Dynamic Workload Adjustments in Human-Machine Systems Based on GSR Features

Jianlong Zhou, Ju Jung, Fang Chen

Dynamic Workload Adjustments in Human-Machine Systems Based on GSR Features

Jianlong Zhou[1], Ju Young Jung[2], and Fang Chen[3]

[1,2,3] National ICT Australia (NICTA), Eveleigh NSW 2015, Australia
[2] The University of Sydney, NSW 2006, Australia
[1,3] {FirstName.LastName}@nicta.com.au
[2] jjun3170@uni.sydney.edu.au

Abstract. Workload is found to be a critical factor driving human behavior in human-machine interactions in modern complex high-risk domains. This paper presents a dynamic workload adjustment feedback loop with a dynamic cognitive load (CL) adaptation model to control workload adjustment during human-machine interaction. In this model, physiological signals such as Galvanic Skin Response (GSR) are employed to obtain passive human sensing data. By analyzing the obtained sensing data in real-time, the task difficulty levels are adaptively adjusted to better fit the user during working time. The experimental results showed that SVM outperformed other methods in offline CL classifications, while Naïve Bayes outperformed other methods in online CL level classifications. The CL adaptation model 1 (average performance is 87.5%) outperformed the adaptation model 2 during the dynamic workload adjustment.

Keywords: Cognitive load, GSR, Dynamic adjustment, Machine learning.

1 Introduction

Cognitive load (CL, also known as workload) refers to the amount of mental demand imposed on a user by a particular task, and has been associated with the limited capacity of working memory and the ability to process novel information [1, 2]. The cognitive load experienced by a user in completing a task has a major impact on her/his ability to acquire information during the task, and can severely influence the overall productivity and performance. High levels of cognitive load are known to decrease effectiveness and performance of users, as well as their ability to learn from their tasks [1]. On the other hand, if a task is very easy and routine, only inducing a low level of cognitive load, it may cause boredom and loss of focus, ultimately resulting in lower performance. In this way, the concept of an optimal range of cognitive load levels is developed, outside of which a subject's ability to learn, perform, and complete a task is likely to be negatively affected [2]. It is crucial to maintain the cognitive load experienced by a user within this optimal range to achieve the highest productivity. Cognitive load measurement (CLM) therefore plays an important role in applications involving human-machine interface. Furthermore, Galvanic Skin Response (GSR) is a measure of conductivity of human skin, and can provide an indica-

tion of changes in human sympathetic nervous system [3]. GSR has attracted researchers' attention as a prospective physiological indicator of cognitive load [3].

In order to keep the user in an optimal state and improve user's engagement and performance, dynamic workload adjustment (DWA) systems automatically modulate the difficulty of tasks and other factors related to tasks in human-machine systems in real-time. By monitoring user's state and adapting the task difficulty levels, a dynamic workload system improves user's performance and helps users maximize their capacity for productive work. However, there are still challenges in utilizing CLM in dynamic workload adjustment. For example, there is no effective model allowing the system to effectively adjust task elements dynamically.

This paper presents a dynamic workload adaptation model to control how workload can be adjusted during human-machine interaction. In this model, physiological modalities such as GSR are employed to obtain passive human sensing data. By analyzing the obtained sensing data in real-time, we adapt task difficulty levels in order to optimize workload in real-time, which allows the system to better fit the task to the user during working time. The dynamic workload adaptation feedback loop helps to balance the task performance and workload levels.

2 Related Work

Shi et al. [4] evaluated users' stress and arousal levels with GSR. The results showed that mean GSR significantly increases when task cognitive load level increases. Moreover, users' GSR readings are found to be lower when using a multimodal interface, instead of a unimodal interface. Son and Park [5] estimated driver's cognitive load using driving performance and skin conductance level as well as other measures in a driving simulator. The results showed that the skin conductance level provides clear changes associated with difficult level of cognitive workload. Nourbakhsh et al. [3] also indexed cognitive load with GSR features of accumulative GSR and GSR power spectrum in arithmetic tasks. Wang et al. [6] indexed cognitive load with GSR features such as mean-difference. GSR data were also tested by the Boosting algorithm with Haar-like features for cognitive load classifications. Furthermore, Afergan et al. [7] used functional near-infrared spectroscopy (fNIRS) to detect task difficulty and optimize workload with a dynamic adaptation. However, few researches use GSR features to index cognitive load dynamically in an adaptive feedback loop. This paper investigates the use of GSR in a dynamic workload adaptation feedback loop in order to improve task performance.

3 Experiment

3.1 Dynamic Workload Adaptation Feedback Loop

We propose a dynamic workload adaptation feedback loop as show in Fig. 1. In this feedback loop, physiological signals such as GSR are recorded when the user is performing a task. The recorded signals are then analyzed and classified as different CL levels. The classified CL levels are input into the adaptation model in order to modu-

late task elements. A new task session is then performed based on the adaptation in order to keep task difficulty on an optimal level for the participants.

