Feasibility and Pragmatics of Classifying Working Memory Load with an Electroencephalograph

David Grimes[1,2], Desney S. Tan[2], Scott E. Hudson[3,2], Pradeep Shenoy[1], Rajesh P.N. Rao[1]

[1]University of Washington
Box 352350, Seattle, WA 98195
{grimes,pshenoy,rao}@
cs.washington.edu

[2]Microsoft Research
One Microsoft Way
Redmond, WA 98052
desney@microsoft.com

[3]Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
scott.hudson@cs.cmu.edu

ABSTRACT

A reliable and unobtrusive measurement of working memory load could be used to evaluate the efficacy of interfaces and to provide real-time user-state information to adaptive systems. In this paper, we describe an experiment we conducted to explore some of the issues around using an electroencephalograph (EEG) for classifying working memory load. Within this experiment, we present our classification methodology, including a novel feature selection scheme that seems to alleviate the need for complex drift modeling and artifact rejection. We demonstrate classification accuracies of up to 99% for 2 memory load levels and up to 88% for 4 levels. We also present results suggesting that we can do this with shorter windows, much less training data, and a smaller number of EEG channels, than reported previously. Finally, we show results suggesting that the models we construct transfer across variants of the task, implying some level of generality. We believe these findings extend prior work and bring us a step closer to the use of such technologies in HCI research.

Author Keywords: Brain-Computer Interface (BCI), electroencephalogram (EEG), cognitive load, memory load, machine-learning, feature selection, classification.

ACM Classification Keywords: H.1.2 [User/Machine Systems]; H.5.2 [User Interfaces]: Input devices and strategies; B.4.2 [Input/Output Devices]: Channels and controllers; J.3 [Life and Medical Sciences].

INTRODUCTION

Human-computer interaction (HCI) researchers continually work on techniques that allow us to measure user states such as cognitive and memory workload, task engagement, surprise, satisfaction, or frustration. Such measures are useful not only for evaluating the efficacy of interfaces, but also for providing real-time information to systems that dynamically adapt and support users' goals [4]. Since we have traditionally interacted with computers through our physical bodies, most of these techniques have been based on observations of user actions and behavior (e.g. [23]). Less frequently, other techniques have utilized physiological signals as indicators of user state [14,21]. While these measures have been reasonably successful, they are rather indirect, especially when the user state in question is of a cognitive nature. Fortunately, advances in cognitive neuroscience and brain-sensing technologies provide us with the ability to interface more directly with the human brain. This is possible through the use of sensors that monitor the electrical and chemical changes within the brain that correspond with certain forms of thought. While using these technologies in HCI research has been previously articulated [12,28], we believe there is an opportunity to further explore practical issues with their use in HCI applications.

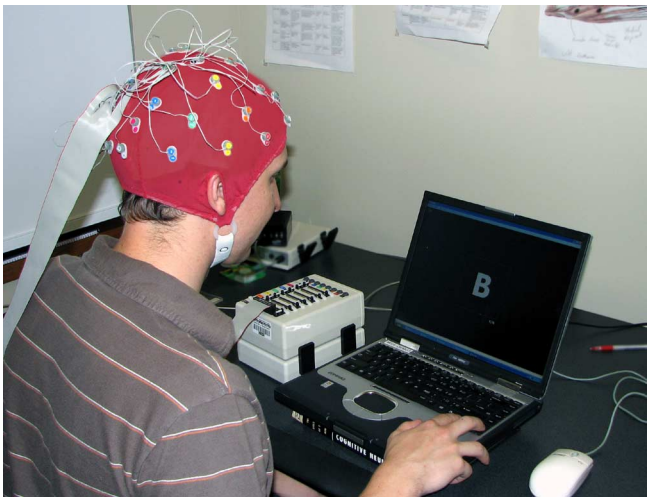

Figure 1: Measuring working memory load with an electroencephalograph. Shot taken from experiment.

In our work, we explore using one of these technologies, an electroencephalograph (EEG), to estimate or *classify* working memory load, or the cognitive effort dedicated to holding information in the mind for short periods of time while performing a cognitive task [1]. Working memory has been shown to be a key component of cognitive load, and is a

CHI 2008, April 5–10, 2008, Florence, Italy.

reasonable measure of how hard a user is working to solve a problem or use an interface. For example, working memory load has long been recognized in HCI to be an important indicator of potential errors as well as a predictive feature of procedural skill acquisition [3]. Given this evidence, interface designers often try to minimize the working memory load required to perform a task, and reliable real-time measures would benefit them greatly.

While various researchers have worked on classifying working memory with EEG (e.g. [6,7,20,21,27]), previous work has typically relied on costly equipment and techniques that make it difficult for non-EEG-experts to replicate and use this work. Additionally, this work has often required experimenters to collect large amounts of classifier training data (sometimes on the order of days), a process that is often prohibitively expensive. While we believe that EEG is complementary to many of the other measures of memory and cognitive load, it is outside the scope of this paper to explore the detailed relationships between these measures. We leave this for future work.

The contributions of this paper are three-fold:

- First, we present our methodology within an experiment we ran to measure working memory load using only EEG signals. The innovation within this methodology is an automatic feature selection scheme that eliminates the need for procedures used in most previous work, such as complex device and physiological drift modeling as well as manual artifact rejection.
- Second, using this methodology, we present classification results using machine learning techniques that replicate and extend prior work in the area. Specifically, we show classification accuracies of up to 99.0% between two load levels, and up to 88.0% between four levels, all with just 8 channels of EEG data. More importantly, we present results showing how classification accuracy varies with different temporal window sizes, amounts of training, and number of EEG channels. Specifically, the results suggest that our techniques allow us to attain accurate classification with less lag, much less training data, and simpler equipment.
- Third, we show how our models work across variants of the memory task, providing encouraging evidence that it might be possible to develop canonical training tasks and to perform general classification of memory load.

