Using a Low-Cost Electroencephalograph for Task Classification in HCI Research

Johnny Chung Lee
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
johnny@cs.cmu.edu

Desney S. Tan
Microsoft Research
One Microsoft Way, Redmond, WA 98052
desney@microsoft.com

ABSTRACT

Modern brain sensing technologies provide a variety of methods for detecting specific forms of brain activity. In this paper, we present an initial step in exploring how these technologies may be used to perform task classification and applied in a relevant manner to HCI research. We describe two experiments showing successful classification between tasks using a low-cost off-the-shelf electroencephalograph (EEG) system. In the first study, we achieved a mean classification accuracy of 84.0% in subjects performing one of three cognitive tasks - rest, mental arithmetic, and mental rotation - while sitting in a controlled posture. In the second study, conducted in more ecologically valid setting for HCI research, we attained a mean classification accuracy of 92.4% using three tasks that included non-cognitive features: a relaxation task, playing a PC based game without opponents, and engaging opponents within the game. Throughout the paper, we provide lessons learned and discuss how HCI researchers may utilize these technologies in their work.

Categories and Subject Descriptors: 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].
General Terms: Human Factors, Experimentation.
Keywords: Brain-Computer Interface, human cognition, physical artifacts, task classification, Electroencephalogram (EEG).

INTRODUCTION

For generations, humans have fantasized about the ability to communicate and interact with machines through thought alone or to create devices that can peer into a person's thoughts. These ideas have captured the imagination of humankind in the form of ancient myths and modern science fiction stories. However, only in recent decades have advances in neuroscience and brain sensing technologies made measurable progress toward achieving that vision. These technologies allow us to monitor the physical processes within the brain that correspond with certain forms of thought.

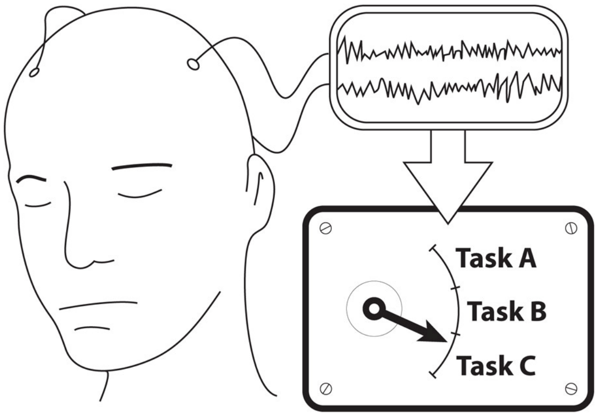

Figure 1 – A conceptual illustration of a Brain-Computer Interface using EEG signals for task classification.

Primarily driven by growing societal recognition for the needs of people with physical disabilities, researchers have used these technologies to build brain-computer interfaces (BCIs), communication systems that do not depend on the brain's normal output pathways of peripheral nerves and muscles [17]. A conceptual illustration of a BCI system is shown in Figure 1. In these systems, users explicitly manipulate their brain activity instead of using motor movements to produce signals that can be used to control computers or communication devices. The impact of this work is extremely high, especially to those who suffer from devastating neurodegenerative diseases such as amyotrophic lateral sclerosis, which eventually strips an individual of all voluntary muscular activity while leaving cognitive function intact.

Although removing the need for motor movements in computer interfaces is challenging and rewarding, we believe that the full potential of brain sensing technologies as an input mechanism lies in the extremely rich information it could provide about the state of the user. Having access to this state is valuable to HCI researchers because it may allow us to derive more direct measures of traditionally elu-

UIST'06, October 15–18, 2006, Montreux, Switzerland.
Copyright 2006 ACM 1-59593-313-1/06/0010...$5.00.

sive phenomena such as task engagement, cognitive workload, surprise, satisfaction, or frustration. These measures could open new avenues for evaluating systems and interfaces. Additionally, knowing the state of the user as well as the tasks they are performing may provide key information that would allow us to design context sensitive systems that adapt themselves to optimally support the state of the user.

The work we present in this paper is an initial step in exploring how BCI technology can be applied to HCI research. First, we demonstrate that effective exploration in this field can be accomplished using low-cost sensing equipment and without extensive medical expertise. An experiment we conducted shows that we were able to attain 84.0% mean accuracy classifying three different cognitive tasks using an off-the-shelf electroencephalograph (EEG) costing only USD$1500. Within this experiment, we present a reusable experimental design adapted from previous BCI work and discuss lessons learned so that other HCI researchers can build upon our experiences to perform their own explorations. Second, we present a novel approach to performing task classification by utilizing both cognitive and non-cognitive artifacts measured by our EEG as features for our classification algorithm. In a second experiment, we attained a mean classification accuracy of 92.4% on three tasks within a more ecologically valid setting, determining various user states while playing a PC based game. We close with a discussion of how this approach can be useful in certain areas of HCI research.