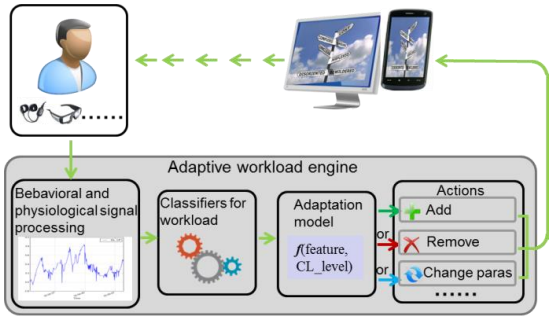

Fig. 1. Diagram of feedback loop of dynamic workload adaptation.

3.2 Task Design

Table 1. Cognitive load level definitions.

CL	*Binary*	*1-Digit*	*2-Digit*	*3-Digit*
0	4	0	0	0
1	3	1	0	0
2	2	2	0	0
3	1	3	0	0
4	0	4	0	0
5	0	3	1	0
6	0	2	2	0
7	0	1	3	0
8	0	0	4	0
9	0	0	3	1
10	0	0	2	2
11	0	0	1	3
12	0	0	0	4

Arithmetic addition task is used in this study. Each task is designed to stimulate a particular CL level from the participant based on the difficulty of the task. An "X" is shown at the beginning on a computer LCD display followed by four numbers in succession, where each number and "X" are displayed for 3 seconds. The participant is required to add these numbers up in his/her head during the task, and must choose one answer from the four options. At the completion of each task, the collected data are analyzed by the system. A new task is followed with a particular CL level controlled by the adaptation model. CL levels are designed as in Table 1 based on [3], where the number in each of the columns represents how many of the particular category of number (binary, 1-digit, 2-digit, 3-digit) are shown in the task.

3.3 Procedures

The experiment is carried out with the following procedures: 1) A computer is firstly setup with the GSR sensor connected and the corresponding drivers installed. 2) The

participant is seated facing the LCD display of the computer. The tips of the index and fourth fingers of the left hand of the participant are connected to the GSR sensor, and the right hand of the participant is used to navigate the mouse to engage with the program. 3) The experiment begins by launching the experiment program. 4) The training stage firstly runs for a total of 8 tasks. 5) After training is completed, 6 testing stages are run, and each testing stage has a total of 20 tasks. Table 2 illustrates the different testing stage scenarios. The numbers displayed for the arithmetic tasks are all randomized, and the difficulties of tasks are controlled by the adaptation model. GSR devices from ProComp Infiniti GSR of Thought Technology Ltd were used in the experiment. 10 participants of university students and research staff were recruited in this experiment.

Table 2. Testing stage scenarios.

Testing Stage	*Initial CL*	*Adaptation Model*
1	0	1
2	6	1
3	12	1
4	0	2
5	6	2
6	12	2

The scenarios of testing stages are as follows:

- The testing stages starting from an initial CL level of 0 and 12 are designed to test how effective the adaptation model is in allowing the difficulty level to shift from one extreme to a more desirable state.
- The testing stages starting from an initial CL level of 6 is designed to test the robustness of the adaptation model, and observe its effectiveness in keeping the CL level stable at the desirable state.
- Two different adaptation models are used for comparison resulting in 6 experiment scenarios.

4 Cognitive Load Classification

4.1 Signal Processing

The raw GSR signals are firstly calibrated in order to account for variations of GSR between individuals and time intervals. As mentioned in Section 3.2, an "X" is shown at the beginning before the actual arithmetic task begins. This period is used as the reference point on which the rest of the data can be calibrated. The calibration is achieved using the relationship: $G_T = \frac{G_t - G_X}{G_X}$, where G_X is the average GSR value during the X-displaying period, and G_t and G_T are the raw and calibrated GSR signals during the task time respectively. Signal smoothing is achieved using a Hann window function as a low pass filter to remove high frequency noise.

4.2 Feature Extraction

Time domain features and frequency domain features are extracted in this study. For the time domain features, we focus on analyzing the nature of the major peaks in the data. In order to normalize the magnitudes of these peaks across all participants, the processed GSR signal for each task is divided by the mean value of all tasks of the particular participant:

$$G_N(i,k,t) = \frac{G_T(i,k,t)}{\frac{1}{m}\sum_{j=1}^{m}\sum_{t=1}^{T_{ij}} G_T(i,j,t)},$$

where $G_T(i, k, t)$ is the result from signal processing and $G_N(i,k,t)$ is the normalised GSR value at time t of task k of subject i.

