Fall 2007

Supporting Real-Time Cognitive State Classification on a Mobile Individual

Michael C. Dorneich
Stephen D. Whitlow
Santosh Mathan
Patricia May Ververs
Deniz Erdogmus, *Oregon Health & Science University*, et al.

Supporting Real-Time Cognitive State Classification on a Mobile Individual

Michael C. Dorneich
Stephen D. Whitlow
Santosh Mathan
Patricia May Ververs
Honeywell Laboratories
Deniz Erdogmus
Andre Adami
Misha Pavel
Tian Lan
Oregon Health & Science University

ABSTRACT: The effectiveness of neurophysiologically triggered adaptive systems hinges on reliable and effective signal processing and cognitive state classification. Although this presents a difficult technical challenge in any context, these concerns are particularly pronounced in a system designed for mobile contexts. This paper describes a neurophysiologically derived cognitive state classification approach designed for ambulatory task contexts. We highlight signal processing and classification components that render the electroencephalogram (EEG) -based cognitive state estimation system robust to noise. Field assessments show classification performance that exceeds 70% for all participants in a context that many have regarded as intractable for cognitive state classification using EEG.

Introduction

ADAPTIVE AUTOMATION, IN WHICH THE AUTOMATION ADAPTS DURING EXECUTION TO THE current task environment, can either provide adaptive aiding, which makes a certain component of a task simpler, or can provide adaptive task allocation, which shifts an entire task from a larger multitask context to automation (Parasuraman, Mouloua, & Hilburn, 1999). Adaptive systems must make timely decisions on how to use varying levels of (adaptive) automation to provide support in a joint human-automation system.

In order for an adaptive system to decide when to intervene, it must have some model of the context of operations, be it a functional model of system performance or possibly a model of the operator's functional state. Currently, many adaptive systems derive their inferences about the cognitive state of the operator from mental models, performance on the task, or external factors related directly to the task environment (Wickens & Hollands, 2000). For example, Scott (1999) developed the Ground Collision-Avoidance System (GCAS) for test on an F-16D. GCAS used the projected

---

ADDRESS CORRESPONDENCE TO: Michael C. Dorneich, Honeywell Laboratories, Human-Centered Systems, 3660 Technology Dr., Minneapolis, MN 55418, 612/951-7488, michael.dorneich@honeywell.com.

Journal of Cognitive Engineering and Decision Making, Volume 1, Number 3, Fall 2007, pp. 240–270. DOI 10.1518/155534307X255618.

time until an aircraft broke through a pilot-determined minimum altitude as an external condition to infer that a pilot's attention was incapacitated, at which point the system would perform a "fly-up" evasive maneuver to avoid a ground collision. In that case, the automation took over control of the aircraft from the pilot.

Neurophysiologically and physiologically triggered adaptive automation offers many advantages over the more traditional approaches to automation by basing estimates of operator state in sensed data directly. These systems offer the promise of leveraging the strengths of humans and machines, augmenting human performance with automation specifically when assessed human cognitive capacity falls short of the demands imposed by task environments. With more refined estimates of the operator's cognitive state, measured in real time, adaptive automation also offers the opportunity to provide aid even before the operator knows he or she is getting into trouble.

Operational Problem

The aim of augmented cognition research is to use physiological and neurophysiological sensors to detect states in which cognitive resources may be inadequate to cope with mission-relevant demands. The goal is to enhance human performance when task-related demands surpass the human's assessed current cognitive capacity, which fluctuates subject to fatigue, stress, overload, or boredom. Efforts have focused on ways to leverage cognitive state information to drive adaptive systems to manage information flow when detected human cognitive resources may be inadequate for the tasks.

The Honeywell team has focused on the dismounted soldier in the future military. The research program described in this article was conducted in support of the U.S. Army Future Force Warrior (FFW) Advanced Technology Development program. The FFW program seeks to push information exchange requirements to the lowest levels, with the goal of enhancing the capabilities of a squad so that it can cover the battlefield in the same way that a platoon now does. A critical element of the FFW program is a reliance on networked communications and high-density information exchange. These capabilities are expected to increase situation awareness at every level of the operational hierarchy. Introducing information technologies within the transformation of the military will facilitate better individual and collaborative decision making at every level. However, effective use of these information sources is constrained by the limitations of the human cognitive system.

This revolutionary concept of operations could dramatically increase the likelihood of information overload that could turn the postulated information superiority into a profound liability. The potential data overload, coupled with the efficiency of information flow required in executing Army doctrine, places an overreliance of critical information throughput on a single point of contact, the individual warfighter. To ensure that warfighters are supported appropriately, there needs to be intelligent information management to ensure that the system can support superior situation awareness on the battlefield.

Adaptive information management systems have an important role in this context. The efficacy of such a system is contingent on reliable and timely cognitive assessment.

An example instantiation of such a system is the Communications Scheduler as described in Dorneich, Whitlow, Mathan, Carciofini, and Ververs (2005). The system changes the information presentation (e.g., high-priority messages preceded by a priming alert, low-priority messages delivered via text messages) based on the message priority and cognitive workload of the soldier during critical times. The system not only reduces the overall number of transmissions at key moments but also improves the likelihood of receipt of essential information with reduced bandwidth and power usage. But such strong mitigations of an adaptive system can be effective only if they are properly tuned to the current cognitive capabilities of the user, as well as thoroughly evaluated with the anticipated users of the system. An accurate real-time classification of the cognitive state of the soldier is an essential first step in this process.

Neurophysiologically driven prototypes for regulating information flow were developed and tested by a team of researchers led by Honeywell (Dorneich, Whitlow, Mathan, Ververs, Pavel, & Erdogmus, 2005; Dorneich, Whitlow, Ververs, Mathan, et al., 2004) to evaluate the potential benefits to the ground soldier who will receive volumes of information from a variety of sensors and sources. Information regarding a soldier's cognitive state was integrated with information systems to manage assets and communications. Cognitive state classification was applied to and focused on those Army roles that require significant cognitive processing, information integration, and information management on the part of the recipient. Such roles include the battlefield commander, robotics noncommissioned officer, platoon leader, and other roles that support the Network Centric Information Environment.

The current FFW approach to cognitive state assessment relies on cardiac and physical sensors to assess general cognitive state based on the level of sleep debt in the last 24 hours and the phase of the circadian cycle (Institute of Medicine of the National Academies, 2004). If a truly adaptive system that manages information flow is to be implemented, a higher degree of fidelity in the cognitive state assessment and temporal resolution is needed.

Cognitive State Classification Techniques

Neurophysiological- and physiological-based assessment of cognitive state has been captured in several different ways, including but not limited to cardiac measures, electroencephalogram (EEG), and functional near-infrared (fNIR) imaging. There is an extensive research history of using cardiac, or electrocardiogram (ECG), measures to evaluate cognitive activity under a variety of task conditions. Measures include heart rate variability in the time domain to assess mental load (Kalsbeek & Ettema, 1963), tonic heart rate to evaluate the impact of continuous information processing (Wildervanck, Mulder, & Michon, 1978), variability in the spectral domain as an index of cognitive workload (Wilson & Eggemeier, 1991), and T-wave amplitude during math interruption task performance (Heslegrave & Furedy, 1979). The fNIR spectroscopy conducts functional brain studies using wavelengths of light introduced at the scalp to measure cognition-related hemodynamic changes, and has been used to assess the cognitive state (Izzetoglu & Bunce, 2004).

Other physiological measures used to inform cognitive state assessment are galvanic skin response (Verwey & Veltman, 1996), eyelid movement (Neumann, 2002; Stern, Boyer, & Schroeder, 1994; Veltman & Gaillard, 1998; Yamada, 1998), pupil response (Beatty, 1982; Partala & Surakka, 2003), and respiratory patterns (Backs & Seljos, 1994; Boiten, 1998; Porges & Byrne, 1992; Veltman & Gaillard, 1998; Wientjes, 1992).

As the gold standard for providing high-resolution spatial and temporal indices of cortical electrical activity from scalp electrodes, EEG has been used in the context of adaptive systems. For instance, researchers have used the engagement index, developed by NASA, in the context of mixed-initiative control of an automated system (Pope, Bogart, & Bartolome, 1995). This method uses a ratio of power in common frequency bands (beta / [alpha + theta]), whereby cognitively alert and focused is represented in beta, wakeful and relaxed in alpha, and a daydream state in theta. Thus higher engagement index values estimate increased levels of task engagement.

The efficacy of the engagement index as the basis for adaptive task allocation has been experimentally established. For instance, under manipulations of vigilance levels (Mikulka, Hadley, Freeman, & Scerbo, 1999) and workload (Prinzel, Freeman, Scerbo, Mikulka, & Pope, 2000), an adaptive system effectively detected states in which human performance was likely to fail and took steps to allocate tasks in a manner that would raise overall task performance. The results associated with the engagement index highlighted the potential benefits of a neurophysiologically triggered adaptive automation. There are several ways in which this promising work needs to be extended in order to be effective in the dynamic, ambulatory contexts of the research reported here.

1. *Individual differences*. As Scerbo and Gustafson (2001) point out, there were unique individual EEG responses to task demands. Although the characterization of the relationship between engagement and EEG activity, in terms of activity within certain frequency bands and sites, was useful for synthesizing broadly observed trends, a given individual's responses may deviate substantially from assumptions derived from averaged data. In response, some researchers have called for an approach that was more sensitive to individual variability in EEG expression (Mathan et al., 2005).

2. *Linear relationships*. The engagement index was based on a linear relationship between power estimates at specific frequency bands. However, there are potentially informative nonlinear relationships across spectral features at various sites that could help discriminate between various cognitive states. Research indicates that more advanced pattern recognition techniques, such as multilayer neural networks, could exploit relationships among features that do not conform to linearity assumptions (Scerbo et al., 2001; Wilson & Russell, 2003).