BACKGROUND AND RELATED WORK

Brain Sensing and EEG Primer

The human brain is a dense network consisting of approximately 100 billion nerves cells called neurons. Each neuron communicates with thousands of others to regulate physical processes and produce thought. Neurons communicate either by sending electrical signals to other neurons through physical connections or by exchanging chemicals called neurotransmitters. Advances in brain sensing technologies enable us to observe the electrical, chemical, or blood flow changes as the brain processes information or responds to various stimuli.

In this paper, we focus on the Electroencephalograph (EEG), a technology used everyday in hospitals and clinics and the most commonly used technology in contemporary BCI research. For general reviews of BCI research, see [4,16,25]. Figure 2 provides a table of alternative brain sensing and imaging technologies and their primary disadvantages for BCI work, especially within the HCI community [20].

EEG uses electrodes placed on the scalp to measure the weak (5-100µV) electrical potentials generated by brain activity. Each electrode typically consists of a wire leading to a gold-plated disk that is attached to the scalp using conductive paste or gel. An EEG records the voltage at each of these electrodes relative to a reference point, which is often simply another electrode on the scalp [7]. Because EEG is a passive measuring device, it is safe for extended and repeated use, a characteristic crucial for adoption in HCI research.

Brain Sensing Technology	Primary Disadvantage
Electrocorticogram (ECoG)	Highly invasive, surgery
Magneto-encephalography (MEG)	Extremely expensive
Computed Tomography (CT)	Only anatomical data
Single Photon Emission Computerized Tomography (SPECT)	Radiation exposure
Positron Emission Tomography (PET)	Radiation exposure
Magnetic Resonance Imaging (MRI)	Only anatomical data
Functional Magnetic Resonance Imaging (fMRI)	Extremely expensive
Event-Related Optical Signal / Functional Near-Infrared (EROS/fNIR)	Still in infancy, currently expensive

Figure 2. A table of current brain sensing technologies and their primary disadvantages for HCI research.

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 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. Each input channel of an EEG includes a multistage amplifier with a typical gain of 20,000.

Unfortunately, this high electrical sensitivity also makes an EEG susceptible to interference from a variety of sources such as physical movement of the person’s body, indoor power lines, and other electronic equipment. BCI researchers have invested a great deal of effort in creating experimental designs, specialized testing facilities and equipment, and software filtering techniques to minimize the presence of these non-cognitive artifacts [5]. However, such a high degree of environmental and experimental control can be impractical for HCI research that aims to eventually function in a typical home or office scenario. In the work presented in this paper, we limited ourselves to a typical office computing environment without any specialized acoustic or electromagnetic insulation. These studies were run in an unmodified office of an active researcher containing multiple computers, fluorescent lights, and other typical sources of signal interference found in an office building.

Because much of the work in BCI has grown out of the rehabilitation engineering and neuroscience domains, a large portion of previous research has used high-end devices costing between USD$20,000-250,000 [e.g. see systems from www.biosemi.com or www.egi.com]. We were unable to find previous examples of successful BCI research performed with low-end EEG systems that are accessible to HCI researchers. While [3,14,22] have explored applying brain sensing technology to HCI related problems, we have found in our interactions that many HCI researchers are hesitant to explore the domain due to the perception of prohibitively high costs associated with owning and maintaining this equipment. Others may feel that the required domain expertise presents a major obstacle. In this paper, we demonstrate that effective BCI research can be accomplished without requiring such high-end and high-cost devices, as well as to provide HCI researchers with a quick primer for entry into the field.

EEG for Task Classification

We focus our attention on EEG work related to the problem of task classification, which has received significant attention because BCI technology is most useful as an input control or communication device if the system is capable of discriminating at least two states within the user. With this ability, a computer can translate the transitions between states or the persistence of a state into a form that is appropriate for controlling an application [13]. Previous methods for accomplishing this can be divided into two approaches: operant conditioning and pattern recognition [21]. Operant conditioning places the user in a tight feedback loop with the system output and the user must learn how to control their brainwaves in order to achieve the desired results. On the other hand, pattern recognition places the burden on signal processing and machine learning techniques in order to recognize the signals associated with mental states or activities of untrained individuals without feedback from the system. The benefit of pattern recognition is that the tedious training and adaptation needed to bridge the gap between human and machine is performed by the computer rather than the human. From an HCI perspective, this approach is much more attractive because it can be applied to detecting and classifying arbitrary states, rather than having the user generate pre-trained states on demand. We utilize this basic approach in our work.

