To appear in 2005 Human Factors and Ergonomics Society Annual Meeting, Orlando, Florida.

Non-intrusive measurement of workload in real-time

Markus Guhe, Wenhui Liao, Zhiwei Zhu, Qiang Ji, Wayne D. Gray & Michael J. Schoelles
Rensselaer Polytechnic Institute

1. Introduction

Workload is a comprehensive concept with many aspects. Simplistic definitions of workload are 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 an important research topic 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 use an "external" notion of workload, namely the number of operations that have to be performed in a given time (and assume that it correlates in a principled fashion with the users' internal, "cognitive" workload). We develop a real-time, non-intrusive system that measures and detects the differences in external workload. We focus on two aspects. First, we study the relations between workload and variance measures (outcome, response time, physical response, and physiological response). Second, we propose a *Bayesian Network* (BN) model, a framework for reasoning under uncertainty, to predict the workload from the various manifesting measures. The model adapts itself to the individual user as well as to a particular task. We consider the latter an important part of our research, because individuals express differences in workload, which is analogous to the differences in corresponding affective states, such as stress (Picard 1997). This is necessary to understand and predict a user's behavior, e.g. for an estimation of how likely the user will make an error. For this reason we currently focus on only a few participants (three), before we validate the system with a large number of participants.

2. Task Design

We use a simple task to set the participants' workload – the auditory 2-back task. This makes it possible to concentrate on the effects of workload, i.e. we can be certain that the changes in our measurements are due to the differences in external workload.

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 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 workload. Tasks are presented in four 10-minute blocks, where each block consists of eight intervals of 72s. The workload is the number of problems the participant has to respond to in an interval (72, 36, 18, 12, respectively). Data are collected in real-time by three cameras, which extract certain facial features, and with the "emotional mouse", a track ball equipped with sensors to collect physiological and behavioral measures.

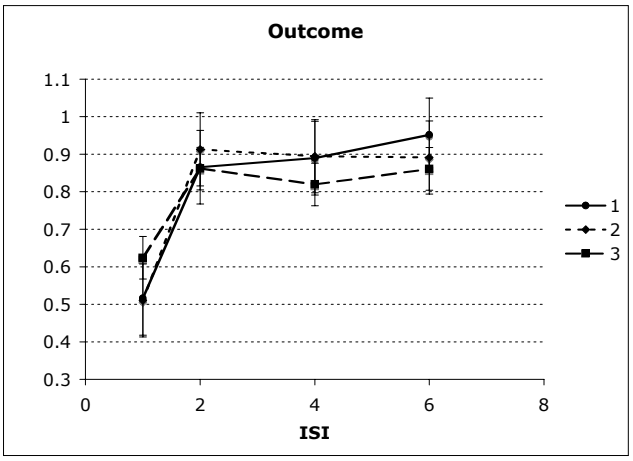
3. Performance Measures

We use two performance measures:

1. *Outcome*: percent 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.

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

a)



b)

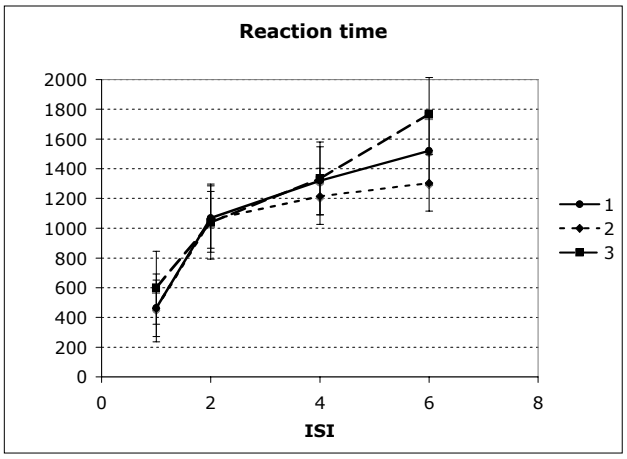


Figure 1: Performance measures for three participants in the auditory 2-back task. Error bars show the standard error

The outcome shows the same pattern for each participant over all four conditions. The most apparent effect is that 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). In contrast to outcome, reaction time shows a steady increase over all four conditions. 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 consistently increases with longer ISIs.