3. *Analysis windows*. The engagement index was designed to estimate cognitive state over an analysis window that was close to a minute in duration. Developers of the engagement index made no claims about its efficacy at temporal resolutions of a few seconds or hundreds of milliseconds. In the authors' own laboratory experience, the engagement index was able to discriminate between periods of high-intensity virtual combat and periods of rest in a first-person video game over the course of analysis

windows that spanned minutes, but not at a resolution of less than 10 s (Dorneich, Whitlow, Ververs, Mathan, et al., 2004). The demands of the task environment may require techniques that provide reliable cognitive state estimates with a fairly high degree of temporal resolution.

4. *Validation context.* Much of the literature associated with cognitive state estimation relies on findings from data collected in relatively stationary laboratory settings (Schmorrow & Kruse, 2002). Data collection in laboratory environments has several attributes that cannot be realized in mobile contexts. For example, the experimental setup can be controlled in order to facilitate better performance, various precautions to improve signal quality can be implemented, and large-scale data collection, analysis, and signal-processing hardware and software can be used. These constraints have to be relaxed in mobile environments. In mobile applications, EEG signals can be very noisy and contaminated by a wide range of artifacts. Furthermore, the system must be portable and able to work in real time.

The work reported here addressed some of the shortcomings that were previously highlighted by creating a system that was optimized to the unique EEG spectral characteristics of each individual in response to specific task demands. Pattern recognition techniques that make no restrictive assumptions about the form of the data being modeled were used. The system provided cognitive state estimates at a high degree of temporal resolution and was designed to work in real time in mobile contexts. Three aspects of the approach are highlighted in the pages that follow: hardware integration into a wireless wearable form factor, real-time signal processing to detect and correct for artifacts, and a nonlinear classification approach.

The remainder of this paper is organized as follows. The next section will discuss the technical challenges in creating and evaluating robust mobile EEG classification. Preliminary laboratory experiments that formed the foundations of the work discussed in this paper will be briefly reviewed. Finally, the mobile field evaluation and results will be discussed in detail, concluding with a discussion of future directions.

Technical Challenges

Realizing the vision of an augmented cognition system in the context of an ambulatory soldier has been constrained by several challenges. First, as Schmorrow and Kruse (2002) noted, processing and analysis of neurophysiological data have been largely conducted offline by researchers and practitioners. However, in order for augmented cognition technologies to work in practical settings, effective and computationally efficient artifact reduction and signal-processing solutions are necessary. Second, inferring the cognitive state of users demands pattern recognition solutions that are robust both to noise and to the inherent nonstationarity in neurophysiological signals (Popivanov & Mineva, 1999). Third, understanding the fluctuations of the cognitive state in applied environments requires the development of means to collect reliable neurophysiological data outside the laboratory. Fourth, experiments must be designed, often under conflicting constraints (e.g., operational realistic tasks vs. well-understood, controlled laboratory tasks), to effectively evaluate classification accuracy. Finally, compact and robust form factors (e.g., size, weight, ruggedness) associated with neurophysiological sensors and processors are a matter of critical concern.

Real-Time Signal-Processing Challenges

Conducting military maneuvers in operational environments such as urban terrain often does not allow an individual to remain stationary and can demand simultaneous cognitive and physical activity. Consequently, difficulties related to the processing of EEG signals in real-world settings include factors associated with both participant motion and the operational environment itself. Thus, utilization of research methods involving EEG in operational environments necessitates the use of real-time algorithms for signal detection and removal of artifacts. Although real-time signal processing and classification of the EEG has been implemented previously (Berka et al., 2004; Gevins & Smith, 2003), it has not been realized in a truly mobile, ambulatory environment.

Inferring cognitive state from noninvasive neurophysiological sensors is a challenging task even in pristine laboratory environments. High-amplitude artifacts, ranging from eye blinks to muscle artifacts and electrical line noise, can easily mask the lower-amplitude electrical signals associated with cognitive functions. These concerns are particularly pronounced in the context of ongoing efforts to realize neurophysiologically driven adaptive automation for the dismounted ambulatory soldier.

In addition to the typical sources of signal contamination, mobile applications must consider the effects of artifacts induced by shock, cable movement, and gross muscle movement. Specifically, artifacts related to participant motion include high-frequency muscle activity, verbal communication, and ocular artifacts consisting of eye movements and blinks; whereas artifacts related to the operational environment include instrumental artifacts such as electrical noise that creates interference with the EEG signal (cf. Kramer, 1991).

Classification Challenges

The use of EEG as the basis for cognitive state assessment was motivated by characteristics such as good temporal resolution, low invasiveness, low cost, and portability. Although EEG offers several benefits, there are shortcomings related to the noise artifacts described previously and the nonstationarity of the neural signal pattern over time. Despite these challenges, research has shown that EEG activity can be used to assess a variety of cognitive states that affect complex task performance. These include working memory (Gevins & Smith, 2000), alertness (Makeig & Jung, 1995), executive control (Garavan, Ross, Li, & Stein, 2000), and visual information processing (Thorpe, Fize, & Marlot, 1996). These findings point to the potential for using EEG measurements as the basis for driving adaptive systems that demonstrate a high degree of sensitivity and adaptability to human operators in complex task environments.

Scenario Design Challenges

In addition to the practical and system configuration challenges faced when moving from the laboratory to field studies, there are issues of experimental control and the characterization of cognitive state in less constrained environments. It is essential to select tasks that are both operationally relevant and afford reasonable adaptations that improve performance. In the laboratory, it is possible to develop simple tasks in which workload is manipulated precisely and consistently. Additionally, a user's performance can be collected and evaluated accurately. This makes it relatively easy to establish ground truth about a user's likely workload.

However, when developing operationally relevant tasks in a field environment, it becomes substantially more difficult to manipulate workload precisely and to interpret and assess a user's performance without compromising operational realism. The mobile field evaluation reported herein had two objectives: first, to determine whether an operationally relevant task load manipulation had a measurable impact on a user's workload; and second, to establish whether a sensor-based classification approach could effectively classify a user's workload in a mobile setting.

System Description

This section describes the mobile classification hardware and software approaches. Subsequent sections will describe how this system was evaluated in a mobile setting.

Hardware

The wireless sensor suite employed by Honeywell was assembled using a variety of off-the-shelf hardware components tied together with a custom agent-based information architecture based on the work of the Institute for Human and Machine Cognition (IHMC; see Dorneich, Whitlow, Ververs, Mathan et al., 2004, for more information). EEG data were collected with a 32-channel BioSemi Active Two system as well as a more deployable six-channel EEG sensor headset made by Advanced Brain Monitoring (ABM). The BioSemi Active Two system integrates an amplifier with an Ag-AgCl electrode, which affords extremely low noise measurements without skin preparation. The ABM system includes two differential channels (FzPOz and CzPOz) and four referential channels (Fz, Cz, POz, and linked mastoids acting as a reference site).

Information from either of the EEG systems was processed on a body-worn laptop that ran the IHMC information architecture. The BioSemi and ABM systems interfaced with the laptop via a USB 2.0 port and Bluetooth® serial port, respectively. The sensor electronics and the laptop were mounted in a backpack worn by the participant (see Figure 1). Sensor data were collected and processed on the laptop computer during the experiment.

Signal Processing

For the BioSemi Active Two EEG system, vertical and horizontal eye movements and blinks were recorded with electrodes below and lateral to the left eye. All channels referenced the right mastoid. EEG was sampled at 256Hz from 7 channels (CZ, P3, P4, PZ, O2, P04, F7), which were selected based on a saliency analysis on EEG

Figure 1. *Body-worn sensor suite and signal-processing system.*

that was collected from various participants performing cognitive test battery tasks (Russell & Gustafson, 2001). EEG signals were preprocessed to remove eye blinks using an adaptive linear filter that is based on the Widrow-Hoff training rule (Widrow & Hoff, 1960). Information from the VEOGLB (electrode that measures vertical eye activity) ocular reference channel was used as the noise reference source for the adaptive ocular filter. DC drifts were removed using high pass filters (0.5 Hz cutoff). A band pass filter (between 2 Hz and 50 Hz) was also employed, as this interval was generally associated with cognitive activity.

The power spectral density (PSD) of the EEG signals was estimated using the Welch method (Welch, 1967). The PSD process used 1-s sliding windows with 50% overlap. PSD estimates were integrated over five frequency bands: 4–8 Hz (theta), 8–12 Hz (alpha), 12–16 Hz (low beta), 16–30 Hz (high beta), and 30–44 Hz (gamma). The classifier received a PSD feature vector of the five bands as input every 100 ms. The particular selection of the frequency bands was based on well-established interpretations of EEG signals in prior cognitive and clinical contexts (e.g., Gevins, Smith, McEvoy, & Yu, 1997).

The ABM system supported an independent signal-processing stream. Six channels were sampled at 256 samples/s with a bandpass from 0.5 Hz and 65 Hz (at 3 dB attenuation) that was obtained digitally with Sigma-Delta A/D converters. Data were transmitted across a Bluetooth RF (radio) link to the collection laptop via an RS232 interface. Quantification of the EEG in real time was achieved using signal analysis techniques that identified and decontaminated eye blinks and identified and rejected data points that were contaminated with electromyographic (EMG) signals, amplifier saturation, and/or excursions attributable to movement artifacts (see Berka et al., 2004, for a detailed description of the artifact decontamination procedures).

