In: Proceedings of the 49th Annual Meeting of the Human Factors and Ergonomics Society

NON-INTRUSIVE MEASUREMENT OF WORKLOAD IN REAL-TIME

Markus Guhe*, Wenhui Liao**, Zhiwei Zhu**, Qiang Ji**, Wayne D. Gray* & Michael J. Schoelles*
*Cognitive Science Department, **Department of Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute, Troy, NY

We present a new method to measure workload that offers several advantages. First, it uses non-intrusive means: cameras and a mouse. Second, the workload is measured in real-time. Third, the setup is comparably cheap: the cameras and sensors are off-the-shelf components. Fourth, we go beyond measuring performance and demonstrate that just using such measures does not suffice to measure workload. Fifth, by using a Bayesian Network to assess the workload from the various manifesting measures the model adapts itself to the individual user as well as to a particular task. Sixth, we use a cognitive computational model to explain the cognitive mechanisms that cause the differences in workload and performance.

INTRODUCTION

Workload is a comprehensive concept with many aspects. Simple definitions of workload include the demand placed upon humans (De Waard 1996) or the portion of the operator's limited capacity required to perform a particular task (O'Donnell & Eggemeier 1986).

Workload measurement is important for a number of problems, such as aviation or traffic. Various measurements for quantifying workload were proposed. These measures can be classified into three categories: *subjective measurement*, such as self-report, e.g. the NASA-TLX, *performance measurement*, where the operator's performance in the tasks is evaluated, e.g. Multiple Resource Theory (Wickens 1992), and *physiological measurement* like Galvanic Skin Response. However, these approaches suffer from limitations in predictive power and general applicability (Wierwille, Rahimi & Casali 1985).

We will distinguish an "external" from an "internal" notion of workload in this paper. The external workload, which we call *task load*, is defined by the task environment. It is the number of problems that have to be processed in a given amount of time. We assume that it correlates in a principled fashion with the users' internal, *cognitive workload*, which is the number of mental operations that have to be performed to process these problems.

We develop a real-time, non-intrusive system that measures and detects the differences in workload. We focus on two aspects. First, we study the relations between task load and variance measures (outcome, response time, physical response, physiological response). Second, we apply a *Bayesian Network* (BN) model – a framework for reasoning under uncertainty – to predict workload from the various manifesting measures. The model adapts itself to a particular task as well as to the individual user. Individual differences are an important aspect of our research, because individuals express differences in workload differently, which is analogous (and most likely related) to the differences in corresponding affective states, such as stress (Picard 1997). Understanding these differences is necessary to understand and predict a user's behavior, e.g. for an estimation of how likely he or she will make an error. For this reason we currently focus on only a three participants, before we validate the system with a large number of participants.

TASK DESIGN

We use a simple task in which we vary the task load: the auditory 2-back task. This makes it possible to focus on the effects of workload, i.e. we can be certain that the changes in our measurements are due to the differences in workload, which correlate to the changes in task load.

In the auditory 2-back task the participants have to determine whether the current letter (*n*) is equal to or different from the letter that was presented two back (*n – 2*). Thus, for the sequence *C–K–C* the correct response is "equal," for the sequence like *M–G–B* "different." We use inter-stimulus intervals (ISIs) of 1s, 2s, 4s, and 6s to define the task load. The task is presented in four 10-minute blocks, where each block consists of eight segments of 72s. The task load is constant throughout each segment and defined by the number of problems the participant has to respond to in an interval (72, 36, 18, 12, respectively). We collect three types of data:

1. *Performance measures*: the response given to a problem and the time of the response,
2. *Facial features*: the users' head and eyes are monitored by three cameras and used to extract certain information,
3. *Physiological measures*: the "emotional mouse" – a track ball equipped with sensors – collects information about the users' physiological state.

