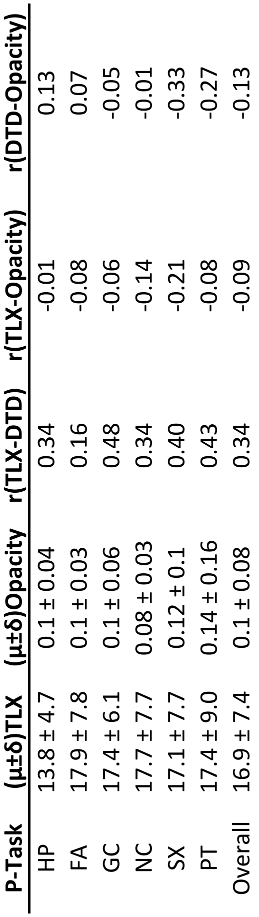
Figure 2: Experimental scenario.

was performed in a quiet, normal-temperature room with one participant at a time. At the beginning of each session, the participants were placed in a comfortable chair in front of a computer monitor and were presented with brief information regarding the experiments. Next, an inexpensive off-the-shelf device (Microsoft Band 2) was put on their left wrist and the rest of the experimental session was recorded in the same chair without any restrictions regarding the participants' hand gestures. Thus, the experimental setup simulates sedentary work on a computer in an office.

The experimental scenario (depicted in Figure 2) consists of Part 1 and Part 2. Part 1 was dedicated to assessing the participants' cognitive capacity and the personality type. For assessing the participants' cognitive capacity, the participants were solving two N-back tasks [12], i.e., 2-back and 3-back task, with a three-minute rest after each of them. In general, in N-back tasks, participants are presented a sequence of stimuli. For each stimulus, they need to decide if the current stimulus is the same as the one presented N trials ago. In our case, the stimuli were fields in a 3x3 grid, one of which was colored at each time step. The participants needed to decide whether the colored field was the same as the one colored n-steps ago. The ratio of correct/incorrect answers provides information about the cognitive capacity of each participant. For assessing the personality type, the participants filled a Hexaco Personality questionnaire [1], which provides information about the participants': Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience. Also, the participants filled a questionnaire with demographic information, the summary of which is presented in Figure 1.

In Part 2 of the experiment, the participants were presented with six cycles of cognitive load tasks. For each cycle, three variations of a randomly selected Primary cognitive-load task (P-task n, where n is in $[1 : 6]$) were presented to the participant. The variations differed in difficulty (designed difficulty as Easy, Medium and Difficult) and thus in the expected cognitive load. After each of the three variations, the participants filled the NASA-TLX questionnaire to assess *subjective* cognitive load posed by the tasks. Also, in parallel with the Primary tasks, an additional Secondary task (S-task) was presented to the participant in order to fill-in the participant's free cognitive resources. The S-task contained a square starting as completely transparent in a random corner of the screen, and then increasing in opacity. The participants' goal was to react, i.e., to click on the appearing square as soon as they notice it. The opacity of the square when clicked was intended to be related to the participant's engagement in the P-task, since more engaged uses were expected to notice the square later, when it is darker [4]. An assumption is that increased engagement corresponds to higher cognitive load put towards the P-task.

For the P-tasks we used the tasks and software presented by [5] in their study of psycho-physiological measures for assessing cognitive load. The software displays the following six primary tasks: Gestalt Completion (GP) test - where the subject is asked to identify incomplete drawings; Hidden Pattern (HP) test - where the subject has to decide whether a model image is hidden other comparison images; Finding A's (FA) test - where the subject has to find the letter 'a' in presented words; Number Comparison (NC) test - Where the subject has to decide whether or not two displayed numbers are the same; Pursuit test (PT) - where the subject has to visually track irregularly curved overlapping lines from numbers on the left side of a rectangle to letters on the opposite side; and Scattered X's (SX) test - Where the subject has to find the letter 'x' on screens containing random letters. Besides solving the P-task, participants



P-Task	(μ±δ)TLX	(μ±δ)Opacity	r(TLX-DTD)	r(TLX-Opacity)	r(DTD-Opacity)
HP	13.8 ± 4.7	0.1 ± 0.04	0.34	-0.01	0.13
FA	17.9 ± 7.8	0.1 ± 0.03	0.16	-0.08	0.07
GC	17.4 ± 6.1	0.1 ± 0.06	0.48	-0.06	-0.05
NC	17.7 ± 7.7	0.08 ± 0.03	0.34	-0.14	-0.01
SX	17.1 ± 7.7	0.12 ± 0.1	0.40	-0.21	-0.33
PT	17.4 ± 9.0	0.14 ± 0.16	0.43	-0.08	-0.27
Overall	16.9 ± 7.4	0.1 ± 0.08	0.34	-0.09	-0.13

Figure 3: Experimental data.

were asked to react as quickly as possible to occasionally appearing S-tasks.