RELATED WORK

EEG Primer

In this paper, we use an Electroencephalograph (EEG), a sensing technology that uses electrodes placed on the scalp to measure electrical potentials related to brain activity (see Figure 1). Each electrode typically consists of a wire leading to a conductive disk that is electrically connected to the scalp using conductive paste or gel. The EEG device records the voltage at each of these electrodes relative to a reference point, which is often another electrode on the scalp. Because EEG is a non-invasive, passive measuring device, it is safe for extended and repeated use, a characteristic crucial for adoption in HCI research. Additionally, it does not require a highly skilled operator or medical procedure to use. For more information about electrical signals generated by the brain as well as EEG, see [5].

The signal provided by an EEG is, at best, a crude representation of brain activity due to the nature of the detector. Scalp electrodes are only sensitive to macroscopic and coordinated firing of large groups of neurons near the surface of the brain, and then only when they are directed along a perpendicular vector relative to the scalp. Additionally, because of the fluid, bone, and skin that separate the electrodes from the actual electrical activity, the already small signals are scattered and attenuated before reaching the electrodes.

EEG data is typically analyzed by looking at the spectral power of the signal in a set of frequency bands, which have been observed to correspond with certain types of neural activity [5]. These frequency bands are commonly defined as 1-4 Hz (delta), 4-8 Hz (theta), 8-12 Hz (alpha), 12-20 Hz (beta-low), 20-30 Hz (beta-high), and >30 Hz (gamma).

EEG for Cognitive State Evaluation

Early researchers observed the sensitivity of EEG to changes in mental effort. For example, Hans Berger [2] and others [11] report observing a decrease in the amplitude of the alpha (8-12 Hz) rhythm during mental arithmetic tasks Other researchers have shown that higher memory loads cause increases in theta (4-8 Hz) and low-beta (12-15 Hz) power in the frontal midline regions of the scalp [17], gamma (>30 Hz) oscillations [8], as well as inter-electrode correlation, coherence, cross phase, and cross power [24].

To test if alpha and theta bands were predictive of memory and cognitive loads in real world computing tasks, Smith et al. [27] compared EEG data when task difficulty was manipulated within a multi-attribute task battery (MATB) multitasking environment. They report successfully creating a user-specific index of task load, the average values of which increase with increasing task difficulty and differed significantly between the difficulty manipulations.

Given this evidence of the existence of reliable indicators of memory load, researchers have attempted to build techniques that utilize these features to measure and classify memory load. Unfortunately, while these indicators may appear to be reliable when data is averaged over large time periods and many users, there is large variability within the signal for any given user at any given point in time. This makes using the features to classify memory loads an extremely difficult task. While it is reasonable to average the data when trying to make statements about the various rhythms, it is less useful when trying to classify user state in real time. For example, Jensen et al. found the increased theta power in only one of their ten subjects, and rather than an alpha decrease, they found that alpha power actually

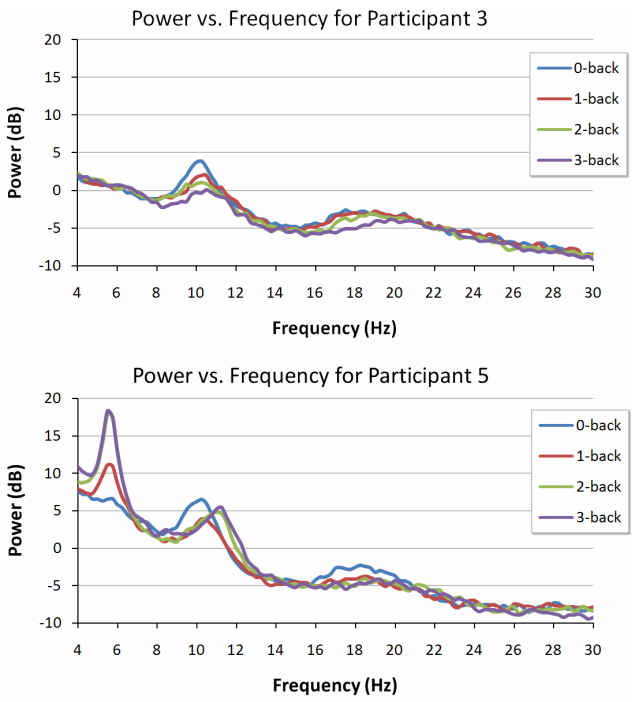

Figure 2: Example of different features and valences of changes showing up within different participants. Participant 3 shows decreased alpha (8-12 Hz) power with increasing load, while participant 5 shows the opposite. Participant 5 also shows strong theta (4-8 Hz) discriminability, but participant 3 does not.

increased in many of their users [10]. However, the theta power they found in the one user was so large as to cause the data to exhibit the predicted trends when averaged across all users. Prinzel et al. [25] report building systems that adapt based on a task load metric they derived from the powers of beta/(alpha+theta) in the EEG signal. This raises questions because it again goes counter to the findings that alpha decreases and theta increases with higher task load.

We have seen similar differences in our work, where EEG from an individual user may exhibit certain predicted characteristics, but not all. For example, taken from our current set of experiments, Figure 2 shows very different frequency response patterns for two of our users. Just as in Jensen et al. and Prinzel et al., we found that users had signals that were inverted from the expected levels. Again, when averaged across large numbers, the data typically exhibited the previously reported trends. In this paper, we assume that these characteristics are due mainly to individual differences and that we can build reliable models if we train on individual users. This also points to the importance of proper feature generation and selection as this is the phase that will account for most of the individual differences.

In their experiments, Gevins and colleagues demonstrate impressive classification results of working memory and cognitive workload using EEG data (see [6] for a summary of this work). In one specific experiment, Gevins and Smith collected data from 8 users over three 6-8 hour sessions and present results showing ~95% classification accuracy between two levels of memory load [7]. They also showed relatively high cross-task and cross-session accuracies.

However, subtle decisions made in their procedure leaves room for improvement. First, collecting 24 hours worth of training data from each user can be prohibitively high for some work. Second, they perform a Laplacian spatial enhancement that requires accurate per-subject head measurements to filter noise from the signal. Third, they manually inspect the data and throw out periods where there are artifacts in the data even after performing an automatic artifact rejection. This is tedious and requires expertise in reading the EEG signals. They report throwing away up to 20% of the data, which is not desirable in our targeted settings where data may be scarce. Furthermore, having to perform this manual step between training and classification has implications on real-time usability of the system. Finally, since their design interleaved different tasks, and used random hold out cross validation, they were training on data that was temporally fairly close to test data and we cannot be certain how well the models would generalize when applied to new data. In our work, we aim to replicate their high classification results and extend their work to further explore the space. We also set out to explore how various parameters such as temporal window size, amount of training data, and number of channels affect the classification. These factors are important to understand if EEG classification is to be used in HCI settings.