Researchers have explored a wide range of neurological phenomena in building BCIs using EEG and pattern recognition. For example, many have studied event-related and evoked potentials which represent distinct voltage fluctuations as a response to specific stimuli [11]. These potentials include signals such as the P300 response, commonly thought to be an index of attention and memory, as well as the N100 response, which is a selective attention wave. Unfortunately, measuring these phenomena typically requires presenting stimuli at regulated timings and under carefully controlled conditions. Additionally, extracting the signal from the noisy data often requires averaging over dozens or hundreds of recordings that are time-locked with the stimulus presentation. While researchers working on extracting these signals from single trials have had some success [23], the general paradigm of tightly controlling stimuli and watching for the absence or presence of a particular response makes these types of signals less useful for task classification and the HCI work we envision.

Fitzgibbon et al. [8] observed statistical differences in the spectral power of EEG signals while subjects performed eight different cognitive tasks. However, statistical differences alone do not necessarily imply the ability to classify between these tasks. Hence, we look to work that has used signal changes to generate features for machine learning algorithms for classifying which cognitive tasks an individual is performing. While early attempts by Gevins et al. did not yield favorable results [12], recent work has been more successful. Keirn and Aunon [13] collected EEG measurements while users performed five different mental tasks: a baseline relaxed state, a multiplication task, a geometric rotation task, a letter composition task, and a visual counting task. These tasks were designed to elicit hemispheric differences on the head derived from the neurophysiological mapping of brain function. Using various feature selection techniques and machine learning algorithms, they were able to achieve 75-90% classification accuracies when comparing pairs of these tasks using within-subject models.

Using Keirn and Aunon's data set, Palaniappan [19] showed that he could obtain up to 97.5% classification accuracy when using the most easily separable pair of tasks for each subject. However, the best performing pair was different for each subject. While differentiating between two known and pre-selected states is sufficient for direct control applications, it is less interesting for general HCI research as it does not allow measuring arbitrary states for any given user. Also using Keirn and Aunon's data set, Anderson and Sijerčić [1] extend this work by using neural networks and temporal averaging to classify the five tasks simultaneously, achieving 33-70% classification accuracy using data from four subjects. The prior, the expected performance of a random classifier, in this case would be an accuracy of 20%. In many cases, a human observer trying to correctly identify the active cognitive task would only be expected to perform as well as a random classifier.

Unfortunately, many researchers working on these problems seem to re-utilize the Keirn and Aunon data set rather than obtain new data. In fact, we have been unable to find work within the last fifteen years replicating these measurements. In order for this approach to be useful to the HCI community, we must be able to reliably replicate the data acquisition procedure, preferably with low-cost equipment that is accessible to HCI researchers. This would provide us with a starting point from which we can perform the experimental manipulations necessary to develop new and relevant applications. Hence, in our first study, we adopt the general experimental methods presented by Keirn and Aunon to collect new data using a subset of their tasks.

EXPERIMENT 1: COGNITIVE TASK CLASSIFICATION

We conducted this experiment to explore the feasibility of using a low-cost EEG to measure and classify brain signals while subjects performed various cognitive tasks.

Tasks

Based on the results from pilot recordings with our system, we chose three tasks from those used by Keirn and Aunon [13]:

Rest – In this task, our baseline, we instructed participants to relax and to try not to focus on anything in particular. We also explicitly instructed them not to continue working on any task that may have preceded the rest task.

Mental Arithmetic – In this task, participants performed mental multiplication of a single digit number by a three digit number, such as 7 × 836. We chose the complexity of the problems so that it was not so difficult as to be discouraging, but also so that it would take most participants more than the allotted time to complete it. We instructed participants to double check their answers if they finished before the time expired. This ensured that they were performing the intended task as well as they could throughout the task period. Since we did not have participants provide us with answers, we confirmed that the problems were keeping them busy for the duration of the task during a debriefing interview.

Mental Rotation – In this task, participants imagined specific objects, such as a peacock, in as much detail as possible and rotating in space. The specific details of the object were left to the participant.

Equipment

We used a Brainmaster AT W2.5, a PC-based 2-channel EEG system [2]. This device retails for approximately USD$1500, comparable to the cost of a laptop computer. The device has a minimum sensitivity of 0.7µV, provides 8-bits per channel, and has a sampling rate of 256 Hz. Prior research has relied on data from much higher-end systems that provide much higher resolution analog-to-digital conversion, higher sampling rates, and more channels. We show in this experiment that this low-cost device is sufficient for measuring the signals of interest.