PERFORMANCE MEASURES

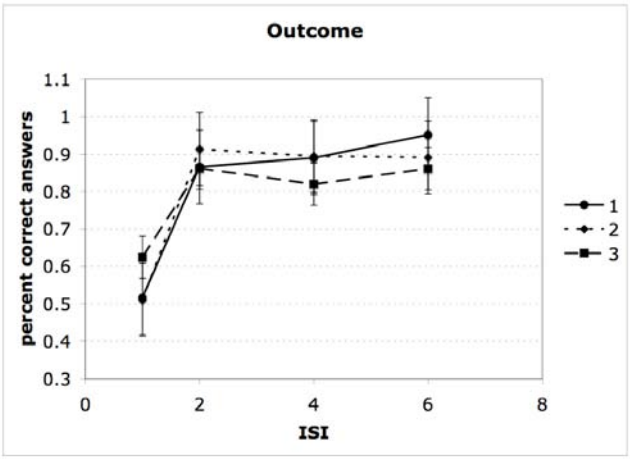
We use two performance measures:

1. *Outcome*: percentage of correct responses over the problems presented,
2. *Reaction time*: the time between the onset of the auditory stimulus and the time of the first response – regardless of whether it was correct or not.

We also analyzed variants of these measures, e.g. the reaction time to the first correct response, but these measures offer no advantages over the ones we decided on using here. Both

measures are similar between the three participants we have analyzed so far, cf. Figure 1. This means, the same task load leads to comparable performance across participants.

a)



b)

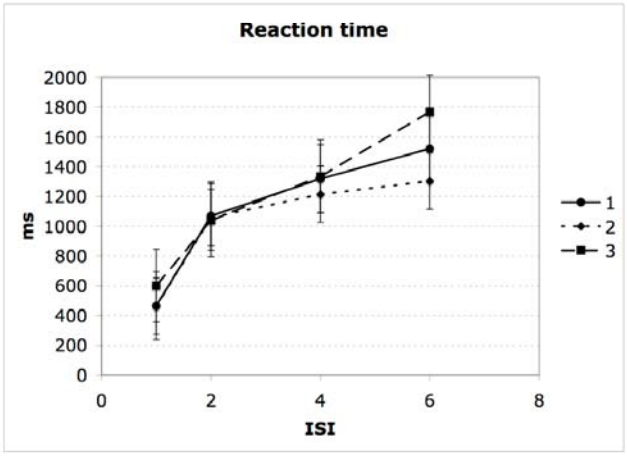


Figure 1: Performance measures for three participants in the auditory 2-back task. Error bars show the standard error

There are two major effects. First, the outcome of all participants in the ISI-2–6 conditions is almost identical (means(n=3): .89, .87, .90, respectively). Only the outcome in the ISI-1 condition is considerably lower (mean(n=3): .54), i.e. it is almost at chance level. Second, in contrast to outcome, reaction time steadily increases over all four conditions. The data also show that intra-individual variability for the reaction time increases for all participants from 48ms on average in ISI-1 to 223ms on average in ISI-6. Inter-individual variability also consistently increases with longer ISIs.

The participants have to perform fewer operations per second the longer the ISIs, i.e. their cognitive workload decreases. However, despite decreasing workload the outcome is constant in the ISI-2, ISI-4, and ISI-6 conditions, which means the participants need decreasing effort to achieve the same outcome. It also means that outcome shows a ceiling effect at approximately .9. Outcome alone is, therefore, not sufficient to measure a user's cognitive workload. This is corroborated by the reaction time data, which suggests that not only the task load but also the cognitive workload differs in these three conditions. Thus, reaction time is a better measure of cognitive workload than outcome, but the large overlap of the standard errors within participants (especially within participant 2) shows that reaction time alone does not suffice to reliably distinguish the four conditions.

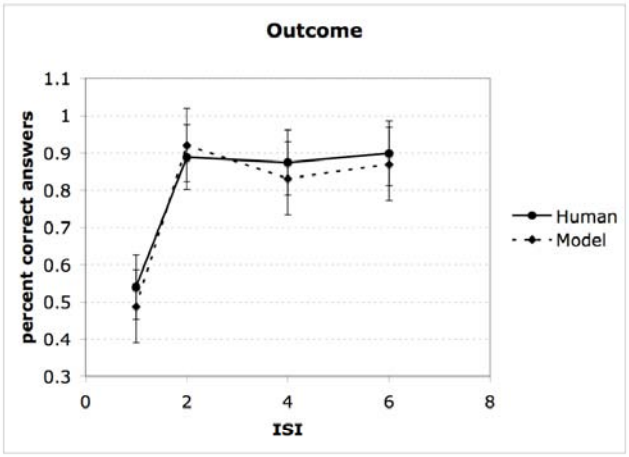
ACT-R MODEL

We created an ACT-R model that interacted with the same task environment as the participants. The model is comparably simple:

1. It waits until it perceives a new letter,
2. Retrieves the previous two letters from declarative memory,
3. Compares the new to the penultimate letter,
4. Decides whether they are equal or not,
5. Presses the corresponding button on the screen, and
6. Stores the new letter in declarative memory.

We ran each ISI condition 10 times; the results are shown in Figure 2.

a)



b)

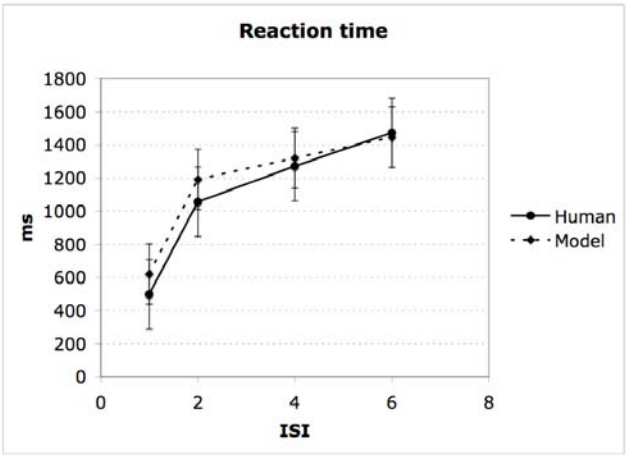


Figure 2: Performance of the ACT-R model in comparison to the average human data

Although we expected to have to build models of each individual, because there are huge differences in the facial features and the physiological measures (see below), the performance we observed in our three pilot participants was too

similar to justify this. We, therefore, built a model of the average performance.

The model fits the data well, in terms of the average values as well as in terms of the standard error. To produce these results we only changed few of ACT-R's default settings. We turned off optimized learning; see Sims & Gray (2004) on this issue. As a result we had to set the base-level constant to 1.0 (default 0.0) and chose a base-level learning rate of 0.7 (default 0.5), which causes a steeper decrease of activation of chunks stored in declarative memory. The consequence is that facts (chunks) stored in declarative memory become inaccessible earlier than with the default setting. To reduce the overall reaction time we also used a value of 100ms for the delay until a new letter is perceived after stimulus onset (default 300ms). We make the last assumption, because the participants expect the stimuli within a certain time window as they adapt to the rhythm in which the stimuli are presented.

We analyzed the behavior of the model and identified the main reason for the large number of errors in the ISI-1 condition to be the fact that it takes ca. 1.5s to produce a response to a new letter. This, obviously, is longer than the time available in this condition (1s), while it causes no problems in the other three conditions. However, the performance data of the ISI-2 condition shows that our participants were obviously able to speed up processing to some extent. In this condition the average reaction time is 1055ms, and despite this speedup they achieved an outcome identical to the slower conditions. However, this capacity for speedup is limited, as they could not reduce their reaction time enough to become better than chance in the ISI-1 condition.

That the reduced reaction time in the ISI-1 condition is not caused by a corresponding speedup of the participants' responses is corroborated by the fact that about 30% of answers occur 0–300ms after stimulus onset. (After 300ms there is a sharp drop in the number of responses.) These early responses are responses to the previous stimulus, because executing the motor movement for clicking the button takes about 300ms, and once the motor movement is initiated it cannot be suppressed, even if the next stimulus is perceived.

The steady increase in reaction time in the other three conditions is caused by an increase in the time it takes to retrieve previously stored letters from declarative memory. The reason is that the longer an element is stored in declarative memory the more its activation decays, and the less activated an element is the longer it takes to retrieve it from declarative memory.