The data show that despite decreasing workload the outcome is constant in the ISI-2, ISI-4, and ISI-6 conditions. Thus, the participants need decreasing effort to achieve the same outcome. It also means that outcome shows a ceiling effect at approximately .9. For this task, outcome alone is, therefore, not sufficient to measure a user's workload, because the reaction time data suggests that not only the external workload but also the cognitive workload (defined as the number of mental operations per time) differs in these three conditions. Reaction time also is a better measure of workload than outcome, but the large overlap of the standard errors within participants (especially within participant 2) shows that no reliable classification into the four workload conditions is possible.

4. Workload Prediction with BN model

4.1 Feature Extraction

We monitor the users' visual features with a visual sensor system (three cameras) and the 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 (PerSac)

4. Gaze Spatial Distribution (GazeDis),
5. Percentage of Large Pupil Dilation (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,
2. Heart Rate,
3. Body Temperature,
4. Pressure used to click the mouse button.

The visual and physiological measures can be used to assess the user's workload (Ward & Marsden 2003). The experimental data show that both types of measures are sensitive to changes in ISI. As ISI decreases, 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 open their mouth; they click the mouse button harder, heat rate increases, and GSR decreases.

4.2 BN Modeling

Although a number of measures are sensitive to workload changes, our study shows that single measures are not reliable. We, therefore, use a Bayesian Network (Figure 2) to combine them to assess workload. Conceptually, a BN is a directed acyclic graph (DAG) that represents a joint probability distribution among a set of variables. Nodes denote variables, links between nodes denote the conditional dependencies between variables. The dependencies are characterized by a *conditional probability table* (CPT) for each node.

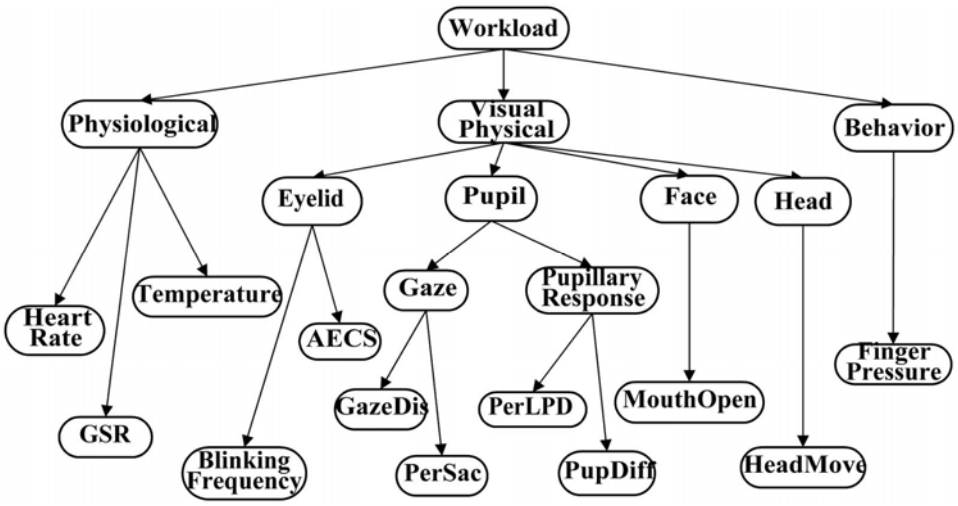


Figure 2: A Bayesian Network for modeling workload

The BN infers workload by modeling the measures (features) from different modalities mentioned above. It addresses uncertainty in the workload modeling, integrates the causes of workload, represents probabilistic relations among causes and various measures from multiple modalities, and provides efficient inference solutions.

4.3 BN parameterization

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 3 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.