It is worth noting that the technology used within EEG is not inherently expensive and the raw electrical components within a low-end device can be acquired for less than USD$100. It is merely a very sensitive digital voltmeter. The OpenEEG project [18] provides instructions for end-users and hobbyists to assemble even lower cost devices than the one we used. However, validating these extremely low-cost do-it-yourself devices was not within the scope of this paper.

EEG Electrode Placement

The 10-20 System is an international standard for EEG electrode placement locations on the human scalp [7]. The system defines a grid relative to physical landmarks on the head, such as the indentation between the nose and forehead (nasion), and the bump on the back of the head (inion) at the occipital protuberance as shown in Figure 3. Electrode locations are defined by either 10% or 20% increments between these landmarks.

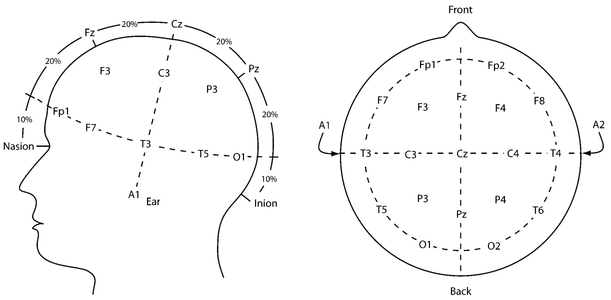

Figure 3. International 10-20 Electrode Placement System

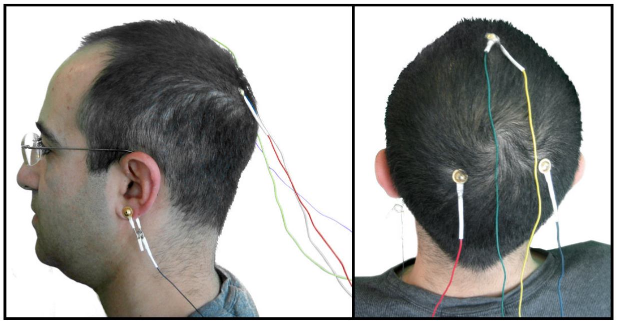

Figure 4. Electrode placement in our experimental setup.

Based on results from pilot recordings, we selected the parietal (P3 and P4) regions as the locations of interest, with both electrode references tied on top of the head at the central region (Cz). The placements of electrodes can be seen in Figure 4. Tying the references for the two EEG channels together allowed us to make meaningful comparisons between the values from each channel. The Brainmaster device also includes a ground electrode connection that we attached to an ear lobe. The purpose of grounding is to provide electrical protection that prevents damaging the sensitive inputs of the device. Its specific location on the head does not impact the recorded signals.

To attach an electrode, we first clean the scalp location with a small amount of Nuprep™, an abrasive skin prepping gel used to remove dirt, oil, and dead skin from the area in order to reduce the impedance of the electrical connection with the scalp. Then, we place a small amount of Ten20™ conductive paste on the electrode and attached the electrode to the scalp. The paste improves the electrical connection and provides a temporary bond that holds the electrode on the scalp. The measured impedance of our electrode connections was approximately 20KΩ. The setup procedure requires about 10 minutes. Once the experiment was complete, we removed the electrodes and subjects could wash off any remaining gel and paste with a brief water rinse.

Procedure and Design

After ensuring that participants were comfortably seated in a chair and attaching the EEG electrodes to their scalp, we explained the three cognitive tasks. We then had them perform several practice trials to ensure they understood the tasks. Participants performed all tasks within this experiment with their eyes closed. We did this to minimize motion artifacts in the EEG signal that would result from blinking or eye movement. Similarly, we instructed participants to keep their head and body as still as they could and not to outwardly vocalize thoughts while performing tasks. This highly constrained subject posture is typical of EEG based BCI work due to the extreme sensitivity EEG has to muscle movement.

A pre-recorded computer driven audio cue indicated when a task was beginning by saying "Rest", "Math", or "Rotate". Following the "Math" prompt, the experimenter read aloud the multiplication problem. For example, they would say "seven times eight-three-six" representing 7 × 836. We presented the stimulus by stating only each digit to remove the irregular language and duration of compound numeric terms. After the "Rotate" prompt, the experimenter verbally provided the object, such as "peacock", for the participant to imagine and mentally rotate. Each task lasted 14 seconds. We used a blocked design similar to the design used by Kiern and Aunon. We grouped each set of three tasks into a trial, six trials into a session, and ran three sessions per participant. This provided 18 recordings for each task. We fully counterbalanced the order of tasks across the trials. Tasks and trials were run back to back, resulting in sessions that took just over 4 minutes. Participants took short breaks between sessions.