Decontaminated EEG was then segmented into overlapping 256 data point windows called *overlays*. An epoch (the temporal window of analysis) consisted of three consecutive overlays. Fast-Fourier Transform (FFT) was applied to each overlay of the decontaminated EEG signal multiplied by the Kaiser window ($\alpha = 6.0$) to compute the power spectral densities (PSD). The PSD values were adjusted to take into account zero values inserted for artifact-contaminated data points. The PSD between 70 and 128 Hz was used to detect EMG artifact. Overlays with excessive EMG artifacts or with fewer than 128 data points were rejected.

The remaining overlays were then averaged to derive PSD for each epoch with a 50% overlapping window. Epochs with two or more overlays with EMG or missing data were classified as invalid. For each channel, PSD values were derived for each 1-Hz bin from 3 Hz to 40 Hz and the total PSD from 3 to 40 Hz. Relative power variables were also computed for each channel and bin using the formula (total band power/total bin power).

Real-Time Classification

Estimates of spectral power formed the input features to a pattern classification system. The classification system used parametric and nonparametric techniques to assess the likely cognitive state on the basis of spectral features; that is, estimate *p(cognitive state | spectral features)*. The classification process relied on probability density estimates derived from a set of spectral samples. These spectral samples were gathered in conjunction with tasks that were as close as possible to the eventual task environment.

The classification system (Figure 2) used a fusion of three distinct classification approaches: K-nearest neighbor (KNN), Parzen windows, and Gaussian mixture models (GMM).

Gaussian mixture models. Gaussian mixture models provided a way to model the probability density functions of spectral features that were associated with each cognitive state. This was accomplished using a superposition of Gaussian kernels. The unknown probability density associated with each class or cognitive state was approximated by a weighted linear combination of Gaussian density components. Given an appropriate number of Gaussian components and appropriately chosen component parameters (mean and covariance matrix associated with each component), a Gaussian mixture model can model any probability density to an arbitrary degree of precision.

The parameters associated with component Gaussians were iteratively determined using the expectation maximization algorithm (Dempster, Laird, & Rubin, 1977). Once the Gaussian parameters were initialized, the system iterated through a two-step procedure for each sample that was associated with each class. In the first step (expectation step), the system computed the probability of a particular training sample belonging to a particular class based on current model parameters (posteriori

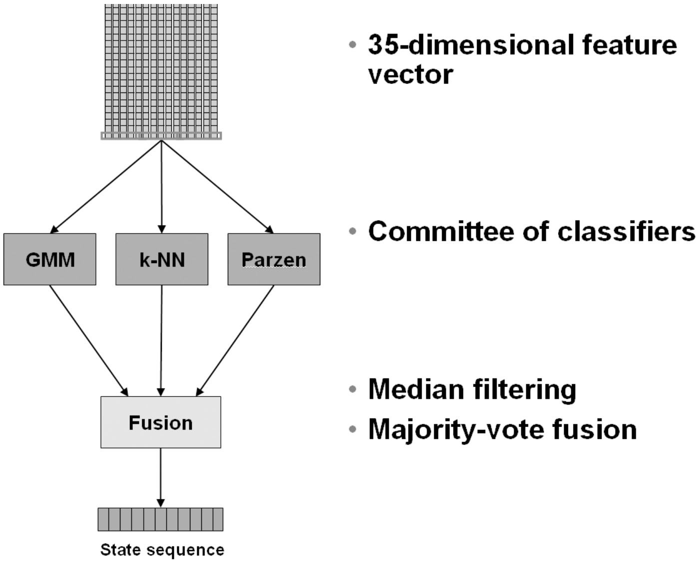

Figure 2. Classification system.

probability). In the maximization step, the model parameters were adjusted in the direction of increasing the class membership likelihood.

Once probability density functions associated with each cognitive state were generated, it became possible to classify individual spectral samples. Each spectral vector was attributed to a class that had the highest posterior probability of representing it. Posterior probabilities were computed using Bayes' rule. For example, Figure 3 shows the probability density functions associated with three distinct classes (i.e.,

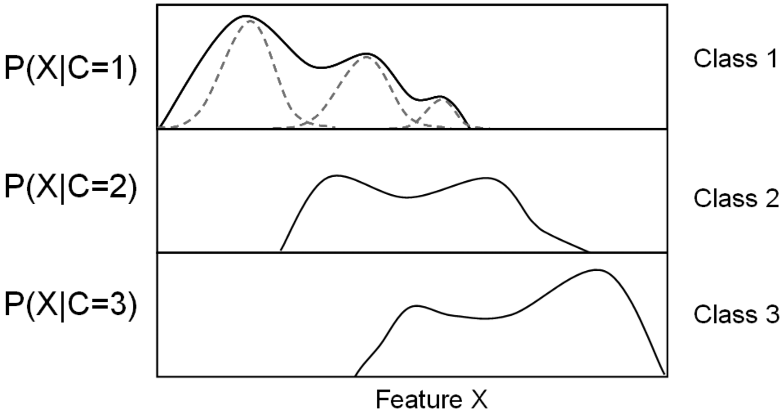

Figure 3. Gaussian mixture models. Small numbers of Gaussian kernels (dotted lines) are used to approximate the distribution of features in each class.

cognitive states). These probability densities are estimated using three Gaussians. Very high values of the data point x are most likely to have come from Class 3, whereas very low values of x are most likely to have come from Class 1.

K-nearest neighbor. The K-nearest neighbor approach is a nonparametric technique that makes no assumption about the form of the probability densities underlying a particular set of data. Given a particular sample x, the classification process identifies k samples whose features come closest (as assessed by Euclidian or Mahalanobis distance metrics) to the features represented in x. The sample x is assigned the modal class of the nearest k neighbors. For example, consider the data point represented by the question mark in Figure 4. Based on $k = 5$, it would be assigned the label associated with the most common class category of its five nearest neighbors.

Parzen windows. Parzen windows (Parzen, 1967) are a generalization of the k-nearest neighbor technique. Instead of choosing the nearest neighbors and assigning a sample x with the label associated with the modal class of its neighbors, each vote is weighed by using a kernel function. With Gaussian kernels, the weight decreases exponentially with the square of the distance. As a consequence, far-away points become insignificant. Kernel volumes constrain the region within which neighbors are considered. Consequently, Parzen windows are a better choice when there are large differences in the variability associated with each class. The data point shown in Figure 5 is assigned to the dominant class in its immediate vicinity.

Composite classifier. These statistical classification techniques were chosen over multilayer neural networks because they required minimal training time. KNN and Parzen windows required no training, whereas the expectation-maximization algorithm used to generate GMMs converged relatively quickly. KNN and Parzen window approaches required all training patterns to be held in memory. Every new feature

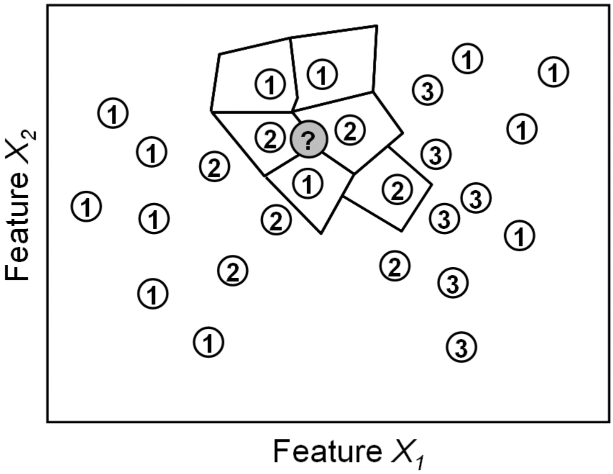

Figure 4. K-nearest neighbor. A given feature vector is assigned the class label associated with the modal class of the n *samples that are the most similar to it.*

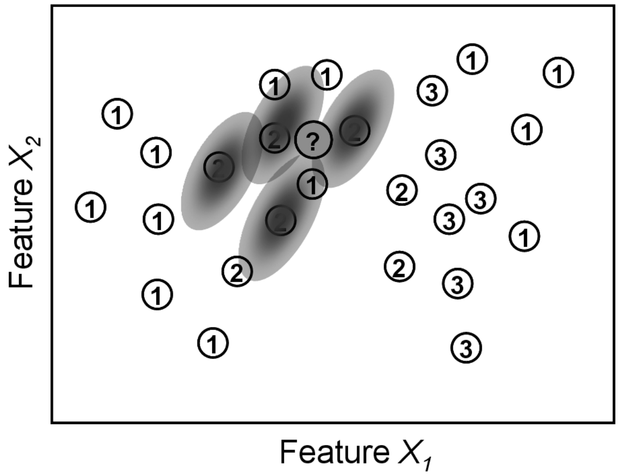

Figure 5. *Parzen windows. Gaussian kernels placed over each data point are used to estimate the distribution of features in each class.*

vector had to be compared with each of these patterns. However, despite the computational cost of these comparisons at run time, the system was able to output classification decisions well within real-time constraints.

The composite classification system regarded the output from each classifier as a vote for the likely cognitive state. The majority vote of the three component classifiers formed the output of the composite classifier. Fusing the outputs of multiple classifiers using a voting scheme is a widely used strategy to increase the robustness of the classification system. The equal weighting of different classifiers implicit in the voting scheme reflected the fact that no single classifier produced consistently superior results across subjects and tasks in pilot experiments. Although simple vote-based fusion improves the overall performance of classification systems (Kittler, Hatef, Duin, & Matas, 1998), there are a variety of alternative options for combining diverse classifiers. Exploring these options will be an objective of future research.

A classification decision was output at a rate of 10 Hz. Outputs from the composite classifier were passed through a modal filter before an assessment of cognitive state was output by the classification system. Modal filtering served to make the cognitive state assessment process more robust to undesirable fluctuations in the underlying EEG signal. Modal filtering was done over a sliding two-second window with the assumption that cognitive state remains stable over that period of time.