EXPERIMENT

The overarching goal of this work is to extend prior work and to bring us closer to understanding the use of electroencephalographs to measure working memory load in real world human-computer interaction applications. In this experiment, we used a simple memory task that provided us with a level of control not otherwise possible with more complex tasks. We describe our classification methodology and demonstrate our ability to accurately classify between different task difficulties. We also describe the accuracy tradeoffs that exist along various dimensions. These dimensions include: temporal window size, which implies lag when performing the classification in real time; training period, since we would like to minimize this; and number of EEG channels, which has implications on device and setup cost. Finally, we show that the models we build generalize to different variants of the task.

Participants

Eight (4 female) university students and recent graduates volunteered for this experiment. The average age of participants was 26.6, ranging from 18 to 34 years of age. All participants had normal or corrected-to-normal eyesight, and none were color blind. Also, none of the participants had any known neurological disorders. Two of the participants were left-handed. The entire experiment took about three hours and participants were paid for their time.

Task

In this experiment, we employed the *n-back* task, an experimental paradigm that has been used extensively in functional neuroimaging studies of working memory and cognitive load [6,7,19]. In the n-back task, participants are presented a series of stimuli (e.g. letters of the English alphabet), one at a time. At each presentation, or trial, they respond with whether or not the current stimulus is the same as the one that they saw some number, n (e.g. 1, 2, 3, etc), presentations ago. Hence, for each trial, participants have to keep a sequence of stimuli in memory, perform a matching task, and then update the sequence with the new stimulus. See Figure 3 for a graphical representation of the 3-back task flow. In the 0-back task, participants have to compare each stimulus with the first one seen in the series, making this solely a matching task with no updating required. Prior work has shown that increasing n, the number of items a user has to remember, increases the working memory load of the user [e.g. 7,19]. A particularly nice property of the n-back task is that the perceptual and motor demands remain constant across difficulty levels. This is important because it allows us to ensure that we are measuring memory load and not some reaction to differing stimuli.

For each trial, we presented the stimulus for 1 second in a 400×400 pixel square centered on the screen, and then removed it and left the screen blank for 3 seconds. Users responded with their answer within this 4 second trial window. Trials were presented back to back every 4 seconds.

We grouped trials into sequences, each containing n preloading trials followed by 30 test trials. Each sequence took 2 minutes and was the unit for which we manipulated task difficulty. Sequences of trials were created by randomly assigning stimuli from a fixed pool of 8 items, with the following constraints: 1) One third of the sequence (10 trials) were matches. That is, the stimulus presented was the same as the one shown exactly n-back in the sequence. 2) One third were non-matches with foils. That is, even though the stimulus presented was not the same as the one n-back, there was a stimulus presented more recently than n-back that matched the current one, hence making the exact location within the sequence important. 3) The remaining third were non-matches that did not include foils. See Figure 3 for examples of each of these cases.

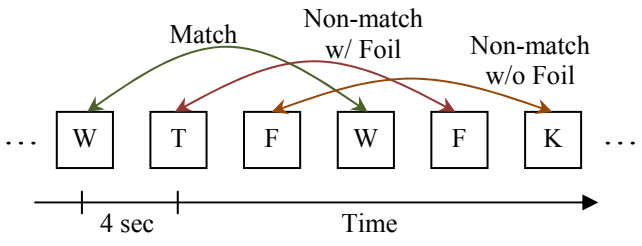

Figure 3: Graphical representation of the 3-back task. This example shows a match (W and W), a non-match with a foil (F and T with a foil, or distractor, F in between), and a non-match without a foil (K and F).

Design and Procedure

We were interested in developing techniques that allow us to classify memory load not only when training and testing on the same task, but also when training on one task and testing on another. Hence, we ran three distinct n-back tasks in this experiment, each drawing from a distinct set of stimuli. The first set consisted of 8 letters, randomly chosen from the set of English consonants. We did not include vowels because we wanted to make it difficult to chunk the letters into a single word or phoneme. The second set consisted of images chosen to have comparable familiarity, complexity, and image ratings as measured by the normative data provided from a revised Snodgrass and Vanderwart object set [26]. All images also had above 95% naming precision, indicating that different people reliably named objects using exactly the same word. The third set consisted of eight spatial locations contained within the 400×400 pixel presentation square. We divided the square into nine equal sub-squares, and colored one of the boundary sub-squares solid white for each of the locations. The center was never colored. We call these tasks the letter task, the image task, and the spatial task, respectively.

We used a randomized block design, each block consisting of a randomly ordered 0, 1, 2, and 3-back sequence. The experiment was within-subjects, with each participant performing multiple blocks in all three tasks. To limit the length of the experiment, we collected more data with the letter task than the other two. This allowed us to use letter task data to explore within-task classification and data in the other tasks for cross-task validation. All participants started by performing a practice block followed by 6 test blocks of the letter task. We always started with the letter task as this seemed representative of how a user may train a system on a standardized task before using it for real-time classification on other tasks. Participants then performed a practice block and 2 test blocks in each of the image and spatial tasks, the order of which was counterbalanced.

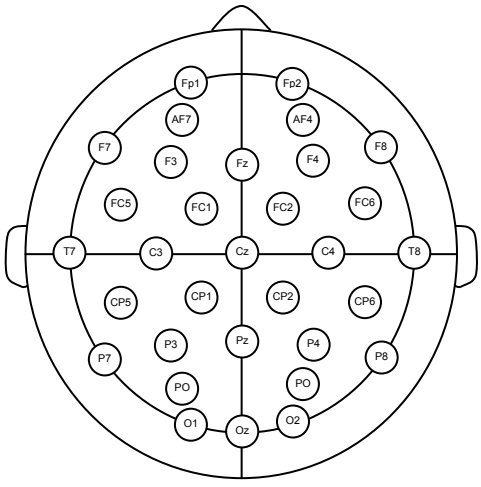

Figure 4: Layout of the channels used in our experiment, according to the internationally accepted 10-20 system of EEG electrode placement.