From the wrist device, we recorded the following type of data: R-R (or inter-beat) intervals, Galvanic Skin Response (GSR), Skin Temperature (ST), barometer data, accelerometer data and data from UV sensor. In this paper, we focus only on the data from the physiological sensors, i.e., R-R, GSR and ST data. The data form the wrist-device was transmitted via Bluetooth and a mobile phone to a server, and the analysis was performed off-line.

Analysis and Results

Each participant filled three questionnaires at each of the six cycles, thus there were 18 NASA-TLX questionnaires per participant. For some questionnaires the sensor data was missing due to technical issues. For the ML analysis, we used only the data from the participants that had sensor data for more than 10 questionnaires, which translated to data from 21 participants. For the analysis, we took 90-seconds segments before each NASA-TLX questionnaire, which was determined experimentally. Figure 3 presents descriptive statistics for the experimental data after the split. The first column represents the name of the Primary task and the second column the number of segments per task. The third column represents the median and the standard deviation for the task load index (TLX). The fourth column represents the median and the standard deviation of the Opacity values measured via the S-task. The final three columns represent the Pearson's correlation coefficient between: TLX and designed task difficulty (i.e. DTD – easy, medium, or difficult); TLX and Opacity; DTD and Opacity. From the figure it follows that TLX varies for different tasks. For example, for the first P-task (HP) the median TLX is 13, whereas for the third P-task (GC) the median TLX is 17. Similarly, the Opacity values vary from 0.1 to 0.14 for different P-tasks. From the coefficients presented in the last three columns we observe: 1) no correlation between the Opacity and the TLX; 2) a low negative correlation between the Opacity and the TLX for two types of P-tasks (SX and PT) – the higher the intensity is, the quicker the participants react to the S-task, which is the opposite of what we expected and; 3) a low correlation between the TLX and the DTD for all P-tasks except for one, the FA task. Thus, the higher the designed difficulty of the P-task is, the higher the participant-reported TLX is.

Data from each segment are filtered and features are extracted using the following steps: the GSR signal is first filtered using a sliding mean filter. Next, the fast-acting component (GSR responses) and the slow acting component (tonic component) are extracted from the filtered GSR signals. The preprocessed GSR signal is used to calculate GSR features: mean, standard deviation, 1st and 3rd quartile, quartile deviation, derivative of the signal, sum of the signal, number of responses in the signal, responses per minute in the signal, sum of the responses, sum of positive derivative, proportion of positive derivative, derivative of the tonic component of the signal, difference between the tonic component and the overall signal. In addition, we calculate the total spectral power of the GSR signal in five frequency bands between 0 Hz and 0.5 Hz with a 0.1 Hz span.

The R-R signal is filtered by removing the R-R intervals that are outside of the interval [0.7*median, 1.3*median]. From the filtered R-R signals, a spectral representation is calculated using the Lomb-Scargle algorithm.

Next, the following HRV features were calculated: the mean heart rate, the mean of the R-R intervals, the standard deviation of the R-R intervals, the standard deviation of the differences between adjacent R-R intervals, the square root of the mean of the squares of the successive differences be-

Target	μ	Best	Best model	Accuracy increase relative to Majority						
	Majority	model	μ Accuracy	HP	FA	GC	NC	SX	PT	μ
TLX	40%	RF	47%	6%	-5%	5%	6%	21%	10%	7%
DTD	33%	NB	51%	27%	11%	10%	22%	14%	24%	18%
Opacity	36%	GB	46%	16%	5%	13%	6%	3%	20%	10%

Figure 4: Leave-one-subject-out evaluation results.

tween adjacent R-R intervals, the percentage of the differences between adjacent R-R intervals that are greater than 20 ms, the percentage of the differences between adjacent R-R intervals that are greater than 50 ms, Poincare plot indicies, the total spectral power of all R-R samples between 0.003 and 0.04 Hz (low frequencies) and between 0.15 and 0.4 Hz (high frequencies), and the ratio of low to high frequency power. Finally, from the ST signal we extract: mean, standard deviation, quartile deviation and derivative of the signal.