Laboratory Evaluation

This section briefly discusses one classification validation experiment conducted in a laboratory setting before moving on to the focus of this paper – mobile field evaluation. The laboratory evaluation described here is representative of the multiple preliminary experiments conducted to validate the approach described in the previous section. For a more detailed discussion of the previous work that provided the foundation of the mobile classification field evaluation, see Dorneich, Whitlow, Ververs, Carciofini, and Creaser (2004); Erdogmus, Adami, Pavel, Lan, Mathan, Whitlow, et al. (2005); Lan, Erdogmus, Adami, Pavel, and Mathan (2005); and Mathan et al. (2005).

Objective

The objective of this experiment was to validate the classification approach using a well-understood laboratory task, the *n*-back task, that has been used to manipulate working memory demands. In addition, two different EEG detection systems were evaluated.

Participants

Data were collected from five participants. All were male researchers at Honeywell.

Apparatus

EEG data were collected with a 32-channel BioSemi Active Two system as well as a more deployable six-channel ABM EEG sensor (see System Description section for details). Three participants wore the BioSemi system and two participants wore the ABM system.

Tasks

The working memory assessment was conducted using the *n*-back task. The *n*-back task required participants to process a sequence of letters presented on a computer screen. With every presentation of a letter, a participant had to both encode the letter in memory and indicate whether the letter corresponded to a letter shown *n* presentations ago. Working memory load encountered by a participant was manipulated by manipulating the value of *n*.

Procedure

Participants were seated and performed the task twice under 1-back and 2-back conditions. Data associated with the first performance under the two conditions were used to train the classifiers. The classifier was tested with data from the second performance under each working memory condition. The features used for classification consisted of estimates of spectral power at theta, alpha, beta, and gamma frequency at each EEG site.

Data Analysis and Results

The accuracy metric used in our evaluations was derived from a *confusion matrix*. The confusion matrix is a square matrix that allows comparison of the accuracy of a classifier by comparing the predicted class membership against actual membership (see Figure 6). Typically, rows represent the actual class, whereas columns represent the predicted class. Counts in each cell provide an indication of how well the classifier performed in classifying each sample in the data set. The counts in each cell are

Counts	Predicted A	Predicted B	Sum True		*Proportions*	Predicted A	Predicted B	
True A	3	1	4		True A	3/4=0.75	1/4=0.25	
True B	2	20	22		True B	2/22=0.09	20/22=0.91	
Sum Predicted	5	21					**Accuracy**	**0.83**

Figure 6. Confusion matrix. The left table counts the number of samples correctly and incorrectly classified. The right table represents the sample proportions (number of samples divided by total population of the true class), and derives an accuracy score based on the average of the accuracy value for each class (0.75 and 0.91, respectively).

weighted by the total count of the samples in each class to produce the proportion of samples correctly and incorrectly classified. The accuracy metric used here is the average of the values in the diagonal; that is, the average number proportion of samples from each class that were correctly classified. See Figure 6 for an example.

Data used for training and testing the classification system were drawn from experimental sessions that were separated by gaps spanning several minutes. The tasks used for the training and testing sessions were identical in nature. The metric used to assess the efficacy of the classification system was the proportion of testing data correctly classified by the classifier as represented by the confusion matrix-derived accuracy metric. The average of the true positive and true negative classification rates of the system reflected both the sensitivity and specificity of the classifier.

The trained classifier assigned each data sample to the 1-back or 2-back category. Results based on chance alone would yield a classification accuracy of 50%. The system was able to classify testing data with an average accuracy of 83% (3 participants, s.d. 10%) with data from the BioSemi system and 75% (2 participants, s.d. 12%) with the ABM system (Lan et. al., 2005). The difference in performance associated with the two systems might lie in the difference in the number of sensors provided by each system. The challenge for the Honeywell team was to test whether the classification method could assist in a mobile, more realistic, environment.

Mobile Field Evaluation: Method

Objective

The objectives of the mobile field evaluation were to test the effectiveness of the cognitive state classification approaches and assess the impact of mobility on classification performance. The tasks were designed to approximate operationally relevant dismounted soldier tasks while still affording some experimental control. The tasks used in the evaluation required the participant to be mobile in all scenarios. The sensors and output of the artifact removal algorithms were required to provide the classifiers with good signals to discriminate between the low and high workload during completion of the scenarios.

It was hypothesized that the scenario design would reliably put participants into high or low states of workload. This hypothesis was tested as part of the evaluation. If the hypothesis were true, then it was expected that the classification algorithms should achieve better-than-chance correct correlations between the cognitive state classification output and the known levels of task load, based on moment-to-moment classification. However, it was anticipated that signal degradation and loss attributable to significant artifacts may preclude the levels of classification performance seen during laboratory studies.

Participants

Eight participants completed the evaluation. All were males between the ages of 21 and 42 (mean = 29.5, *SD* = 7.8), with between 16 and 21 years of education (mean = 18.6, *SD* = 1.8). None had military experience. All had normal or corrected 20/20 vision and normal hearing.

Apparatus

Efforts focused on deployment of a cognitive state sensor system in a mobile experimental test environment. The primary challenge was fielding an integrated sensing, computational, and interactive system within a mobile hardware ensemble. The prototype ensemble was organized around the U.S Army MOLLE (Modular Lightweight Load-Carrying Equipment) backpack that provided the framework on which to integrate multiple sensors, interface devices, network adapters, and data collection computer.

Transitioning from a laboratory environment with computer simulations to a field exercise required network communications to support experimental requirements such as scripting and stimuli presentation. During the field experiment, a remote computer ran scripts that played prerecorded radio broadcasts to simulate communication traffic to a dismounted infantry leader. Initially, all sensed data were transmitted wirelessly to a remote desktop computer, which calculated the cognitive workload state of the participant and triggered the adaptive automation. The remote computer also logged data for post hoc analysis.

However, network connectivity and reliability across the experimental test field posed a considerable challenge and motivated the migration of all data logging and reasoning to be done on the backpack laptop carried by the participant. After streamlining the EEG signal conditioning algorithms, migrating all hardware interfaces to the backpack laptop, and integrating and testing other external hardware modules, we performed an early system integration test. Subsequently, all software components for signal processing, adaptive automation reasoning, and data logging were migrated to the backpack computer.

Tasks

The design of the scenarios to empirically assess classification accuracy was subject to a multitude of sometimes contrary constraints, as noted previously. Tasks were chosen to be “classifiable,” meaning the tasks within the scenario reliably put participants in the cognitive workload state of interest. The Honeywell team worked with the U.S. Army Natick Soldier Center to develop an operational scenario that closely aligned with operational doctrine, training, and execution of military missions.

Each participant played the role of a platoon leader navigating along a known and secure route to an objective while communicating over the radio. Each of the participants completed four experimental trials, each with periods of low and high task loads. The navigation task increased the overall task complexity as well as testing the performance of the neurophysiological and physiological sensors and cognitive state classifiers while the participant was mobile. In addition to navigation, participants performed the following tasks:

- *Maintain Radio Counts*. The participant kept a running total of civilians, enemies, and friendlies reported to him over the radio by the company commander while ignoring the counts reported to two other platoon leaders. The participant was periodically prompted to report his counts.
- *Mission Monitoring*. The participant monitored three virtual squads moving in bounded overwatch (one squad moves while the other two squads provide protection). When all three squads reported that they were in position, the participant ordered the appropriate squad to move forward. The order of the squads reporting, as well as the squad to move forward, was randomized.
- *Interruption Task*. A series of math problems were periodically (one problem/min) presented to the participants as an interruption task during the scenario. This task was representative of any type of unanticipated interruption that requires significant cognitive resources and an immediate response from the platoon leader. Once the interruption task started, participants had 10 s to answer the problem correctly.
- *Maintain Situation Awareness*. In addition to the situation awareness they needed to perform on the other tasks listed here, participants were asked about the content of additional low-priority messages they received.

Stressors were used to make the scenarios more representative of the actual environment in which soldiers operate. Stressors included time pressure to complete tasks (for example, the countdown clock on the mathematical task) and the increased rate of messages in the high task load elements of the scenario. Participants were encouraged to keep moving throughout the scenarios. The stress and anxiety brought on by competition was explored by offering a monetary award for the highest score at the end of the evaluation.

Procedures

Independent variable. Task load was either high or low. Within each scenario there were blocks of high and low task load conditions that lasted approximately 5 min and 3 min, respectively. The primary difference between high and low task load periods was the pace of radio communications. The composite rate of Maintain Count and Mission Monitoring messages was approximately 2.4 times faster in the high task load period (8.7 messages/min) than the low task load period (3.6 messages/min).

Experimental design. This was a single-factor (task load block: high/low) within participants' design. Each scenario had four task load blocks in a fixed order: high, low, high, low.

Training trials. There were two components to the training that were conducted before the participant performed the experimental trials. The first training session was to ensure that all participants had basic familiarity and proficiency with all the tasks they were to perform in the experiment. The second training session was to collect data with which to train the cognitive state classifiers. After collecting between 5 and 10 min of EEG spectra data for both low and high task load training conditions, we submitted the data to the composite classification system to identify patterns to distinguish the workload conditions. This was done on the same day as the evaluation.

Experimental trials. Scenarios were run in a large grassy field surrounded by light forest situated behind Honeywell in northeast Minneapolis, Minnesota. Participants primarily interacted with a handheld radio and a personal digital assistant (PDA). Input for the mission monitoring and the Maintain Counts tasks came over the radio, and they responded over the radio as well. The math interruption task, which was completed on a PDA, occurred at equal frequencies under both task load conditions. At the end of each block, participants were asked to fill out subjective workload surveys.