Equipment and Dependent Measures

We collected EEG data using a Biosemi ActiveTwo 32-channel system (www.biosemi.com), sampling at a rate of 2048 Hz. As seen in Figure 4, we placed electrodes at approximately evenly spaced locations on the scalp using the internationally accepted 10-20 system [9]. We did not control or eliminate any of the traditionally considered noise elements (e.g. 60 Hz power hum, etc.) found in the experimental environment, a university office. Before beginning, the experimenter explained the EEG device and requested that participants try to reduce unnecessary physical movements during the testing phases of the experiment. Since this was not enforced, the extent to which motion artifacts impact performance is fully manifested in our results.

Seven of the participants performed the tasks on an IBM Thinkpad laptop with its 14.1" LCD screen running at 1024×768 pixels. Due to technical problems, the last participant used a Compaq Evo N800C laptop with a 15" LCD screen running at 1024×768 pixels. Analysis showed no significant behavioral or classification differences between this participant and the other seven. Each participant sat at a comfortable distance from the laptop (about 30") and provided answers with the left ("non-match") and right ("match") arrow keys on the laptop keyboard. All users used their right hand to provide input. As visual feedback, their answer highlighted at the bottom of the screen when they hit a key. They were free to change this as many times as they liked during the 4 second trials.

We collected response times and the accuracy of answers as dependent measures. We also collected subjective ratings of the difficulty of the tasks and participant confidence in their answers after each set of 2 test blocks. At the end of the experiment, we collected subjective ratings of difficulty and participant confidence levels across the three tasks.

Behavioral and Subjective Results

In this section we describe the performance and subjective results, which provide evidence that our Difficulty manipulation was effective.

Performance Results

We performed independent 4 (Difficulty: 0-back v. 1-back v. 2-back v. 3-back) × 3 (Task: letter v. image v. spatial) repeated measures analysis of variance (RM-ANOVA) for the response time and accuracy measures.

For response time, we observed a main effect of Difficulty ($F_{3,21}$=23.885, p<.001), with higher difficulty leading to longer response times. See Figure 5 (top) for a chart of the medians. Posthoc tests revealed that all Difficulty levels were significantly different from each other (p<.001), except for the adjacent ones, the 1-back v. 2-back and the 2-back v. 3-back. All posthoc tests we report in this paper were corrected using Bonferonni adjustment for multiple tests. We also observed a main effect of Task ($F_{2,14}$=5.013, p=.023), driven by longer response times in the spatial task than the image task (means: 869.45 v. 739.11 seconds respectively, p<.05). Finally, we found an interaction between Difficulty and Task ($F_{6,42}$=3.268, p=.01), driven by responses in the 2-back and 3-back being comparatively slower in the spatial task than the other two tasks.

Analysis of accuracy data revealed a main effect of Difficulty ($F_{3,21}$=10.824, p<.001), driven by lower accuracies in the 3-back than the 0-back (p=.084), 1-back (p=.049), and 2-back (p=.062), see Figure 5 (bottom). We found no other main effects or interactions in the accuracy data.

We also looked at learning effects as users progressed through the experiment. We found learning effects within the letter task, but not within any of the other tasks or between the sets of tasks. Within the letter task, response time showed significant effects of trial number ($F_{5,15}$=7.981, p<.001), driven by significantly slower responses in the first trial than the third (p=.026), forth (p=.044), fifth (p=.010), and sixth (p=.003). Accuracies showed marginally significant effects of trial within the letter task ($F_{5,15}$=2.093, p=.07), driven by lower accuracies in the first trial than the fifth (p=.01) or sixth (p=.01). These results suggest that users were still adjusting to the task in the first trial or two. While dropping these from our EEG classification efforts might have increased our classification accuracies, we decided to leave them in to get a conservative lower bound estimate of how well we can do, even with noisy data.

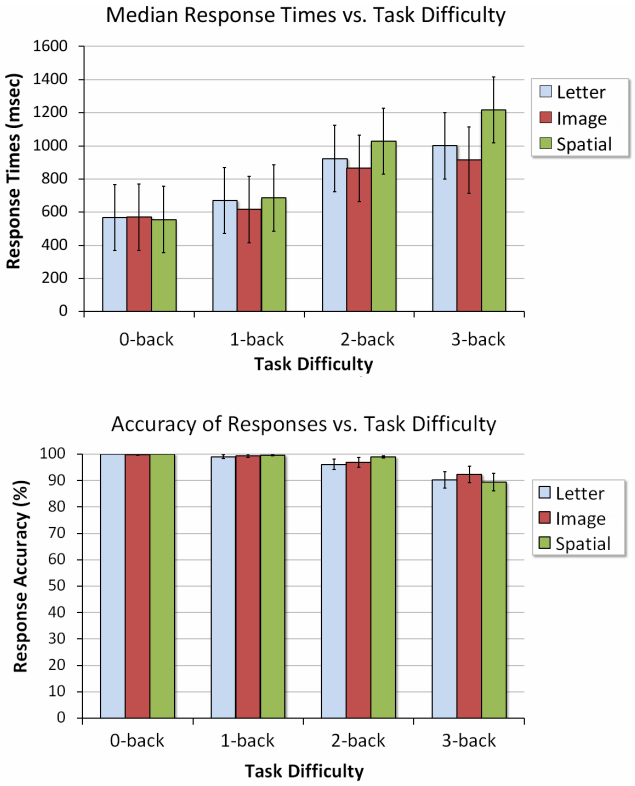

Figure 5: Response times significantly increase (top) and responses become less accurate (bottom) when memory set size, n, increases in the n-back task. Error bars in all graphs represent +/- SEM.