WORKLOAD PREDICTION WITH BN MODEL

Feature Extraction

We monitor the users' visual features with a visual sensor system consisting of three cameras and their physiological measures with the "emotional mouse." All these measures are obtained non-intrusively and in real-time. The visual sensor system extracts eight visual features (Ji, Zhu & Lan 2004):

1. Blinking Frequency (BF),
2. Average Eye Closure Speed (AECS),
3. Percentage of Saccadic Eye Movement over time (PerSac),
4. Gaze Spatial Distribution (GazeDis),
5. Percentage of Large Pupil Dilation over time (PerLPD),
6. Pupil Ratio Variation,
7. Head Movement,
8. Mouth Openness

The "emotional mouse" is a track-ball with integrated sensors, which measure:

1. Galvanic Skin Response (GSR),
2. Heart Rate,
3. Body Temperature,
4. Pressure used to click the mouse button

The visual and physiological measures can be used to assess a user's cognitive workload (Ward & Marsden 2003). The experimental data show that both types of measures are sensitive to changes in ISI. Although there are significant individual differences, as we will describe below, there are also general observations. With decreasing ISI participants blink less frequently, the eyes close faster, the pupils dilate more often, and the eye gaze focuses on the screen more often and remains longer. Participants less frequently move their head, and do not open their mouth as often; they click the mouse button harder, their heat rate increases, and the Galvanic Skin Response decreases.

BN Modeling

Although a number of measures are sensitive to workload changes, using only single measures is not reliable for workload assessment. Therefore, we use a Bayesian Network (Figure 3) to fuse the various measures. Conceptually, a BN is a directed acyclic graph (DAG) that represents a joint probability distribution among a set of variables. In this graph nodes denote variables, and links between nodes denote the conditional dependencies between variables. The dependencies in the BN are characterized by a *Conditional Probability Table* (CPT) for each node. One benefit of the BN is that it is capable of representing the dependencies and semantics at different levels of abstraction between the variables.

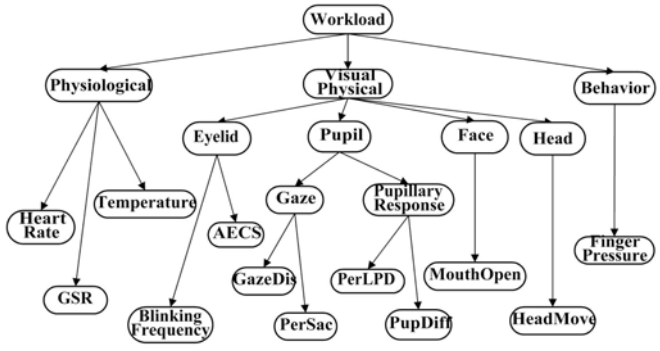


Figure 3: A Bayesian Network for modeling workload

The BN infers the cognitive workload by modeling the measures (features) from different modalities mentioned above. To model correlations among measures from the same modality, an intermediate node is introduced for each type of

measures, e.g. "Visual Physical" links "Workload" and the visual features. As stated above, we assume that the task load correlates with the cognitive workload. The cognitive workload influences the users' physical state, which, in turn, influences their observable visual and physiological features such as blinking frequency and the pressure with which they press the mouse button. For similar reasons we introduce intermediate nodes for "Eyelid", "Pupil", "Face", and "Head" to model the correlations among variables in the same group. The BN addresses the uncertainties of the workload measures, integrates the causes of workload, represents probabilistic relations among causes and various measures from multiple modalities, and provides efficient inference solutions.

BN parameterization

The BN parameterization determines the CPTs for each node. The accuracy of the parameters directly determines the performance of the BN. Currently domain experts initialize the CPTs. Then, the EM learning algorithm (Lauritzen 1995) refines the CPTs based on the data from different participants. Since the sensitivity of each individual feature to workload varies with individual participants, the learned BN parameters are particular to each participant. Figure 4 illustrates the quantified sensitivity, which is calculated as the mutual information between the workload and each feature from the BN model. The mutual information *I(W; F)* indicates how much information the random variable *F* tells about another one *W*.