The significance of each peak i is quantified using the duration of the peak S_{d_i}, magnitude S_{m_i}, and area $S_{a_i} = S_{d_i}S_{m_i}$. The time domain features include: 1) Sum of peak durations $S_d = \sum S_{d_i}$; 2) Sum of peak magnitudes $S_m = \sum S_{m_i}$; 3) Sum of peak areas $S_a = \sum S_{a_i}$; 4) Number of peaks S_f; and 5) Time taken to choose answer T_c.

For frequency-domain feature extraction, Z-score normalisation is firstly applied:

$$G_{Z(i,j)} = \frac{G_T(i,j) - \mu_{G_T(i,j)}}{\sigma_{G_T(i,j)}},$$

where G_T is the calibrated and smoothened GSR signal and G_Z is the normalised signal of task j of subject i. μ and σ are the mean and standard deviation of $G_T(i,j)$. Since each task is normalised in this way using its own mean and standard deviation, the magnitudes and range of G_Z of each task become standardised, thus the frequency features can be isolated more effectively. The power spectrum is extracted using:

$$P(\omega) = \frac{1}{N} Y(\omega) Y^*(\omega),$$

where P is the power spectrum, ω is frequency, N is the length of the signal, and Y and Y^* are the frequency spectrum and its complex conjugate respectively. The average power below 1Hz was calculated for each task, as this frequency region contained the most non-zero values.

4.3 Offline Cognitive Load Classifications

Using all of the features mentioned above, three different machine learning (ML) methods were used to classify CL levels: SVM, Naïve Bayes, and Random Forest. Data was collected from 12 participants who performed an offline variation of the experiment with no real-time adaptation, using the same task design. For each participant, equal number of tasks were performed for CL levels {0,4,8,12}. Leave-one-out cross-validation across the participants was used to evaluate ML performance. For the collected offline GSR signals, SVM slightly outperformed the other two with the accuracy of 78.1% compared to 71.9% for Naïve Bayes and 76.0% for Random Forest for the 2-class classification. For the 4-class classification, SVM also outperformed the other two ML methods (49.0%), compared to 40.6% and 36.5% for Naïve Bayes and Random Forest respectively. Therefore, SVM was chosen as the ML method during the dynamic workload adaptation.

4.4 Online Cognitive Load Classifications

The online cognitive load classification during the dynamic workload required a slightly different cross-validation method compared to the offline classifications. In the online cognitive load classifications, the ML model was trained using a calibrated and therefore *personalized* version of the same static data used in the offline cross-validation. After extracting the features from the data, they were calibrated using the mean and variance of the features extracted from a short training stage conducted by the particular subject. In this way, the participant did not have to run an extremely long training stage, and the classification model would still take into account the subjective differences between participants to a certain extent.

Correctness and therefore *accuracy* of the classifications also need to be more clearly defined for online classifications as the CL levels are all integer values ranging from 0-12 for the tasks, but classifications are only made from the set of {0,4,8,12}. *Correctness* in the online 4-class classification is defined as follows: for any CL level that falls between two of {0,4,8,12}, a classification is considered *correct* if it matches one of these two *neighboring* values. For example, if the CL level is 3, a classification of either 0 or 4 is considered *correct*. For a 2-class problem, it is difficult to use a similar logic, thus a *correctly* classified value was defined to be within 6 levels of the true CL level. The *correctness* of classification for the online analysis was purposely defined more loosely due to issues that arise with stricter definitions given the nature of the task design. As a consequence of the inherent randomness presented in the tasks, the CL level can only be used as a good indicator of the difficulty rather than an exact metric. This is especially more relevant for the online classification problem involving 13 different CL levels as the margins between each level are not as distinguished. For the offline case which only involved levels {0,4,8,12}, this issue is minimized as the margins between levels are less disputable.

Table 3 shows the comparison of classification accuracy between offline classifications and online classifications. The results show that all ML methods have similar or even better performance in online CL classifications as in offline CL classifications. All three ML methods can effectively classify CL levels of tasks in real-time.

Table 3. Accuracy of offline classifications vs. online classifications.

Algorithm	***2-Class (%)***		***4-Class (%)***	
	offline	***online***	***offline***	***online***
SVM	78.1	81.7	49.0	44.0
Naïve Bayes	71.9	89.2	40.6	63.6
Random Forest	76.0	88.4	36.5	56.6

5 Dynamic Workload Adjustment

5.1 Adaptation Models

The objective of the adaptation model is to keep the CL level within the range 4-8 (middle range of CL levels) during the dynamic workload adaptation. Two adaptation models are designed in this study.