Subjective Results

At the end of each set of two blocks, we asked participants to rate both the perceived difficulty as well as confidence in their answers for each of the Difficulty levels. Participants responded on a 7-point Likert scale, with increasing ratings representing less difficulty and higher confidence in their answers. We ran a Friedman's Chi-Square test on these subjective ratings. We observed a main effect of Difficulty in the perceived difficulty metric ($\chi 2(3)=22.48$, $p<.001$) as well as the confidence in answers metric ($\chi 2(3)=21.08$, $p<.001$). As expected, our higher Difficulty levels resulted in subjective ratings of higher difficulty (mean ratings: 6.85, 6.7, 5.33, 2.88, for 0, 1, 2, and 3-back respectively) and lower confidence (mean ratings: 6.68, 6.5, 5.33, 3.1). At the end of the experiment, participants rated the perceived difficulty and confidence across each of the three tasks. We found no significant effects with this data, though means seem to support performance results showing that the spatial task might have been slightly harder than the rest (mean ratings: 5.0, 4.5, 2.5, for the letter, image, and spatial tasks) and had participants less confident of their results in this task (mean ratings: 5.0, 5.0, 4.0).

Classification Methodology

With the above results suggesting that our task was working as expected, we move to classification of the EEG data. In this section we describe the different phases of our classification methodology. First, we perform basic signal processing to remove extraneous data and transform the time series data into a time independent data set. Next, we compute a set of features based on prior neurophysiological evidence, some of which we mathematically combine. Then, we use a novel feature selection process to prune the feature set, keeping only those that add the most information to the classifier without over-fitting, representing uninteresting artifacts such as drift, or overlapping with each other. Finally, we use these features to build a Naïve Bayes density model and perform classification. We describe each of these phases in the following subsections.

Basic Signal Processing

We measured EEG data for training and classification during each block of *n*-back trials excluding the 4-12 second pre-loading period in which it is difficult to reason about the user's cognitive load. This provided us with 4800 seconds (80 minutes) of useful 32-channel data for each participant: 2880 seconds in the letter task, and 960 seconds each in the image and spatial tasks. While we could have discarded data from erroneous responses or used event-related potentials (e.g. [17]) to derive more signal, we feel that detecting and linking behavioral responses as well as discrete stimuli presentation to the data may not always be possible. As a result, we do not demarcate any data as being special in our analysis.

As the recording rate of 2048 Hz is higher than required given our intended feature set, we down-sampled the data using a low pass filter to obtain a 256 Hz signal. We then divide the EEG signal into multiple overlapping windows, as done in previous work [15], and transform data from each window into the frequency domain using power spectral density (PSD) estimation. Each of these windows is treated as a semi-independent data-point.

Feature Generation

Adopting select features used in previous work [7,15], we compute the following for each instance: signal power in each of the standard six EEG frequency bands for each channel as well as phase coherence (similarity in mean phase angle) in each band across select pairs of channels. Additionally, we calculated signal powers from the 4-13 Hz band in 1 Hz intervals, from 13-31 Hz in 2 Hz intervals, and from 32-50 Hz in 4 Hz intervals. We projected that this would provide higher resolution estimates and more distinct features in these smaller power bands, where we expected the most information. Since prior work provides evidence of elevated theta powers and depressed alpha powers with increasing cognitive load, we also included theta/alpha and theta/(alpha+beta) features for each channel. This led to anywhere from a few hundred to over a thousand features depending on the number of channels considered.

Feature Selection and Classification

Once we have generated our full set of features, we apply an efficient feature selection process to select the most predictive and robust features. To do this, we use a relative information gain criteria evaluated for each feature. Information gain is estimated by discretizing each feature into 15 equally spaced bins and calculating mutual information based on a Naïve Bayes density model [16]. We label the information gain that represents how predictive the features are of the Difficulty condition, IG_c.

EEG data is often affected by device and physiological drift. This observation is not new, and many researchers are working on carefully modeling device and physiological drift so that they can somehow subtract it, leaving only useful signal [13]. Unfortunately, modeling the drift accurately is difficult in controlled settings, and even more so in the uncontrolled settings such as a typical office environment. We propose selecting features that have not been corrupted by slowly varying signal noise which initially appear predictive of the Difficulty condition.

To do this, we use the same process for calculating IG_c to calculate the information gain that represents how predictive the features are of a particular contiguous sequence of n-back trials, IG_s. Due to the randomized block design, in which sequences of trials would be repeated a number of times in random order throughout the experiment, we can distinguish between robust and spurious correlations. Since we would like to maximize how predictive the features are of any condition and minimize how predictive they are of contiguous sequence, we use the ratio IG_c/IG_s to obtain our final information gain metric. We found that our gain met-

ric was also able to automatically remove features which were corrupted by muscle artifacts such as eye blinks.

We then use the 30 features with the highest IG_c/IG_s values to train our classifier. This number was empirically chosen from pilot data, but the models do not seem sensitive to this. Classification uses the same discretized Naïve Bayes density model to select the maximum a posteriori class.

Classification Results

We evaluated the classification accuracy using a cross validation scheme that accounted for the block design of the experiment. With traditional cross validation, random samples are held out from the training set and used to test the accuracy of the model. Unfortunately, doing this would provide training instances that occur adjacent to test instances. This produces significantly overestimated accuracy estimates and would not be representative of real world, real time use in HCI settings. We caution readers to be aware of this when comparing results between studies. In our block cross validation scheme, we hold out an entire contiguous sequence of trials for testing in each fold. This is more representative of performance if a new sequence were recorded and tested. In this experiment, we conducted a 24-fold block hold out cross validation for the letter task. In this section we report on the classification accuracies for within-task training and testing using the letter task, followed by the cross-task results in which we train on the letter task and test on the image and spatial tasks.

Within-Task Results

We present the within-task classification accuracies as a series of parametric results, varying temporal window size, amount of training, and the number of EEG channels used. Calculating and presenting all permutations of these factors would require an inordinate amount of time and space and would not yield much more information than we present here. Hence when we were not explicitly varying a particular factor we defaulted settings to 10 second window size, 8 channels, and 5 blocks worth of training data. Since we present results from classifications with different numbers of Difficulty levels, we should note that 5 blocks of training data is 20 minutes in a 2-way classification, 30 minutes in a 3-way classification, and 40 minutes in a 4-way classification. We highlight these parameters where it makes sense in all graphs of the within-task classification accuracies. Likewise, rather than presenting classification results for all 11 tests we ran, which included six pairwise classifications, four 3-way classifications, and the 4-way classification, we choose two 2-way, two 3-way, and the 4-way classification. Other tests we ran suggest that these numbers are representative of the space and that logical extensions can be made to draw legitimate conclusions.