Data Analysis

The principal goal of the data analysis was twofold: (a) determine whether the difference in task load invoked a concomitant difference in cognitive workload, and (b) validate that the cognitive state classification algorithms can distinguish these differences in task load.

Subjective workload ratings of mental demand, physical demand, temporal demand, performance, effort, and frustration were taken via the NASA-TLX rating scale (Hart & Staveland, 1988). NASA-TLX was given at the end of each experimental task load block. Successful cognitive workload manipulation was assessed by comparing the subjective workload ratings with the task load manipulation. In addition, objective performance measures on the tasks were compared across low and high task load blocks as another indication of differentiated workload. Objective measures included

- *Maintain Counts*: Reported versus actual counts of civilians, enemies, and friendlies.
- *Mission Monitoring*: Errors in choice of which squad to send forward, and errors in the timing of move command.
- *Tertiary Mathematical Task*: Response time to initiation alert, time to solve the problem, and response accuracy.

Classification accuracy was assessed by comparing the cognitive state classification accuracy across the low and high task load periods within each block. The classification system provided cognitive state assessments every 2 s, providing a moment-to-moment assessment. As mentioned earlier, the accuracy metric used to evaluate the classifier was derived from a confusion matrix.

Mobile Field Evaluation: Results

Subjective Results

Workload was manipulated by varying the task load (rate of incoming messages) over a block of time. The NASA-TLX was administered to confirm that the participants experienced a change in perceived workload. The TLX scores were compared in the high and low task load blocks (see Figure 7). An analysis of variance (ANOVA) was performed on the measures to study within-participants contrasts. Differences were considered significant for alpha < .05. During the high task load blocks, participants recorded a significant increase in mental demand, $F_{1,7} = 13.4$, $p < .01$; temporal demand, $F_{1,7} = 23.5$, $p < .01$; performance, $F_{1,7} = 20.0$, $p < .01$; effort, $F_{1,7} = 25.9$, $p < .01$; and frustration, $F_{1,7} = 15.0$, $p < .01$, as compared with the low task load blocks. The only measure that did not change significantly was physical demand, $F_{1,7} = .006$, $p > .10$, which was expected, as the scenario design did not vary the physical demands in the two task load conditions.

Performance Results

Figure 8 (page 258) illustrates the task-related ANOVA results (alpha < .05) in the low and high task load blocks. Participants showed reduced accuracy on the Mission Monitoring task in the high task load periods (67.4%) as compared with the low task load periods (95.8%). This difference was significant, $F_{1,7} = 24.7$, $p < .01$. The difference in the Maintain Counts performance was not significant. On the math

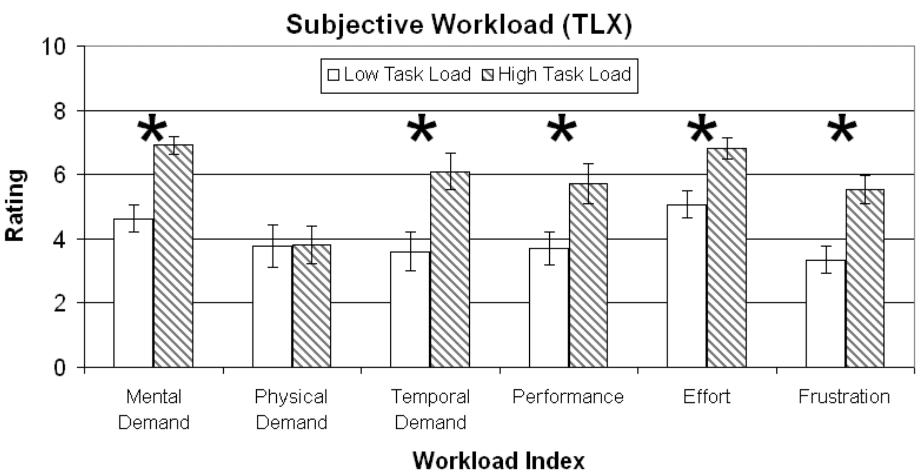

Figure 7. Subjective assessment of workload in the high and low task load blocks; significant differences are denoted with an asterisk.

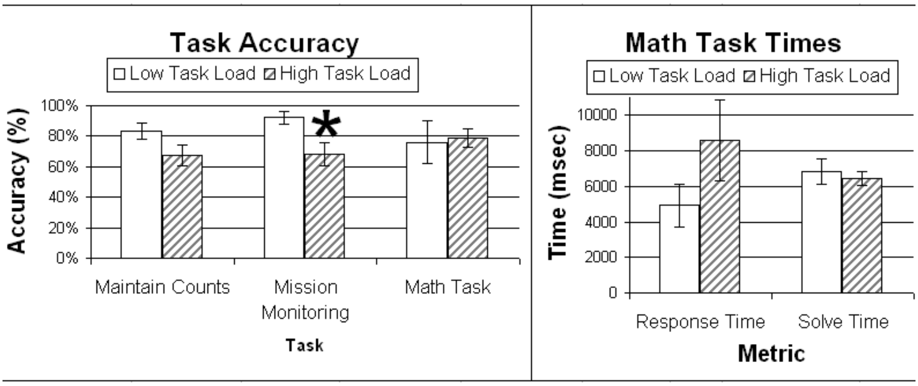

Figure 8. Task metrics across task load conditions; significant differences denoted with an asterisk.

interruption task, participants responded faster in the low task load block (loss of data left only $n = 4$, so the difference was not significant), and the solve time and accuracy showed no difference.

The subjective ratings of workload, as well as the behavioral results from the Mission Monitoring task during the low and high task load blocks, all lend confidence to the hypothesis that the scenario design did indeed create two distinct levels of cognitive workload among the participants. The ability of the real-time cognitive state classification system to correctly characterize the task load blocks is the topic of the next section.

Classification Results

A crucial component of classification in field settings was a systematic procedure for selecting a subset of EEG features that was robust to potential artifacts and provided a basis to discriminate between workload classes. One way to do this was through an exhaustive selection of every possible feature combination drawn from the training data. Then the feature subset producing the best classification performance could be selected for classifying cognitive state in the field. However, such an exhaustive search would result in 2^n searches, where n represents the number of features. Instead, *backward elimination* (Langley, 1994) was used, a heuristic procedure that searches the space of possible feature subsets to identify those that would provide reliable classification. Feature selection was based on the training data that were obtained prior to the testing data and under the same task conditions. With an appropriate selection of channels, by this approach it was possible to classify cognitive state with an accuracy that exceeded 70% for all participants. The mean classification accuracy was 74.4% with a standard deviation of 9.01%. A classification accuracy as high as 95% was observed for one participant (see Figure 9). Data from one participant (s6) were lost because of a system malfunction.

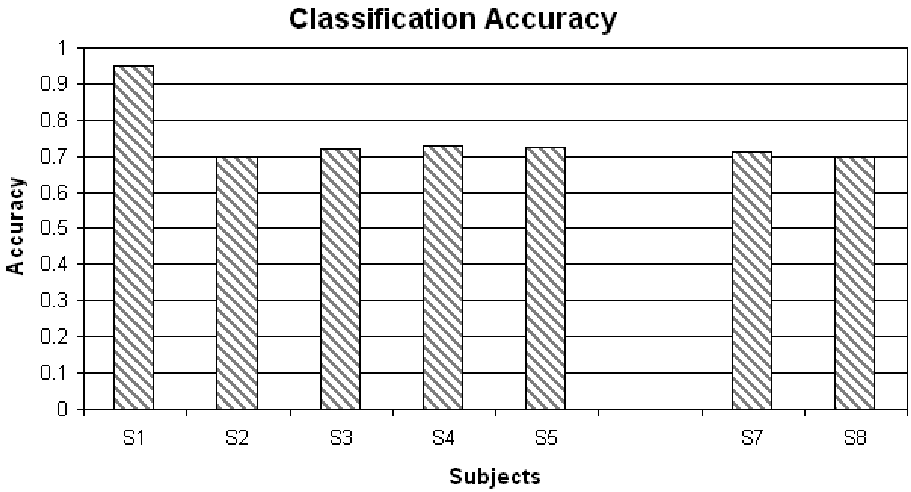

Figure 9. Moment-to-moment classification accuracy for each participant.

Performance with both the BioSemi (participants s7 and s8) and ABM (5 participants: s1–s5) system was close to identical in the field environment. This finding was in contrast to lab assessments, in which the 32-channel BioSemi system provided better performance relative to the six-channel ABM system (Dorneich, Whitlow, Ververs, Mathan et al., 2005). A possible explanation for this discrepancy may be because of the differences in the hardware design. The large number of relatively unconstrained cables that are associated with the BioSemi might have been susceptible to movement-induced vibration, which may have been a potential source of noise. Any benefits of additional channels that the BioSemi system provided may have been lost because of vulnerabilities to these movement artifacts. In contrast, the ABM system was specifically designed for mobile use. If these results are replicated with a larger group of participants, there may be a need for hardware specifically designed to withstand the rigors of mobility in the field.

Discussion

Findings

A series of substantial scientific and engineering issues needed to be successfully addressed in order to deliver compelling results for the mobile cognitive state classification. First, participants needed to be placed in reliably high and low task loading conditions within an operationally relevant mobile task scenario. This was validated across multiple performance and subjective measures. These results lent confidence that the classification assessment approach was tested against task conditions that were perceived and elicited performance commensurate as low and high cognitive task loads.