In this design, we cycle through many blocks of short tasks rather than using a single long recording for each task because EEG signal properties naturally drift over short periods of time [7]. This drift occurs both in the DC component as well as in the higher frequency spectrum of the signal. While some researchers have attributed this to such factors as changes in conductivity on the scalp due to minor perspiration or persistent neurological shifts, the reasons behind the drift are generally not well understood. However, through observations within our pilot experiments and reviewing prior literature describing this phenomenon, we can be fairly certain that it is not a direct result of cognitive activity brought about by the different tasks in this study. Hence, cycling through shorter recordings in counterbalanced order reduces the correlation of any unrelated drift with the tasks, and minimizes the risk of inadvertently using this feature for classification. This ensures we are creating a task classifier rather than a temporal drift detector.

Participants

Eight individuals (3 females) volunteered for this study. Participants ranged from 29 to 58 years of age (mean age of 37.8). All were cognitively and neurologically healthy, and all were right handed, except for one participant who had a slight nerve injury in his right hand and who had trained himself to depend more on his left hand. The study took about 30 minutes and participants were given a small gratuity for their time. The individual used for pilot experiments did not participate in these studies.

Data Analysis and Classification Results

In order to classify the signals measured from our EEG, we performed some basic signal processing to transform the time series data into a time independent data set. We then computed a set of base features that we mathematically combined to generate a much larger set of features. Next, we used a feature selection process to prune the feature set, keeping only those that added the most useful information to the classifier and to prevent over-fitting. Our feature generation and selection process was similar to that used by Fogarty et al. in their work on modeling task engagement to predict interruptibility [9]. We used these features to train a Bayesian Network and perform the classification. Finally, we discuss how averaging may be used to enhance the classification accuracies leading us to our final results. Each of these steps is described in the following subsections.

Basic Signal Processing

Since the brain exhibits a characteristic electrical response to all forms of sensory stimuli, the auditory prompting used in the experimental setup likely introduces unwanted response artifacts into the EEG data. For this reason, we remove the first 4 seconds from each task recording. This is sufficient time for both the computer and the experimenter prompts to complete and allows a moment for task onset to occur within the participant. The remaining 10 seconds then contains only signals during which the task was actually being performed. This provides 180 seconds of 2-channel 8-bit EEG data for each task and each participant.

Since most machine learning algorithms do not handle time-series data well, we must convert the data into a time independent dataset. To do this, we adopt a technique used in previous work [e.g. 1,13]. We slice the EEG signal into small overlapping windows and compute features based the content of each window. Specifically, we divide each 10-second task recording into 2-second windows overlapping by 1 second. This provides 9 windows per task period and a total of 486 windows for each participant. This set of windows becomes the set of training instances used by the machine learning algorithm for constructing and validating the classifier model.

Feature Generation

EEG data is typically analyzed by looking at the spectral power of the signal in a set of six standard frequency bands which have been observed to correspond with certain types of neural activity [7]. These frequency bands are: 1-4Hz (*delta*), 4-8Hz (*theta*), 8-12Hz (*alpha*), 12-20Hz (*beta-low*), 20-30Hz (*beta-high*), and 30-50Hz (*gamma*). Taking a Fourier transform of the EEG data provides us with the frequency content of the signal. It is worth noting that the *gamma* band is sometimes defined as not having an upper frequency bound (e.g. simply >30Hz). However, the data recorded from our EEG has very little signal above 50Hz with the exception of 60Hz interference from indoor power lines. As a result, we decided to limit the frequency range used for our *gamma* band.

Adopting features used in previous work [e.g. 1,13], we compute the following for each window: *signal power* in each of the six frequency bands for each channel, *phase coherence* (similarity in mean phase angle) in each band across channels, and each *band power difference* between the two channels. This results in 24 features that are commonly used for EEG signal analysis. In addition to these features, we also compute the following set of more general signal properties for each input channel: *mean spectral power*, *peak frequency*, *peak frequency magnitude*, *mean phase angle*, *mean sample value*, *zero-crossing rate*, *number of samples above zero*, and the *mean spectral power difference* between our two input channels. In total, this provides with 39 base features for each window. We then compute the product and division of each pair of base features resulting in 1482 additional features. Non-linear manipulations of features such as this is a common machine learning technique used to compensate for a potential lack of expressiveness in the statistical model use by the classifier [9].

Because our output variable is a three-valued nominal variable (*Rest*, *Math*, and *Rotate)*, it is useful to divide our continuous valued input variables into discrete bins that might provide meaningful discriminations in the output variable. Doing this would allow the model, for example, to determine that having a *beta* band power above a particular threshold might be indicative of a *Math* task rather than attempting to find a correlation in the continuous values that would be useful in separating all the three tasks. We did this for our data using Fayyad and Irani's Minimum Description Length method [6] as implemented by Weka, an open source data mining and machine learning tool [24].