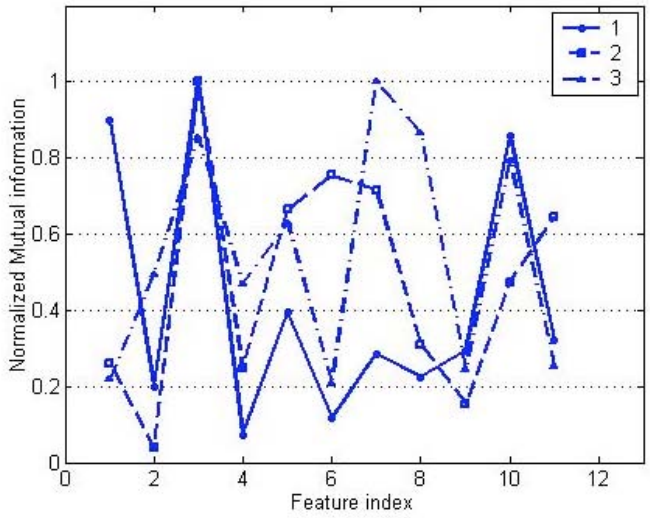


Figure 4: The sensitivity of each feature to workload for three subjects. X-coordinate is the feature index: 1-heart rate; 2-GSR; 3-Finger Pressure; 4-AECS; 5-BF; 6-GazeDis; 7-PerSAC; 8-PerLPD; 9-PupDiff; 10-MouthOpen; 11-HeadMove. The y-coordinate indicates the mutual information between the task load and each feature. The values are normalized to the range of [0 1].

BN inference

In the BN the "Workload" node is in one of four states, representing different task load levels. The input is a 12-dimensional vector quantifying the 12 measures (evidences) extracted from each interval. This corresponds to the twelve leaf nodes in the BN. The output is the posterior probability of the task load given these evidences, which is a 4-dimensional vector, *[p(WL=1), p(WL =2), p(WL=3), p(WL =4)]*, where *WL* stands for the *workload level*. Figure 5 shows the mean inferred task load for each type of ISI. This vector then is multiplied with the vector $[1\ 2\ 3\ 4]^T$ to map the probability values into different ranges. The result is the *inferred workload index* (Figure 5). Ideally, the mean values should be 1, 2, 3, and 4. As the figure shows, the mean values of the workload indexes are close to the desired values. The standard deviations are very small, which means the inferred results are stable and robust. These results show that the BN model successfully integrates various measures from multiple modalities and infers the task load level.

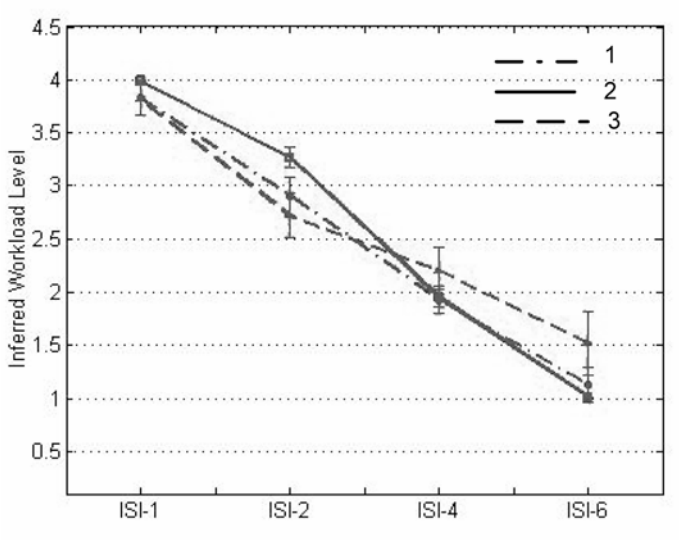


Figure 5: Inferred workload levels including error bars from the BN model. It shows a steady decrease in inferred workload with longer ISIs.

DISCUSSION

The Bayesian Network can identify the task load level simply by observing the users' behavior, i.e. without having access to information from the task environment. Since we assume a direct correlation between task load and cognitive workload for our task, the system is capable of picking up different levels of workload for each participant. This is the more remarkable as the BN uses different configurations for each participants and it learns these configurations on its own.