Second, the evaluation confirmed that the signal-processing and classification algorithms not only ran on a mobile computing platform in real time but delivered moment-to-moment cognitive state classification performance greater than 70% for all participants. In order to deal with poor signal-to-noise ratio under mobile EEG collection, real-time signal processing was developed to remove eye blinks, exclude data contaminated by muscle artifacts, account for eye movements, correct for DC drift, eliminate spikes, and remove motion-induced, high-frequency components. The net result was that the signal-processing solution preserved sufficient signal quality to decipher differences in the EEG spectral dynamics under low and high cognitive loads.

There are many reasons why these results constituted a significant contribution to the emerging field of augmented cognition as well as to the broader field of experimental neuroscience. First, the granularity of the classification performance was at the 2-s resolution and did not depend on larger samples to classify disparate states. Classification performance represented the percentage of all data samples, approximately 300 in high blocks and 180 in low blocks for all participants who were correctly classified. It was a far more common practice to report average classification performance over an entire experimental block or time windows substantially greater than 2 s, neither of which was a particularly germane measure when evaluating a system that adapts in real time.

Second, the task conditions were far more heterogeneous, variable, and ecologically valid than was typically seen in prior classification studies in which participants performed a single well-defined laboratory task. In both low- and high-task conditions, participants were required to perform three separate tasks along with requisite task switching and working memory rehearsal. As is the case for most "cognition in the wild," participants adopted different strategies to manage the multiple-task execution (as evidenced in postexperimental questionnaire responses). To achieve reasonably good classification rates under these conditions indicates that the utility of EEG in classification was likely to extend to more ecologically valid task conditions.

Third, the classifiers were trained with data from a distinct period that was completed before the test phase. In many classification studies, researchers sample training and test samples from the same block, often from temporally adjacent samples. It is well known that EEG baselines drift over time, described as *nonstationarity*, as is common to many physiological processes; therefore, running a classifier on training data from a previous period was a technical risk but resulted in validating the approach in a more rigorous manner.

Fourth, the system used data from relatively few sensor sites – six sites from the ABM system and seven from the BioSemi system – because any imagined field deployment needs to minimize the number of sensors. Many researchers strive to maximize the number of sensors to ensure adequate coverage to provide them with the spatial resolution to capture subtle differences across the cortex. These findings suggest that even relatively sparse EEG arrays provide sufficient coverage to distinguish between the two task loading conditions.

Fifth, the current study achieved encouraging classification between two states that are very similar in the classes of cognitive processing that is required, such as working memory, but differ substantially in the intensity or tempo of processing required. This suggests that the approach detected differences in executive functions that supported the management of multiple tasks over time.

Finally, all these findings indicate that this reported EEG approach will be an effective means of triggering adaptive systems in real-world applications. This approach provides the temporal resolution to respond to short-term changes in cognitive state that would be required for applications such as communications scheduling (Dorneich, Whitlow, Mathan, Carciofini et al., 2005). In this study, the communications scheduler (adaptation) applied messaging techniques that included drawing attention to higher-priority items with additional alerting tones or visual text messages and deferring lower-priority messages to a commander's display device for later review. Communication scheduling significantly increased the accuracy in maintaining counts in high task load conditions (67.4% accuracy unmitigated, 95.7% mitigated). Likewise, the communications scheduler significantly increased the accuracy of mission monitoring in high task load when mitigation was available (68.2% unmitigated, 95.8% mitigated). Because the focus of this article was the feasibility of assessing cognitive state in a mobile participant, space constraints precluded full discussion of the adaptive automation performance results in the current evaluation (see Dorneich, Whitlow, Mathan, Ververs et al., 2005).

Lessons Learned

In addition to the performance findings discussed previously, many practical lessons were learned in the assembling, fielding, and evaluating of EEG-based classifiers in a mobile setting. A summary of lessons learned in this work is presented in the table below.

Lessons Learned

Area	Lesson Learned
Task definition	Consult domain experts. The U.S Army Natick Soldier Center was consulted in designing "operationally relevant" tasks. This saved considerable time, and results will be better received due to their ecological validity. The use of representative tasks lends more confidence that the findings will be transferable to an actual domain.
Task definition	Baseline tasks early and often to ensure that representative participants perform and perceive different task loads as low and high. Initial assumptions about what participants could handle in terms of a "high" communications tempo were quickly challenged by the data collected with pilot participants.
Signal processing	Develop the capability to collect data in an actual environment. A novel stability control was created to improve filtering of ocular activity. When faced with the extreme artifacts in a mobile environment, most adaptive filters would become unstable and unusable.

Continued on next page

Lessons Learned (continued)

Area	Lesson Learned
Signal processing	Critically review similar research to understand application to the target domain. Findings from prior research were quickly identified as inadequate for identifying relevant EEG sites for use in applied operational domains. Given the dynamic multitasking nature of the mobile task environment, the limited relevance of controlled laboratory studies in down-selecting to a subset of channels was discovered. Most studies involved well-defined, homogeneous, stationary tasks that typically reported averaged results and not moment-to-moment classification accuracy. Collect sufficient data to determine how much training data are required to provide good classification performance. Use pilot studies to determine how much training data were required to provide robust classification performance. The amount of data needed varies depending on the nature of the task environment, signal-to-noise ratio, and classification techniques used.
Classification	Fit the approach to the constraints of the environment. Explore multiple temporal windows in considering the constraints imposed by the sensor density, computational efficiency, precise task adaptation needs, and the high degree of classification accuracy during ongoing research studies. Determine the ideal number of sensors by considering the processing demands, operational environment, and generalizability of the classification across multiple situations. It was determined that more sites were not always better for machine learning classification. Once the classifier approach goes beyond the most informative features (site by frequency band), the classifier begins to overfit to noise and degrade classification performance – much like adding unnecessary parameters to a regression model.
System integration	Ruggedize the equipment for testing in a field environment. Most ruggedized laptops not only come with shock-mounted hard drives to protect your data but include better thermal management, which is sorely lacking in traditional laptops (as was found one warm, sunny spring day). Select an EEG system that preamplifies the signal at the electrode site to enable low noise measurements.
Program management	Whenever possible, simplify the experimental design to reduce the complexity of conducting field studies. Inevitably the system integration phase will take three times longer than expected. By limiting the number of research questions of interest and avoiding rolling up everything in a single study, implementation of overall findings for the study is more manageable. This ambitious study involved making a novel system fieldable, creating realistic operational tasks with separable cognitive task loads, and adapting a classification approach to the operationally relevant tasks, all of which seriously challenged timetables, budgets, and overall resources.
Risk management	Consider an experimental design that includes segments with severable benefits (meaning that if something breaks or if it starts raining, the data collected up to that point will be usable) so that a lengthy data collection does not become "all or nothing." With a lengthy, elaborate experiment using an elaborate system, the probability of running start to finish without some glitch approaches zero.

Lessons Learned (continued)

Area	Lesson Learned
Participant recruitment	Within the bounds of any Institutional Review Board (IRB) agreement, recruit motivated participants for lengthy experiments of this nature. From the time the participants arrived until they cleaned the EEG gel out of their hair, these experimental sessions lasted a minimum of 5 to 7 hr, during which they wore a 35-lb backpack and an EEG sensor headset with gel, walked the navigation course for at least an hour, and performed very challenging cognitive tasks. Fortunately, this study recruited individuals who were intrinsically motivated, competitive, and highly intelligent.

Limitations

Although the classification results reported here are promising, several shortcomings need to be addressed in future work. First, some of the results we described will have to be validated against larger groups of participants. Second, although the classification approach seems to generalize over periods spanning minutes and hours, it is unclear whether the system can generalize over larger temporal gaps between training and testing. Third, the work reported here has focused only on EEG; however, considering other information sources, such as cardiac sensors and fNIR, may make cognitive state estimation more robust under circumstances in which EEG may be compromised. Fourth, the cognitive state classifiers evaluated in this study use Bayes rule to make decisions about the cognitive state, based on EEG feature vectors. However, making optimal classification decisions within a Bayesian framework also requires consideration of the prior probability of various workload states and the cost of actions that are associated with the cognitive state-related decisions. The current implementation assumes equal priors for each state and does not weigh the cost of actions. Consideration of priors and costs will be an important priority as the technology described here is transitioned into an operationally relevant system.

Next Steps

As the technology transitions from mobile, experimental scenarios to future operational integration events, the Honeywell classification approach will be tailored to address likely deployment challenges. Feedback from Army partners indicates the Honeywell sensor and computational component must address the following high-level requirements:

- Provide reliable performance under harsh dismounted conditions.
- Integrate with other FFW subsystems in a manner that does not appreciably increase weight, size, power consumption, network bandwidth utilization, or computational resources.
- Garner very high levels of user acceptance and operational acceptance.

Classification Accuracy

The classifier approach will continue to be developed to address some of the limitations discussed earlier. Evaluating the classification approach with a larger set of users who are operating in their natural task environment will be the focus of the next evaluation. In addition, cardiac sensors as well as EEG sensors will be assessed with the goal of fusing the sensor streams to provide more robust, reliable, and accurate classification. In future work, we will also look at the consideration of priors and costs in the classification decision.

System Reliability

Maintaining system reliability under harsh conditions is the reality of the dismounted soldier domain. In addition to the common challenge for all electronics in the battlefield to be ruggedized, a system that measures neurophysiological signals must confront the considerable noise introduced by motion, sweating, and muscle activity. In previous sections, we discussed the means by which these artifacts were addressed for the participants operating in the mobile, multitasking scenarios.

The next steps to improve system reliability will involve rigorous testing within dismounted operational environments, which will expose the system to increased physical stress and a variety of environmental conditions and will likely introduce new classes of signal artifacts that have not yet been encountered. This would provide an opportunity to improve signal processing by isolating and addressing, either by advanced data filtering or physical integration improvements, the new sources of noise.