Feature Selection

Once we have generated our full set of 1521 features, we apply a feature selection process to eliminate non-predictive features and to prevent over-fitting our data. First, we applied Weka's *CfsSubsetEval* operator, which evaluates subsets of features favoring those that have a high correlation with the output variable while having low inter-correlation among the features within the selected set. This provides a computationally inexpensive method of identifying a relatively small subset of useful features for the classification problem. On average, this process reduced the number of features to 51 for the 3-task classifiers and 39.2 features for the pair-wise classifiers, the models that only discriminate between two tasks. We then applied a more computationally expensive wrapper-based feature selection process, which builds a classifier model beginning with an empty set of features and then incrementally adds or removes features based on their impact on overall classification accuracy. This further reduced the number of features used for classification, resulting in an average of 23 features for the 3-task classifiers, which have 486 example windows, and 16.4 features for the pair-wise classifiers, which have 324 example windows.

	3 task	**Math v. Rotate**	**Rest v. Math**	**Rest v. Rotate**
subject 1	67.9%	83.3%	88.0%	85.8%
subject 2	70.6%	82.7%	91.4%	84.3%
subject 3	77.6%	88.3%	93.8%	86.7%
subject 4	63.6%	69.4%	84.9%	86.7%
subject 5	66.5%	91.0%	81.2%	80.9%
subject 6	59.3%	80.6%	80.2%	68.5%
subject 7	71.4%	87.3%	90.4%	86.7%
subject 8	69.8%	87.7%	82.4%	83.6%
Mean	**68.3%**	**83.8%**	**86.5%**	**82.9%**

Figure 5. Classification accuracies per subject for the three mental tasks used in study 1.

We applied this process of feature generation and selection to the data from each participant separately, catering the set of most predictive features to the individual. Constructing per-participant models is common in previous work [1,13,19] due to the high-variance in EEG signal properties between individuals. This is analogous to early days of speech recognition when systems had to be trained to particular individuals. However, close inspection of our final selected sets revealed that the base components *mean spectral power*, *alpha*, and *beta lo* frequently appeared among the most predictive features for many participants. This could be the basis of exploring cross-user classifiers in future work.

Baseline Classification Results

After performing the feature selection procedure for each participant, we used a Bayesian Network classifier to identify which task was being performed during a given test window. Rather than using standard 10-fold cross validation to estimate the classification accuracies of the models, we used 18-fold cross validation to control for the block design of the data collection procedure. For each fold, the model trained on 17 of the 18 available trials and reserved one trial for testing. A trial contains 9 contiguous windows for each task. Each of our reported results is the mean classification accuracy after repeating this process 18 times using a different test trial for each fold. This is more representative of the performance that we would expect if a new trial was recorded and tested than if we used standard 10-fold cross validation.

The Bayesian Network classifiers for three mental tasks yield classification accuracies of between 59.3 and 77.6% (μ=68.3, σ=5.5), depending on the user. The prior for these classifications, or the expected result of a random classifier, is 33.3%. The pair-wise classifiers have a prior of 50% and yield accuracies of between 68.5 and 93.8% (μ=84.4, σ=6.0). Figure 5 presents a full breakdown of the classification accuracies for each subject.

Enhanced Classification Results

The classification accuracies presented in the previous subsection are the accuracies of the models attempting to classify each of the 2-second windows in isolation. They ignore the fact that each window is temporally adjacent to and overlapping several other windows. In this section, we discuss reintroducing the temporal nature of the data by averaging the classifier output over several adjacent windows. Doing this reduces the impact of spurious signals in the EEG data and substantially improves the overall classification accuracy.

To perform averaging, we must first decide the size of our averaging kernel (how many windows we should average over before producing the final classifier output). In scenarios where the classifier has no a priori knowledge of when tasks begin and end, the best strategy may be to choose a fixed kernel size, e.g. 5. If all five windows come from the same task, averaging reduces the impact of noise in the windows and the final output is more likely to be correct. However, if the five windows do not come from the same task then the averaging kernel will contain conflicting information. The classification accuracies will be poorer adjacent to transitions between tasks. Thus, the benefit of averaging the classifier output will depend on the density of task transitions in the data. More precisely, the impact averaging will have on the overall classification accuracy is directly related to the ratio between the kernel size and the expected duration of tasks. To illustrate, we have computed the resulting classification accuracies using a few averaging scenarios that we believe might occur in typical HCI applications.