Window Size. The first factor we vary is the window size, or the amount of temporal data that we transform into the frequency domain and treat as a single data instance. We chose window sizes of 2, 4, 10, 20, 30, 60, and 120 seconds, where 120 seconds is the length of the entire n-back sequence.

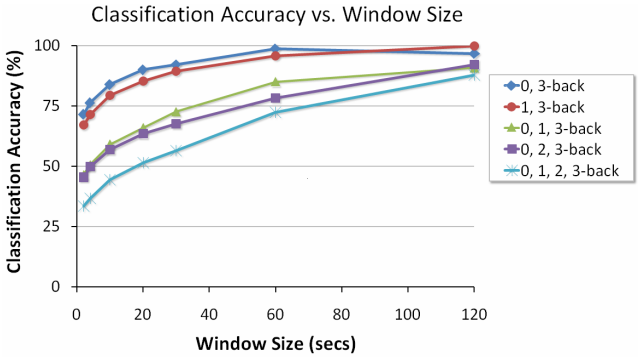

Figure 6: Classification accuracy increases as windows size (lag) increases. Interestingly, accuracies converge at very high levels with long enough windows.

First, as shown in Figure 6, we were able to replicate classification accuracies reported in previous work (e.g. our accuracies of 92.3% accuracy at 30 seconds is comparable to Gevins' 2-way accuracy of ~95% using 27.5 second windows [7]). The interesting thing about this is that because of the robustness of our feature selection process, we were able to attain these accuracies without performing any explicit artifact rejection. For example, Gevins et al. report using both automatic and manual artifact detection and throwing away nearly 20% of their data before training and classifying. We were also able to attain these results with far less training data than previously used. For the above data point, we use 20 minutes worth of training data compared to several days' worth as reported in previous work. As we will show, we could drastically reduce this amount of training data without affecting classification very much.

Second, we demonstrate the tradeoff that exists between window size and classification accuracy. This is important as the window sizes imply lag if the classification is operating in real time, as they might be in many HCI applications. As one might expect, the curves rise much more steeply as smaller window sizes increase to medium, and then seem to mostly plateau at the larger window sizes. The other interesting thing to note is how all the classifications, even the 3-way and 4-way ones, seem to converge at greater than 90% accuracies as the window sizes get larger. We are not aware of previous work that has shown accuracies for more than a 2-way classification.

Amount of Training Data. The second factor we vary is the amount of training data used to build the model. We picked levels of 1, 2, 3, 4, and 5 training blocks, using a 10 second window and 8 channels of EEG data. We added training blocks starting from the first one the user performed and progressing in time. We believe that this is a conservative estimate of accuracies given that the behavioral results suggest that participants were still learning the task in the first block or two. However, boosting our accuracy results by eliminating data or starting from the last block would be

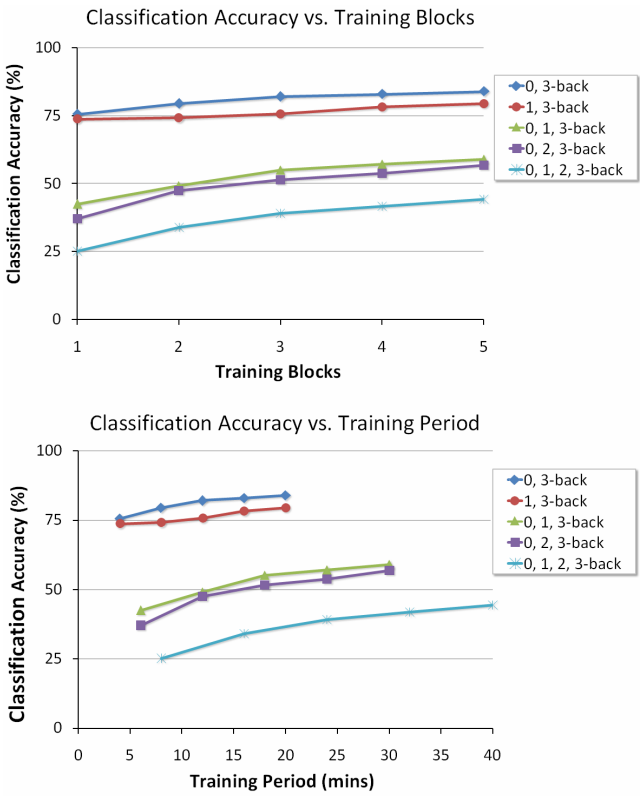

Figure 7: Classification accuracy increases marginally as more training data is added. This is seen both with number of blocks (top) as well as with raw training times (bottom).

artificial, possibly masking the fact that participants would still need to be trained with initial blocks before attaining that level of expertise.

Surprisingly, we see a much smaller reduction in accuracy than we had expected as we reduce the available training data. As seen in Figure 7 (top), with only 4 minutes worth of training data (58 training instances), we are able to attain 75.7% accuracy for the 0-back vs. 3-back classification, a drop of only about 9% when training with 5 times as much data. Figure 7 (bottom) shows this same classification accuracy data plotted against time rather than the number of blocks. These results suggest that our classification methodology allows us to attain relatively high accuracies with very little training data, a result that is extremely important if this is to be used in HCI settings.

Number of EEG Channels. The third factor we vary is the number of EEG channels used. We tested our classification scheme with 1, 2, 8, 16, and 32 channels of EEG data, corresponding to commonly used sets that propose uniform distribution balanced across hemispheres. For a list of channels used, see Table 1. Again somewhat surprisingly, while much previous work has utilized relatively high channel EEG systems with complex schemes to eliminate inter-channel noise (e.g. Laplacian spatial filtering), we found that using unfiltered data from 2-channels was nearly as good as using all 32 channels. See Figure 8 for a graph of these results. Various simple filtering schemes either did not improve classification much, and even sometimes hurt it. This implies that using our feature selection and classification schemes allow us to use small numbers of channels, and potentially inexpensive off-the-shelf EEG systems.