System Fieldability

Effective integration with FFW component systems essentially implies the need to continue to reduce the hardware, software, computational, and power footprint of the system. In a matter of 2 years, the computational platform has transitioned from a five-desktop immobile system to a fully wearable mobile system that relies on only a laptop computer in the participant's backpack (see Figure 10). In addition to the dramatic hardware reduction, the sensing and signal-processing requirements have been streamlined to be tractable on a single standard laptop. There will be continued efforts to streamline the sensing system to ensure that it is as small, power-efficient, and reliable as possible.

In the future, much of the signal-processing and classification calculations could be done on dedicated hardware rather than utilizing software processing capacity. The determining factor in the computational load of the classification system is the number of sensor sites necessary for robust classification. Toward that goal, the system has transitioned from using the BioSemi Active Two system with 32 channels of EEG to the ABM 6-channel sensor headset.

Furthermore, reducing computational requirements will be explored by encoding neurophysiological signal processing onto a hardware system that would require less software computation from the FFW wearable computer. Finally, potential network protocols that utilize the minimum bandwidth while still transmitting the requisite volume of feedback to provide value to the FFW suite will be explored. This requires secure, efficient, and wireless data transmission from the integrated sensors

Figure 10. Initial (left) and current (right) systems.

to a local signal processor for managing artifacts and spectrally decomposing signals for subsequent classification. Ultimately, a fielded FFW augmented cognition system will likely require advanced sensors, integrated hardware signal processing, and highly efficient software agents running on the FFW mobile computer. Such a system would be capable of triggering adaptations to the warfighters' task environment based on their cognitive state.

The next steps to improve fieldability include exploring sensor options that have a reduced footprint compared with current sensing systems. For example, free-field or minimal-preparation EEG electrode-based systems that are easily integrated into a helmet liner or embedded within helmet pads will be considered.

System Form and Function Acceptability

In order for a system to be successfully fielded, user acceptance is critical to ensure use in the battlefield environment. User acceptance for an augmented cognition system includes ease of donning and doffing, comfortable integration with Advanced Combat Helmet (ACH), and satisfaction of functional expectations. The ACH is the replacement of the old Kevlar Army helmet and is designed to be lighter, stronger, and compatible with current night vision devices, communications packages, and nuclear, biological, and chemical defense equipment and body armor (Global Security, 2006). Specifically, the system would need to be seamlessly integrated into the ACH to a degree that a warfighter could simply don the helmet to enable the sensors that are either integrated within the helmet liner or helmet padding, without adhesives or electrolyte gel. The sensor-enabled helmet must be reasonably comfortable to wear for extended durations.

Finally, the augmented cognition system should deliver value and satisfy functional expectations to justify the addition, however small, of power, weight, and computational requirements. Initial implementations of the augmented cognition system would involve providing cognitive state information to remotely located commanders or key leaders to assess the cognitive combat readiness of their subordinates.

The next step in addressing these challenges is experimentation in an operational environment that will further constrain the form and functional requirements. This step will also provide a test environment to perform cognitive classification studies with considerably more ecological validity, further proving the feasibility and utility of determining cognitive states of interest in an operational environment.

Adaptive System Triggering

Work continues on building adaptive systems that use cognitive state assessment as triggers. Automation is an effective means to allow users to conserve cognitive resources to allocate to other higher priority tasks (Dixon & Wickens, 2004; Rovira, Zinni, & Parasuraman, 2002). Using an assessment of the cognitive state of the user on which to base decisions when to apply automation is one method of adaptive automation. The work described here focuses on real-time assessment of a human's capacity to understand and use information while under high task load conditions, in which cognitive capacity can fluctuate greatly. In task management, mitigation strategies might include intelligent interruption to improve limited working memory, attention management to improve focus during complex tasks, or cued memory retrieval to improve situational awareness and context recovery.

Ultimately, the goals of adaptive automation are similar to those of automation in general: improve overall performance while avoiding "operator out of the loop" conflicts or mistrust in the automation. Such technologies not only have the potential to significantly reduce the strain on soldiers' cognitive resources but also provide the opportunity to improve overall decision making by better managing information flow (Schmorrow, Raley, & Ververs, 2004). The overall result is a benefit by making smarter decisions about what information gets presented, when it is presented, and how it is presented.

Acknowledgments

The authors thank Danni Bayn, Jim Carciofini, Natalia Mazaeva, Trent Reusser, James Sampson, and Jeff Rye for their contributions to this work. We also thank Glenn Wilson for an early review of the manuscript, as well as the anonymous reviewers, all of whom provided excellent comments and feedback.

This research was supported by a contract with DARPA and funded through the U.S. Army Natick Soldier Center under Contract No. DAAD16-03-C-0054, for which Dylan Schmorrow and Amy Kruse served as the program managers of the DARPA Improving Warfighter Information Intake Under Stress/Augmented Cognition program and Henry Girolamo was the U.S. Army program manager and DARPA agent. Any opinions, findings, conclusions, or recommendations expressed herein are those of the authors and do not necessarily reflect the views of DARPA or the U.S. Army.

References

Backs, R. W., & Seljos, K. A. (1994). Metabolic and cardiorespiratory measures of mental effort: The effects of level of difficulty in a working-memory task. *International Journal of Psychophysiology, 16*, 57–68.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91*, 276–292.

Berka, C., Levendowski, C., Cvetinovic, M. M., Petrovic, M. M., Davis, G., Lumicao, M. N., Zivkovic, V. T., Popovic, M. V., & Olmstead, R. (2004). Real-time analysis of EEG indices of alertness, cognition, and memory acquired with a wireless EEG headset. *International Journal of Human-Computer Interaction, 17*(2), 151–170.

Boiten, F. A. (1998). The effects of emotional behaviour on components of the respiratory cycle. *Biological Psychology, 49*(1–2), 29–51.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, 39,* 1–38.

Dixon, S. R., & Wickens, C. D. (2004). Automation reliability in unmanned aerial vehicle flight control. In D. Vincenzi (Ed.), *Proceedings of Human Performance and Situation Awareness and Automation (HPSAA) II*. Daytona Beach, FL: HPSAA.

Dorneich, M. C., Whitlow, S. D., Mathan, S., Carciofini, J., & Ververs, P. M. (2005). The communications scheduler: A task scheduling mitigation for a closed loop adaptive system. *Proceedings of the 11th International Conference on Human-Computer Interaction (HCI International 2005).* Mahwah, NJ: Erlbaum.

Dorneich, M. C., Whitlow, S. D., Mathan, S., Ververs, P. M., Pavel, M., & Erdogmus, D. (2005). DARPA improving warfighter information intake under stress: Augmented cognition phase III final report. *Technical report for DARPA Augmented Cognition Phase 3 (contract DAAD 16-03-C-0054)*.

Dorneich, M. C., Whitlow, S. D., Ververs, P. M., Carciofini, J., & Creaser, J. (2004). Closing the loop of an adaptive system with cognitive state. *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting* (pp. 590–594). Santa Monica, CA: Human Factors and Ergonomics Society.

Dorneich, M. C., Whitlow, S. D., Ververs, P. M., Mathan, S., Raj, A., Muth, E., Hoover, A., DuRousseau, D., Parra, L., & Sajda, P. (2004). DARPA improving warfighter information intake under stress. Augmented cognition: Concept validation experiment (CVE) analysis report for the Honeywell team. *Technical report for DARPA Augmented Cognition Phase 2B (contract DAAD 16-03-C-0054)*.

Erdogmus, D., Adami, A., Pavel, M., Lan, T., Mathan, S., Whitlow, S., & Dorneich, M. (2005). Cognitive state estimation based on EEG for augmented cognition, presented at the 2nd IEEE EMBS International Conference on Neural Engineering, Arlington, VA, March 16–19.

Garavan, H., Ross, T. J., Li, S.-J., & Stein, E. A. (2000). A parametric manipulation of central executive functioning using fMRI. *Cerebral Cortex, 10*, 585–592.

Gevins, A., Smith, M. E., McEvoy, L., & Yu, D. (1997). High resolution EEG mapping of cortical activation related to working memory: Effects of task difficulty, type of processing, and practice. *Cerebral Cortex*, *7*, 374–385.

Gevins, A., & Smith, M. E. (2000). Neurophysiological measures of working memory and individual differences in cognitive ability and cognitive style. *Cerebral Cortex, 10*, 829–839.

Gevins, A., & Smith, M. (2003). Neurophysiological measure of cognitive workload during human-computer interaction. *Theoretical Issues in Ergonomics Science, 4*(1–2), 113–132.

Global Security. (2006). Advanced combat helmet (ACH). Retrieved October 24, 2007, from *http://www.globalsecurity.org/military/systems/ground/ach.htm*.

Hart, S. G., & Staveland, L. E. (1988). Development of a multi-dimensional workload rating scale: Results of empirical and theoretical research. In P. Hancock & N. Meshkati (Eds.), *Human mental workload*. Amsterdam: Elsevier.

Heslegrave, R. J., & Furedy, J. J. (1979). Sensitivities of HR and T-wave amplitude for detecting cognitive and anticipatory stress. *Physiology & Behavior, 22*(1), 17–23.

Institute of Medicine of the National Academies. (2004). Strategies for monitoring cognitive performance. In *Monitoring metabolic status: Predicting decrements in physiological and cognitive performance* (pp. 171–172). Washington, DC: National Academy Press.

Izzetoglu, K., Bunce, S., Onaral, B., Pourrezaei, K., & Changem, B. (2004). Functional optical brain imaging using near-infrared during cognitive tasks. *International Journal of Human-Computer Interaction, 17*, 211–227.