Comparing the workload inferred by the BN with the reaction time data (and taking reaction time as indication of workload) shows that the reaction time decreases when the workload increases. Reaction time often is a good indicator of workload (Wierwille, Rahimi & Casali 1985). However, it is apparent that this relation is not the simple "higher workload causes longer reaction time." Rather, the inverse is true. The longer a letter (an ACT-R chunk) has to be remembered, i.e. the longer the intervals between the stimuli are, the longer it takes to retrieve it (Anderson & Lebiere 1998).

Taking Figures 1b) and 5 together, one can see that the workload level inferred by the BN provides a good measure of the task load as well as of the user's cognitive workload. In addition, Figure 4 shows that the features the BN uses to infer the workload level differ substantially with the individual.

So, what do we measure, task load or cognitive workload? We measure both. Each user is subjected to different task loads, which influences him or her in different ways. This means, each ISI condition has a different effect on the user. The ACT-R model shows that the differences are differences in cognitive workload, i.e. a difference of the number of mental operations over time. The differences in workload cause different externally observable behavior. These externalized differences are picked up by the sensing systems, i.e. they manifest themselves as differences in facial and physiological properties. From these differences the BN infers that the user is currently experiencing a particular workload. Since the BN has no access to information about the current ISI condition, i.e. the current task load, what it measures must be the cognitive workload of the user. However, since the classification into the four different experimental conditions (the four different task loads) works almost perfectly, this means (1) task load and cognitive workload do indeed correlate, (2) the BN measures not only task load but also workload.

This view is consistent with our ACT-R model, which offers a convincing explanation of the differences in performance. The difficulty the participants experience in the ISI-1 condition is due to the fact that generating a response without time pressure takes about 1.5s, which is longer than the participants have available in this condition. Even though they are able to speed up their responses to some extent (1055ms on average in ISI-2) this does not suffice to generate an outcome better than chance. The increase in reaction time over all conditions is caused by the time between stimuli: the longer this time span the longer it takes to retrieve previous stimuli from memory.

CONCLUSION

We present a new method to measure workload that offers several advantages. First, it uses non-intrusive means: cameras and the emotional mouse. Second, workload is measured in real-time. Third, the setup is comparably cheap: the cameras are standard "webcams" and the emotional mouse contains off-the-shelf sensors. Fourth, we go beyond simply measuring performance (outcome and reaction time) and demonstrate that just using such measures (outcome in particular) does not suffice to measure workload, because the same outcome can be achieved despite a different workload. Fifth, since we use a BN model to assess the workload from the various manifesting measures, the model adapts itself to the individual user as well as to a particular task. Sixth, we use a cognitive computational model to gain insights into the underlying cognitive mechanisms the participants use, and this model corroborates our explanations.

ACKNOWLEGMENTS

Support for the work reported here was provided by the Defense Advanced Research Projects Agency (DARPA) through ONR grant #N00014-03-1-1003.

REFERENCES

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Hillsdale, NJ: Lawrence Erlbaum.

De Waard, D.(1996). *The measurement of drivers' mental workload*, Ph.D. thesis, University of Groningen.

Ji, Q., Zhu, Z.W., & Lan, P.L. (2004). Real Time Non-intrusive Monitoring and Prediction of Driver Fatigue. *IEEE Trans. Vehicle Technology*, July 2004.

Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis, 19*, 191-201.

O'Donnell, R.D., & Eggemeier, F.T. (1986). Workload assessment methodology, *Handbook of perception and human performance*, 2(42), 1-49, 1986.

Picard, R. W. (1997). *Affective computing*. Cambridge, MA: MIT Press.

Sims, C., & Gray, W. D. (2004). Episodic versus semantic memory: An exploration of models of memory decay in the serial attention paradigm. In M.C. Lovett, C.D. Schunn, C. Lebiere, & P. Munro (Eds.). *Proceedings of the 6th international conference on cognitive modeling – ICCM-2004*, 279–284. Pittsburgh, PA.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, 3(2), 159–177.

Wierwille, W. W., Rahimi, M., & Casali, J. G. (1985). Evaluation of 16 Measures of Mental Workload using a Simulated Flight Task Emphasizing Mediational Activity. *Human Factors*, 27(5), 489–502.