# Ch	Channels Used
1	Cz
2	Fz, Pz
8	2 ch + F3, F4, Cz, P3, P4, Oz
16	8 ch + F7, F8, T7, T8, C3, C4, P7, P8
32	All channels seen in Figure 3

Table 1: Channels selected for use in each analysis.

Cross-Task Results

While we have shown accurate classification results for training and testing on the same task, there exist scenarios in which we would like to train the system on one task and classify on another. For example, this would be useful since we could not train on the same task if we were trying to profile the user's load as they learn a brand new task. Training and classifying across tasks also allows us to evaluate the generality of the models that we are building.

In order to validate the cross-task classification we selected features and trained on the letter task and then used this model to classify data collected from the image and spatial tasks. To keep presentation brief, we report accuracies for the 0-back vs. 3-back classification using a 10 second window, 8 EEG channels, and 20 minutes of training data. As can be seen in Figure 9, the accuracy within the letter task does not fall significantly when applied to cross-task classification (means: 84.0% vs. 80.4% and 77.04% for within-task vs. testing with the image and spatial tasks, respectively). Further analysis shows that the average cross-task tradeoff curves look very similar to the within-task ones. For example, with 4 second windows, 4 minutes of training, and 2 EEG channels, the 0-back vs. 3-back accuracy falls to 62.5%, compared to 69.4% with 10 second windows, 12

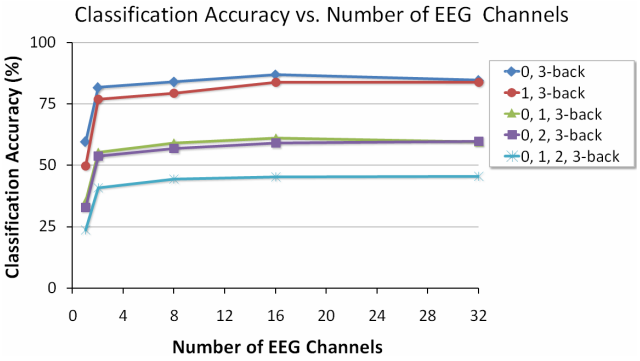

Figure 8: Classification accuracy increases drastically from up to 2 channels but seems to plateau with the addition of more channels.

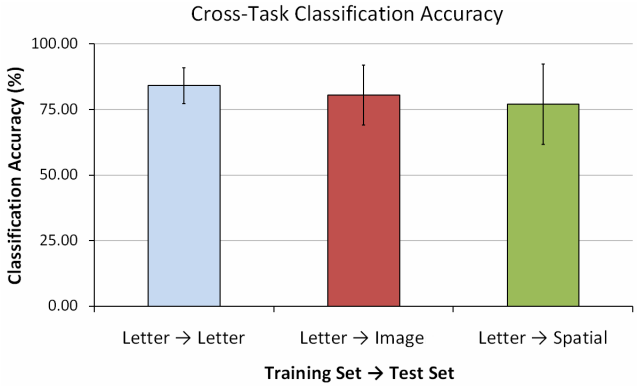

Figure 9: Classification accuracy remains relatively high even when training on one task variant (letter) and testing on the others (image or spatial).

minutes of training, and 8 EEG channels, or 70.1% when we increase that to 30 second windows.

Summary of Classification Results

We have presented results showing that we are able to attain relatively high classification accuracies for memory load using our feature selection and classification scheme. Within this scheme, we have also shown the tradeoffs that exist between accuracy and window size, which translates into classification lag when using this with real-time HCI applications. Furthermore, results show that our accuracies are not significantly reduced when we reduce the amount of training data or the number of EEG channels, suggesting that our classification methodology could lead to lower equipment and time overhead than previously articulated. Additionally, we have shown that our models transfer reasonably well to different variants of the n-back task. These results, when combined, bring us closer to being able to utilize EEG technologies in HCI work.

DISCUSSION AND FUTURE WORK

In this work, we have presented a methodology that yielded high classification accuracy for working memory load, both within task as well as across related tasks. We specifically tailored our methodology as well as our experiment and validation so that the results represent the potential for using these techniques in HCI research. However, this is just a first step and much future work remains.

In the experiment, we have assumed that increasing the number of items a user had to remember uniformly increased the working memory load for the entire period. This provided 'ground truth' for our memory load measurements and allowed us to cleanly label the data and test the classifier. However, while this might be true on average, it is almost certainly not true in all instances, especially when the temporal windows used were relatively small. In some of those cases, we may actually be classifying the load correctly even though it may not match the label we have assigned based on the average task difficulty. In future work, we aim to validate the temporal resolution that we can attain with our classification. We hope that this validation would alleviate the need for collecting ground truth in real applications and provide a measure that would allow us to continuously monitor how working memory load varies, even within a single task.

The stimuli used in the task variants for cross-task validation were fairly different perceptually, and possibly cognitively. However, the task structure was the same across all of them. While it is encouraging that we are able to classify across these similar tasks, a significant amount of work remains to explore how much task variation we can accommodate. We would also eventually like to develop a set of canonical classification tasks that researchers can use for various cognitive measures.

Cursory explorations into cross-user classification indicate that a naïve implementation in which we trained on several users and tried to apply the model to a new user did not work very well. Given the individual differences we saw across users, this is not altogether surprising. While we would have liked to have found a set of features that could robustly be used across users, analysis on features that were selected by our system suggest that these in fact differed quite drastically. However, our feature selection process provides us with the ability to evaluate entirely new features that may generalize across users. Another approach that may yield interesting results is one that first tries to cluster people who exhibit similar characteristics in their signal and then apply different models to classify their load. Exploring both of these approaches remains future work.

CONCLUSION

In this paper, we have described our EEG classification methodology, including a novel feature selection scheme that maximizes information gain while minimizing the effects of drift. This seems to eliminate the need to build complex models and understanding of any drift that exists in the data. We have also described an experiment exploring the use of an EEG to classify working memory load. In this experiment, we show that we are able to attain high memory load accuracies for 2-way, 3-way, and 4-way classifications. We also demonstrate the tradeoffs that exist between the accuracies and temporal window size, amount of training data, and number of EEG channels. Results suggest that we can attain relatively high classification accuracies even with shorter windows, less training data, and a smaller number of channels than previously reported. This is encouraging and is a step towards using these technologies in HCI environments, due to a reduction in costs and complexity of application and analysis. Finally we show that the models we build transfer across variants of the task, again providing encouraging evidence of the generality of our techniques.