Kalsbeek, J. W. H., & Ettema, J. H. (1963). Scored irregularity of the heart pattern and measurement of perceptual or mental load. *Ergonomics*, *6*, 306–307.

Kittler, M., Hatef, R., Duin, R. P. W., & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(3), 226–239.

Kramer, A. (1991). Physiological metrics of mental workload: A review of recent progress. In D. Damos (Ed.), *Multiple task performance* (pp. 279–328). London: Taylor & Francis.

Lan, T., Erdogmus, D., Adami, A., Pavel, M., & Mathan, S. (2005). Salient EEG channel selection in brain computer interfaces by mutual information maximization. *27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 7064–7047). Los Alamitos, CA: IEEE.

Makeig, S., & Jung, T.-P. (1995). Changes in alertness are a principal component of variance in the EEG spectrum. *NeuroReport, 7*(1), 213–216.

Mathan, S., Mazaeva, N., Whitlow, S., Adami, A., Erdogmus, D., Lan, T., & Pavel, M. (2005). Sensor-based cognitive state assessment in a mobile environment. *Proceedings of the 11th International Conference on Human-Computer Interaction*. Mahwah, NJ: Erlbaum.

Mikulka, P., Hadley, G., Freeman, F., & Scerbo, M. (1999). The effects of a biocybernetic system on vigilance decrement. *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting* (p. 1410, abstract). Santa Monica, CA: Human Factors and Ergonomics Society.

Neumann, D. L. (2002). Effect of varying levels of mental workload on startle eyeblink modulation. *Ergonomics, 45*, 583–602.

Parasuraman, R., Mouloua, M., & Hilburn, B. (1999). Adaptive aiding and adaptive task allocation enhance human-machine interaction. In M. W. Scerbo & M. Mouloua (Eds.), *Automation technology and human performance: Current research and trends* (pp. 129–133). Mahwah, NJ: Erlbaum.

Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies, 59*(1–2), 185–198.

Parzen, E. (1967). On estimation of a probability density function and mode. In *Time series analysis papers*. San Diego, CA: Holden-Day, Inc.

Pope, A. T., Bogart, E. H., & Bartolome, D. (1995). Biocybernetic system evaluates indices of operator engagement. *Biological Psychology, 40*, 187–196.

Popivanov, D., & Mineva, A. (1999). Testing procedures for non-stationarity and non-linearity in physiological signals. *Mathematical Biosciences*. *157*(1–2), 303–320.

Porges, S. W., & Byrne, E. A. (1992). Research methods for measurement of heart-rate and respiration. *Biological Psychology, 34*(2–3), 93–130.

Prinzel, L. J., Freeman, F. G., Scerbo, M. W., Mikulka, P. J., & Pope, A. T. (2000). A closed-loop system for examining psychophysiological measures for adaptive automation. *International Journal of Aviation Psychology, 10*, 393–410.

Rovira, E., Zinni, M., & Parasuraman, R. (2002). Effects of information and decision automation on multi-task performance. *Proceedings of the Human Factors Society 26th Annual Meeting* (pp. 327–331). Santa Monica, CA: Human Factors and Ergonomics Society.

Russell, C. A., & Gustafson, S. G. (2001). *Selecting salient features of psychophysiological measures* (Air Force Research Laboratory Technical Report AFRL-HE-WP-TR-2001-0136).

Scerbo, M. W., Freeman, F. G., Mikulka, P. J., Parasuraman, R., DiNocero, F., & Prinzel, L. J. (2001). *The efficacy of psychophysiological measures for implementing adaptive technology* (NASA Technical Report TP-2001-211018, June 2001). Hampton, VA: National Aeronautics and Space Administration, Langley Research Center.

Schmorrow, D. D., & Kruse, A. A. (2002). Improving human performance through advanced cognitive system technology. *Proceedings of the Interservice/Industry Training, Simulation and Education Conference*. Washington, DC: National Defense Industrial Association.

Schmorrow, D., Raley, C., & Ververs, P. (2004). Toward effective warfighting in stressful environments. Poster presented at Human Performance and Situation Awareness and Automation (HPSAA) II. Daytona Beach, FL.

Scott, W. B. (1999). Automatic GCAS: You can't fly any lower. *Aviation Week and Space Technology*, *150*(5), 76–79.

Stern, J. A., Boyer, D., & Schroeder, D. (1994). Blink rate: A possible measure of fatigue. *Human Factors*, *36*, 285–297.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*, 520–522.

Veltman, J. A., & Gaillard, A.W. K. (1998). Physiological workload reactions to increasing levels of task difficulty. *Ergonomics, 41*, 656–669.

Verwey, W. B., & Veltman, H. A. (1996). Detecting short periods of elevated workload: A comparison of nine workload assessment techniques. *Journal of Experimental Psychology-Applied, 2*, 270–285.

Welch, P. D. (1967). The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, *15*(2), 70–73.

Wickens, C. D., & Hollands, J. (2000). *Engineering psychology and human performance* (3rd ed.). New York: Prentice Hall.

Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. *IRE WESCON Convention Record* (pp. 96–104). New York: Institute of Radio Engineers.

Wientjes, C. J. E. (1992). Respiration in psychophysiology: Methods and applications. *Biological Psychology, 34*(2–3), 179–203.

Wildervanck, C., Mulder, G., & Michon, J. A. (1978). Mapping mental load in car driving. *Ergonomics, 21*, 225–229.

Wilson, G., & Russell, C. (2003). Operator functional state classification using multiple psychophysiological features in an air traffic control task. *Human Factors, 45*, 381–389.

Wilson, G. F., & Eggemeier, F. T. (1991). Physiological measures of workload in multi-task environments. In D. Damos (Ed.), *Multiple-task performance* (pp. 329–360). London: Taylor & Francis.

Yamada, F. (1998). Frontal midline theta rhythm and eyeblinking activity during a VDT task and a video game: Useful tools for psychophysiology in ergonomics. *Ergonomics, 41*, 678–688.

Michael Dorneich (Ph.D., cognitive engineering, University of Illinois, 1999) is a senior research scientist at Honeywell Laboratories, Human Centered Systems Group. His expertise includes design of mixed-imitative systems, adaptive human-machine interaction in complex domains, interactive learning environments, and collaborative decision support systems. Recent work includes developing neuroadaptive systems in the DARPA Augmented Cognition program, for which he is co-PI. For NASA he has designed alerting systems, decision aids, and displays/controls for the Orion program.

Stephen D. Whitlow (M.S., cognitive psychology, University of Illinois, 1995) is a senior research scientist at Honeywell Laboratories, Human Centered Systems Group. His technical expertise includes neuroadaptive systems, mixed-initiative design, and data visualization. He has worked in domains including aviation, medical systems, cyber-security, military operations, unmanned vehicles, satellite control, and petrochemical processing. He is the PI of Honeywell's Augmented Cognition program, which is developing adaptive systems that manage information flow based on real-time sensed cognitive state.

Santosh Mathan (Ph.D., human-computer interaction, Carnegie Mellon University, 2004) is a senior research scientist at Honeywell Laboratories, Human Centered Systems Group. His projects focus on sensor-based approaches to cognitive state estimation in real-world application contexts. He is PI on DARPA's Neurotechnology for Intelligence Analysts program and leads pattern recognition and signal-processing efforts under DARPA's Augmented Cognition program. He is also exploring the potential for using sensor-based workload estimation techniques in human computer-interaction assessment contexts.

Patricia May Ververs (Ph.D., Honeywell Aerospace, research scientist) is a principal research scientist in the Human Centered Systems section at Honeywell Aerospace in Columbia, MD. She received her Ph.D. in engineering psychology from the University of Illinois at Urbana-Champaign in 1998. Since joining Honeywell in 1998, she has led programs in the areas of high-speed research, cockpit alerting, advanced primary flight displays, augmented cognition, and neurotechnologies. Ververs serves as the Synthetic Vision System Integrated Product Team lead for the DARPA Sandblaster program, developing technologies to support helicopter landings in brownout conditions.

Deniz Erdogmus is an assistant professor jointly with the Computer Science and Electrical Engineering Department and the Biomedical Engineering Department of Oregon Health and Science University. His research focuses on information-theoretic adaptive signal processing and biomedical applications, including brain interfaces. He received two B.S. degrees in electrical and electronics engineering and mathematics in 1997 and an M.S. in electrical and electronics engineering in 1999 from Middle East Technical University, Turkey. He received his Ph.D. in electrical and computer engineering from the University of Florida (2002).

Andre Gustavo Adami is an assistant professor in the Computer Science Department of the University of Caxias do Sul, Brazil. He received his B.Sc. in computer science in 1994 from the University of Caxias do Sul, his M.Sc. in computer science in 1997 from the Federal University of Rio Grande do Sul, Brazil, and his Ph.D. in electrical engineering in 2004 from the Oregon Health and Science University.

Misha Pavel is a professor in the Department of Biomedical Engineering at Oregon Health and Science University. He is the director of Point-of-Care Laboratory, focused on unobtrusive monitoring and neurobehavioral assessment and modeling. His research interests include analysis and modeling of perceptual and cognitive processes, pattern recognition, information fusion, and decision making in healthy and impaired individuals. He received his PhD. in experimental psychology from New York University and earned his M.S. in electrical engineering from Stanford University.

Tian Lan received both a B.S. and an M.S. in electronics engineering from Beijing University of Aeronautics and Astronautics in 1997 and 2000, respectively. He is a Ph.D. candidate and research assistant in electrical engineering at OGI School of Science & Engineering, Oregon Health & Science University. His research work focuses on signal processing, adaptive systems, machine learning, and information theory and their applications in biomedical signal processing, including EEG, MRI, and brain-computer interfaces.