- Adaptation model 1:
  - If classified CL level CL_t at time t is to be 8 or 12, decrease the CL level by 1;
  - If CL_t is classified to be 0 or 4, increase CL level by 1.
- Adaptation model 2:
  - If CL_t is classified to be 12, or CL_t and CL_{t-1} are both classified to be 8 or above, decrease CL level by 1;
  - If classified level is 0, or CL_t and CL_{t-1} are both classified to be 4 or below, increase CL level by 1.

Adaptation model 1 can be regarded as the more *dynamic* variation model, while the adaptation model 2 is clearly more *stable* as it has to meet slightly more strict criteria to change levels. The design of both adaptation models is also kept consistent with their objective by ensuring that their behaviors are symmetrical about the mean value of the desired range, i.e. the level 6.

5.2 Performance Evaluation of Adaptation Models

Performance of the adaptation model is defined as the percentage of tasks which have $4 \leq CL \leq 8$ out of all tasks performed during a testing stage. It is regarded as the *desirable* range in the dynamic workload adaptation. Fig. 2 shows the changes of average CL levels during dynamic adaptation process. Each point (n, CL) on a curve represents the average CL level for the corresponding task number (n) of a particular testing stage scenario, differentiated by their initial CL and the adaptation model used. In Fig. 2, it is clear to see that all adaptation models were able to successfully drive and maintain the CL level within the desired range, and only minor differences are observed between the two adaptation models. For a quantitative comparison, the mean performances for each scenario are summarized in Table 4.

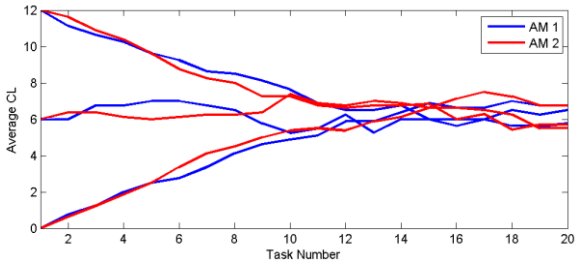

Fig. 2. CL Changes throughout testing stages.

Table 4. Average performance of adaptation models.

Initial CL	*AM1 (%)*	*AM2 (%)*
0	53.1	56.9
6	87.5	86.3
12	41.9	51.5

Significant differences are only observed for $CL_{initial} = 12$, where adaptation model 2 achieved around 10% higher performance than adaptation model 1. Reasons

for poorer and more disparate performance levels for this scenario could be attributed to the subjectivity inherent in the experiment that could not be completely removed from the task design. Subjectivity is a more significant issue for higher difficulty level tasks, as the GSR responses are likely to show greater variations between participants due to differences in arithmetic ability. With this increased diversity, accurate classification becomes more challenging, and translates to poorer performance.

6 Conclusions and Future Work

This paper investigated the use of GSR features in dynamic workload adjustment. Both time domain and frequency domain features were extracted and used for CL level classifications. The experimental results showed that SVM, Naïve Bayes, and Random Forest were all able to provide reasonable accuracies of CL level classifications. The classification results could be used as inputs to CL adaptation models for a dynamic workload adjustment environment, where the CL level is driven and maintained around an optimal level. Future work will focus on designing more complex adaptation models as well as the generalization of the dynamic workload adjustment on a wide variety of real working environments.

7 Acknowledgements

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. The authors thank Lucas Mattos for help in setting up the experiment.

8 References

1. Chandler, P., & Sweller, J. (1991). Cognitive Load Theory and the Format of Instruction. *Cognition and instruction*, *8*(4), 293–332.
2. Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. M. (2003). Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. *Educational Psychologist*, *38*(1), 63–71.
3. Nourbakhsh, N., Wang, Y., & Chen, F. (2013). GSR and Blink Features for Cognitive Load Classification. In P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, & M. Winckler (Eds.), *Human-Computer Interaction – INTERACT 2013* (pp. 159–166). Springer Berlin.
4. Shi, Y., Ruiz, N., Taib, R., Choi, E., & Chen, F. (2007). Galvanic Skin Response (GSR) As an Index of Cognitive Load. In *CHI '07 Extended Abstracts on Human Factors in Computing Systems* (pp. 2651–2656). New York, NY, USA: ACM. doi:10.1145/1240866.1241057
5. Son, J., & Park, M. (2011). Estimating Cognitive Load Complexity Using Performance and Physiological Data in a Driving Simulator. In *Proceedings of AutomotiveUI'11*. Austria.
6. Wang, W., Li, Z., Wang, Y., & Chen, F. (2013). Indexing Cognitive Workload Based on Pupillary Response Under Luminance and Emotional Changes. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces* (pp. 247–256).
7. Afergan, D., Peck, E. M., Solovey, E. T., Jenkins, A., Hincks, S. W., Brown, E. T., … Jacob, R. J. K. (2014). Dynamic Difficulty Using Brain Metrics of Workload. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3797–3806). New York, NY, USA: ACM.