ACKNOWLEDGMENTS

We thank Ed Cutrell, Mary Czerwinski, Brent Field, Eric Horvitz, Dan Olsen, John Platt, Greg Smith, and Dan Weld for their support and discussions of this work.

REFERENCES

1. Baddeley, A.D., & Hitch, G. (1974). Working Memory. In G.H Bower (ed.) *Recent Advances in Learning and Motivation, 8*. Academic Press: New York.
2. Berger, H. (1929). Uber das elektroenzephalogramm des menschen. Arch Psychiatr, 87, 527-570.
3. Card, S., Moran, T., & Newell, A. (1983). *The Psychology of Human-Computer Interaction*. Lawrence Erlbaum Associates: Hillsdale, NJ.
4. Chen, D., & Vertegaal, R. (2004). Using mental load for managing interruptions in physiologically attentive user interfaces. *Extended Abstracts of SIGCHI 2004 Conference on Human Factors in Computing Systems*, 1513-1516.
5. Fisch, B.J. (2005). *Fisch & Spehlmann's EEG primer: Basic principles of digital and analog EEG*. Elsevier; Amsterdam.
6. Gevins, A., & Smith, M.E. (2003). Neurophysiological measures of cognitive workload during human-computer interaction. *Theoretical Issues in Ergonomics Science, 4*, 113-131.
7. Gevins, A., Smith, M.E., Leong, H., & McEvoy, L. (1998). Monitoring working memory load during computer-based tasks with EEG pattern recognition methods. *Human Factors, 40(1)*, 79-91.
8. Howard, M.W., Rizzuto, D.S., Caplan, J.B., Madsen, J.R., Lisman, J., Aschenbrenner-Scheibe, R., Schulze-Bonhage, A., & Kahana, M.J. (2003). Gamma oscillations correlate with working memory load in humans. *Cerebral Cortex, 13*, 1369-1374.
9. Jasper, H.H. (1958). The ten-twenty electrode system of the international federation. *Electroencephalography and Clinical Neurophysiology, 10*, 371-375.
10. Jensen, O., Gelfand, J., Kounios, J., & Lisman, J.E. (2002). Oscillations in the alpha band (9-12 Hz) increase with memory load during retention in a short-term memory task. *Cerebral Cortex, 12*, 877-882
11. Kilmesch, W., Schimke, H., & Pfurtscheller, G. (1993). Alpha frequency, cognitive load and memory performance. *Brain Topography, 5(3)*, 241-251.
12. Kitamura, Y., Yamaguchi, Y., Imamizu, H., Kishino, F., & Kawato, M. (2003). Things happening in the brain while humans learn to use new tools. *Proceedings of SIGCHI 2003 Conference on Human Factors in Computing Systems*, 417-424.
13. Kohlmorgen, J., Muller, K.R., & Pawelzik, K. (1998). Analysis of drafting dynamics with neural network hidden markov models. *Conference on Advances in Neural Information Processing Systems*, 735-741.
14. Kramer, A.F. (1991). Physiological metrics of mental workload: A review of recent progress. In *Multiple Task Performance* (ed. Damos, D.L.), 279-328.
15. Lee, J., & Tan, D.S. (2006). Using low-cost electroencephalograph for task classification in HCI research. *Proceedings of the 19th ACM Symposium on User Interface Software and Technology*, 81-90.
16. MacKay, D.J.C. (2002). *Information Theory, Inference, & Learning Algorithms.* Cambridge University Press: USA.
17. Mecklinger, A., Kramer, A.F., & Strayer, D.L. (1992). Event related potentials and EEG components in a semantic memory search task. *Psychophysiology, 29*, 104-119.
18. Onton, J., Delorme, A., & Makeig, S. (2005). Frontal midline dynamics during working memory. *Neuroimage, 27*, 341-356.
19. Owen, A.M., McMillan, K.M., Laird, A.R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping, 25*, 46-59.
20. Palaniappan, R. (2005). Brain computer interface design using band powers extracted during mental tasks. *Proceedings of the 2nd International IEEE EMBS Conference on Neural Engineering*, 321-324.
21. Picard, R. (2000). *Affective Computing*. MIT Press: Cambridge.
22. Picton, T.W., Bentin, P., Berg, P., Hillyard, S.A., Johnson, J.R., Miller, G.A., et al. (2000). Guidelines for using human event-related potentials to study cognition: Recording standards and publication criteria. *Psychophysiology, 37*, 127-152.
23. Piolat, A., Olive, T., Roussey, J., Thunin, O., & Ziegler, J.C. (1999). SCRIPTKELL: A tool for measuring cognitive effort and time processing in writing and other complex cognitive activities. *Behavior Research Methods, Instruments, & Computers, 31(1)*, 113-121.
24. Pleydell-Pearce, C.W., Whitecross, S.E., & Dickson, B.T. (2003). Multivariate analysis of EEG: Predicting cognition on the basis of frequency decomposition, inter-electrode correlation, coherence, cross phase and cross power. *Hawaii International Conference on System Sciences*, 11-20.
25. Prinzel, L.J., Pope, A.T., Freeman, F.G., Scerbo, M.W., & Mikulka, P.J. (2001). Empirical analysis of EEG and ERPs for psychophysiological adaptive task allocation. *NASA Technical Report TM-2001-211016*.
26. Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object set: The role of surface detail in basic-level object recognition. *Perception, 33*, 217-236.
27. Smith, M.E., Gevins, A., Brown, H., Karnik, A., & Du, R. (2001). Monitoring task loading with multivariate EEG measures during complex forms of human-computer interaction. *Human Factors, 43(3)*, 366-380.
28. Velichkovsky, B., & Hansen, J.P. (1996). New technological windows into mind: There is more in eyes and brains for human-computer interaction. *Proceedings of the SIGCHI 1996 Conference on Human Factors in Computing Systems*, 496-503.