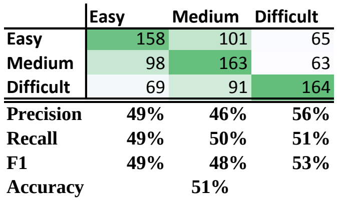
After the feature extraction, the segments are represented by features ready to be fed into machine learning (ML) algorithms. We experiment with the following seven ML algorithms: Random Forest (RF), Support Vector Machine, Gradient Boosting Classifier (GB), AdaBoost Classifier (with a Decision Tree as the base classifier), KNN Classifier, Gaussian Naive Bayes (NB) and Decision Tree Classifier.

Since the statistical analysis presented in Figure 3 discovered that the type of the Primary task significantly influences data, we experiment with task-specific models for each of the three cognitive load measures: NASA-TLX, Designed Task Difficulty (DTD) and the Opacity measured from the Secondary task. In order to have comparable experimental results, the three cognitive load measures are represented as a three-class problem. For the DTD, the classes are Easy-Medium-Difficult, as inherited from the designed difficulty of the Primary task itself. For the TLX and for the Opacity, there were three continuous values per task per person. From the three values, the segment with lowest TLX or Opacity value was labeled as Easy, the segment with highest value was labeled as Difficult, and the segment with the value in between was labeled as Medium. If the values for two segments were equal, both segments were labeled with the same label (Easy or Difficult, depending whether the third value was higher or lower). Thus, there is a slight difference in the majority class between the three targets (TLX, DTD and Opacity).

All models are evaluated using leave-one-subject-out cross validation. Figure 4 summarizes the results. The first column represents the target for the ML models. The second column represents the majority. The third column represents the best model out of the seven tested. The fourth column presents the mean accuracy achieved by the best model. The next six columns represent the per-task accuracy relative to the task majority. Finally, the last column represents the mean accuracy relative to the target majority. For five out of the six Primary tasks, the ML model that predicts the DTD achieve better accuracy compared to the models that predict the TLX or Opacity. Only for one Primary task (the SX), the TLX model is better. The highest mean accuracy of 51% is achieved by the NB, which is an 18 percent improvement compared to the majority. In addition, depending on the type of the P-task, the accuracy can vary for up to 17 percents. For example, the minimum accuracy achieved by the NB is 43% (for the task GC) and the maximum is 60% (for the task HP). Figure 5 presents the confusion matrix for the best performing task-specific model build with the NB algorithm. The precision, recall and the accuracy are near 50%. While these figures are not very high, it is encouraging that the model tends to confuse the neighboring labels, i.e., Easy-Medium and Medium-Difficult more than the distant labels (i.e. Easy-Difficult).

Conclusions

25 participants were were subjected to varying cognitive load and physiological data was collected with an inexpensive physiological sensing wristband. Features were extracted and then used to infer the cognitive load using ML. One subjective measure - the NASA-TLX, and two objective



	Easy	Medium	Difficult
Easy	158	101	65
Medium	98	163	63
Difficult	69	91	164
Precision	49%	46%	56%
Recall	49%	50%	51%
F1	49%	48%	53%
Accuracy		51%	

Figure 5: Confusion Matrix for the best performing model for predicting Task Intensity.

measures: designed task difficulty and reaction time on a secondary task, were used to label the cognitive load level.

The statistical analysis discovered that for two Primary task types (SX and PT) the higher the design difficulty is, the quicker the participants react to the Secondary task. This may indicate that for some tasks, when there is a higher intensity, the participants may be more focused on the screen, thus they can react faster to the Secondary task. This seems counter-intuitive, since one would expect the higher the difficulty of the Primary task is, the slower the reaction would be on the Secondary task [4]. We plan to study this relation in future work.

Using task-specific ML models, we compared the accuracy of seven ML algorithms for predicting the three different cognitive load measures. The results showed that the ML models are most accurate when predicting the designed task difficulty (DTD), second is the Opacity measure, and the models perform poorly when predicting the NASA-TLX. This may mean that the ML models can better capture objective measures (DTD and Opacity) compared to the subjective one. However, this question needs to be studied with more participants in order for the results to be fully reliable.

The ML experiments also confirmed that the type of the Primary task significantly influences the experimental results, i.e., the accuracy of the ML models significantly varies depending on the type of the Primary task. For example NB achieved a minimum of 43% and a maximum of 60% accuracy, depending on the type of the Primary task.