In the first scenario, the duration of each task may not be significantly longer than the kernel size. For example, if we choose an averaging kernel size of 5 windows for our tasks, the kernel requires 6 seconds worth of data while each task lasts only 10 seconds. This results in 40% of the eventual classifications made by averaging over conflicting data. This provides us with a 5.1% average improvement in the 3-task classifiers, while the pair-wise classifiers only marginally improved over their baseline accuracies.

In the second scenario, the task duration may be significantly longer than the kernel size. In this case, the percentage of poor classification results due to task transition points is reduced. If we simulate this scenario by doing the same averaging over 5 windows but leave out those sets that span transitions, we attain a 12.7% improvement over the 3-task baseline and a mean improvement of 8.0% over the pair-wise baselines.

Lastly, if the classifier knows or can estimate when tasks begin and end, we can further improve classification by expanding the kernel size to span the entire task period. For our study, this meant averaging over the 9 windows generated from each 10-second task recording. This yielded mean 3-task classification accuracies of 75.9 to 90.7% (μ=84.0, σ=6.0) and mean pair-wise accuracies of 75.0 to 100% (μ=94.4, σ=7.4) across all eight participants. This is a mean improvement of 15.7% and 10.0% over their respective baseline accuracies. Figure 6, shows the impact of these averaging schemes under these scenarios relative to the baseline accuracy of the raw classifier output with no averaging. It is interesting to note that no particular pair of tasks was substantially easier to classify than any other pair.

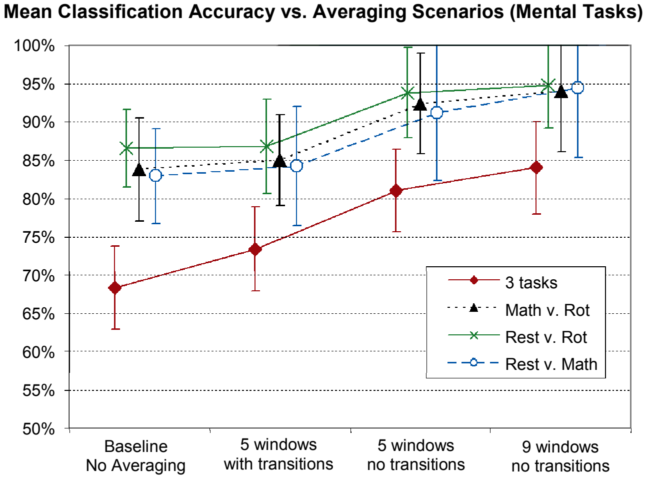

Figure 6. Plot of overall classification accuracies for cognitive tasks in experiment 1 under various averaging scenarios. Error bars represent standard deviation.

We should also note that performing averaging in this manner intrinsically increases the classification latency since measurements from multiple windows are needed by the model to produce an output. While we might be able to mitigate this latency by dynamically changing the kernel size according to various heuristics, exploring such schemes remains future work.

Are We Really Reading Minds?

The classification accuracies we achieve in this study suggest that we can indeed reliably measure and identify performance of our three mental tasks. However, we cannot be certain that the phenomena providing the classification power is entirely generated by neuronal firings in the brain. For example, it is entirely possible that the sensitive EEG is detecting muscular, or Electromyograph (EMG), artifacts from minute subconscious motor movements in various parts of the body. Non-cognitive artifacts detected by EEG include blinking, eye movement, head movement, scalp galvanic skin response, jaw and facial EMG, gross limb movements, and sensory evoked potentials.

We believe that various cognitive tasks are involuntarily coupled with physiological responses [15] and that it is difficult, if not impossible, to fully isolate cognitive activity using EEG in healthy neurologically-intact individuals. This is problematic for researchers aiming ultimately to apply the technology to disabled individuals, as they have to guarantee that the features of interest are generated solely by the brain. For this reason, many researchers have conducted extensive work to remove these ‘confounds’ introduced by physical artifacts before classification [10] or have limited their data collection to include only participants who suffer from the same disabilities as those as their target users. However, since we are aiming to apply this to a generally healthy population, we only need to determine the reliability of the features in predicting the task. This concept was briefly explored by Chen and Vertegaal who sensed motor activity in order to model mental load in their physiologically attentive user interfaces [3]. If non-cognitive artifacts are highly correlated with different types of tasks or engagement, we can exploit these artifacts to improve our classification power. This is in contrast to the neuroscience community which has spent significant efforts to reduce and remove these artifacts from their recordings.

EXPERIMENT 2: GENERAL TASK CLASSIFICATION

We conducted a second experiment to explore using both cognitive and non-cognitive artifacts to classify tasks in a more ecologically valid setting. The tasks we chose involved playing a PC-based video game. This task was chosen because it places the participant in a typical personal computing environment while encouraging a relatively high degree of motor activity with the mouse and keyboard. This experiment also serves as a demonstration of the how the same experimental design can be used as a generalized approach for a much broader range of experimental tasks and potential applications.