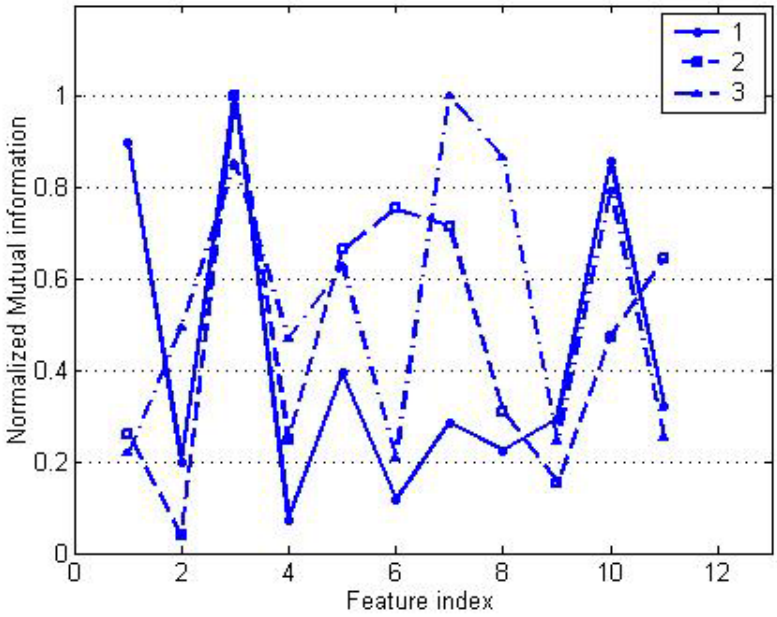


Figure 3: The sensitivity of each feature to workload for three subjects. X-coordinate is the feature index: 1-heart rate; 2-GSR; 3-Finger Pressure; 4-AECS; 5-BF; 6-GazeDis; 7-PerSAC; 8-PerLPD; 9-PupDiff; 10-MouthOpen; 11-HeadMove. The y-coordinate indicates the mutual information between the workload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

4.4 BN inference

In the BN the "Workload" node has four states, representing different workload levels. The input is a 12-dimensional vector quantifying the 12 measures (evidences) extracted from each interval, which corresponds to the twelve leaf nodes in the BN. The output is the posterior probability of the workload given these evidences, which is a 4-dimensional vector, *[p(WL=1), p(WL =2), p(WL=3), p(WL =4)]*, where WL represents "workload level". Figure 4 shows the mean inferred workload under each ISI level. 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 4). 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. The results show that the BN model successfully integrates various measures from multiple modalities and infers the workload level.

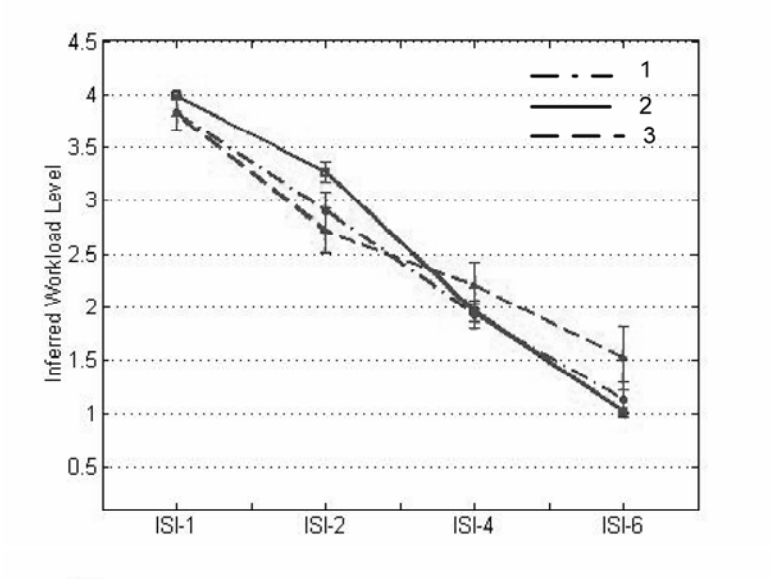


Figure 4: Inferred workload levels including error bars from the BN model. It shows a steady decrease in inferred workload with longer ISIs.

5. Discussion

Comparing the workload that is inferred by the Bayesian Network with the reaction time 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. A more plausible explanation is, therefore, that the participants assess the time they have to respond to the stimuli and that with longer ISIs their general level of arousal decreases. A decrease in arousal increases the time for memory retrievals and also causes the other processes involved in performing this task to slow down (Anderson & Lebiere 1998).

Taking Figures 1b) and 4 together, one can see that the workload level inferred by the BN provides a good measure of the external as well as the internal workload of the user. In addition, Figure 3 shows that the features the BN uses to infer the workload level differ substantially with the individual.

6. 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, the 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 differences in 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.

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. Veh. 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.

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.