The work presented here is related to the study presented by Novak et al. [7], where they performed statistical analysis on physiological data, and the recent study presented by Schaule et al. [11], where they presented a system for interruption management. Both of these studies are using data from Microsoft Band. In contrast to their work, we evaluated ML models for inferring three different cognitive load measures for six different cognitive load tasks, which provides deeper insight into the dynamics of the cognitive load.

Acknowledgement

We thank SeungJun Kim, Anind K. Dey, Eija Haapalainen and Jodi F. Forlizzi for sharing their software for displaying the Primary cognitive tasks with us. We would also like to thank Martin Frlin for his help with programming the environment and conducting user studies.

REFERENCES

1. M.C. Ashton, K. Lee, M Perugini, P Szarota, R.E. de Vries, L Di Blas, K Boies, and B. De Raad. 2004. A six-factor structure of personality-descriptive adjectives: Solutions from psycholexical studies in seven languages: Solutions from psycholexical studies in seven languages. *Journal of Personality and Social Psychology* 86, 2 (2004), 356 – 366. DOI: http://dx.doi.org/10.1037/0022-3514.86.2.356 Article.

2. Jelmer P. Borst, Niels A. Taatgen, and Hedderik van Rijn. 2010. The Problem State: A Cognitive Bottleneck in Multitasking. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 36, 2 (2010), 363–382.

3. Jelmer P. Borst, Niels A. Taatgen, and Hedderik van Rijn. 2015. What Makes Interruptions Disruptive?: A Process-Model Account of the Effects of the Problem State Bottleneck on Task Interruption and Resumption. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 2971–2980. DOI: http://dx.doi.org/10.1145/2702123.2702156

4. Krista E DeLeeuw and Richard E Mayer. 2008. A Comparison of Three Measures of Cognitive Load: Evidence for Separable Measures of Intrinsic, Extraneous, and Germane load. *Journal of Educational Psychology* 100, 1 (2008), 223–234.

5. Eija Haapalainen, SeungJun Kim, Jodi F. Forlizzi, and Anind K. Dey. 2010. Psycho-physiological Measures for Assessing Cognitive Load. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing (UbiComp '10)*. ACM, New York, NY, USA, 301–310. DOI: http://dx.doi.org/10.1145/1864349.1864395

6. Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology* 52 (1988), 139–183. DOI: http://dx.doi.org/https://doi.org/10.1016/S0166-4115(08)62386-9

7. G. Jakus K. Novak, K. Stojmenova and J. Sodnik. 2017. Assessment of cognitive load through biometric monitoring. Society for Information Systems and Computer Networks, Ljubljana, Slovenia.

8. Fred Paas, Juhani E Tuovinen, Huib Tabbers, and Pascal WM Van Gerven. 2003. Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist* 38, 1 (February 2003), 63–71.

9. Rahul Rajan, Ted Selker, and Ian Lane. 2016. Task Load Estimation and Mediation Using Psycho-physiological Measures. In *Proceedings of the 21st International Conference on Intelligent User Interfaces (IUI '16)*. ACM, New York, NY, USA, 48–59. DOI:http://dx.doi.org/10.1145/2856767.2856769

10. Dario D. Salvucci and Niels A. Taatgen. 2008. Threaded cognition: An integrated theory of concurrent multitasking. *Psychological Review* 115, 1 (January 2008), 101–130.

11. Florian Schaule, Jan Ole Johanssen, Bernd Bruegge, and Vivian Loftness. 2018. Employing Consumer Wearables to Detect Office Workers' Cognitive Load for Interruption Management. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 32 (March 2018), 20 pages. DOI: http://dx.doi.org/10.1145/3191764

12. Florian Schmiedek, Martin Lövdén, and Ulman Lindenberger. 2014. A task is a task is a task: putting complex span, n-back, and other working memory indicators in psychometric context. *Frontiers in psychology* 5 (2014).

13. Gašper Urh and Veljko Pejovic. 2016. TaskyApp: Inferring Task Engagement via Smartphone Sensing. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct (UbiComp '16)*. ACM, New York, NY, USA, 1548–1553. DOI: http://dx.doi.org/10.1145/2968219.2968547

14. Manuela Züger and Thomas Fritz. 2015. Interruptibility of Software Developers and Its Prediction Using Psycho-Physiological Sensors. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 2981–2990. DOI: http://dx.doi.org/10.1145/2702123.2702593