Tasks

The game we selected for this experiment was Halo, a PC-based first person shooter game produced by Microsoft Game Studios. The game involves navigating a 3D environment using the keyboard and mouse in an effort to engage opponents using various weapons. Participants were allowed to move their head and body freely as they would naturally while playing the game. The tasks we tested within the game were as follows:

Rest – In this task, again used as our baseline, participants were instructed to relax and fixate their eyes on the cross-hairs located at the center of the screen. During this task, they did not interact with the controls, nor did any of the game elements interact with them.

Solo – In this task, participants navigated the environment, interacting with passive objects or collecting ammunition scattered throughout the scene. None of the enemies were visible or engaged the participant in this task. This was designed to emulate some of the physical movement involved in playing the game without evoking the task engagement of shooting at an opponent.

Play – In this task, participants navigated the environment and engaged an opponent controlled by an expert player. The same expert played against all participants. Game elements and expert player behavior were designed to ensure subjects were properly engaged in performing the task throughout the task period.

Setup, Design, and Procedure

The experimental setup, design, and procedure were similar to the first study. All participants were given a tutorial on the game and allowed to practice until they felt comfortable with the controls. We repeated each of the three tasks 6 times in fully counterbalanced order for each session. Because of the additional time necessary to navigate the virtual environment and engage in the activity, we lengthened the task durations to 24 seconds. Due to the extended duration of the tasks, we ran only two sessions of the game task per participant. The same eight participants that completed the first experiment also took part in this experiment. Six of these participants were novices and two were moderately experienced players.

Results

The data preparation, feature generation and selection, as well as the machine learning procedure were identical to that used in the first experiment. We removed the first 4 seconds from each task period and divided the remaining data into 2-second windows overlapped by 1 second. This provided 19 windows per task period and 684 windows per participant. We computed 1521 initial features which were reduced by feature selection to an average of 20.25 features for the 3-task classifiers and 16 features for the pair-wise classifiers. We estimated the accuracies of the Bayesian Network classifiers using 12-fold cross-validation since only 2 sessions were run providing 12 trial recordings.

The baseline classification accuracies were between 65.2 and 92.7% (μ=78.2, σ=8.4) for the 3-task classifiers and 68.9 and 100% (μ=90.2, σ=8.5) for the pair-wise classifiers. After averaging, we were able to achieve 83.3 to 100% accuracies (μ=92.4, σ=6.4) for 3-task and 83.3 to 100% (μ=97.6, σ=5.1) for pair-wise comparisons. Figure 7 shows the mean baseline accuracy of the classifiers as well as the

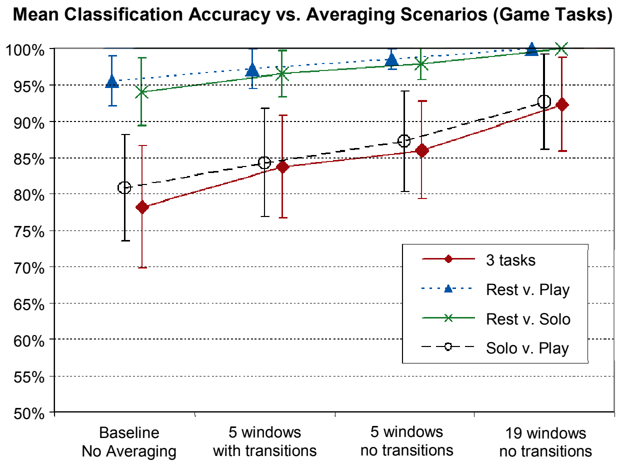

Figure 7. Plot of overall classification accuracies for three game tasks in experiment 2 under various averaging scenarios. Error bars represent standard deviation.

impact of the various averaging schemes described in the previous experiment.

These classification accuracies were substantially higher than our prior experiment using three cognitive tasks. It is also interesting that the classifier for *Solo* vs. *Play* did not do as well as the classifiers comparing against *Rest*. In the *Solo* and *Play* conditions, the amount of motor activity was moderately similar causing a reduction in classification accuracy, while the comparisons against the *Rest* task were easily distinguished by the classifier. This partially illustrates the dramatic impact that varying degrees of motor activity has on the ease of discriminating between different tasks using EEG data. While it may be desirable to speculate on the difference in the level of engagement between the *Solo* and *Play* conditions and the classifiers ability to distinguish them, we cannot make any conclusive claims about mental engagement given the data we collected. We can only state that we were able to achieve a mean accuracy of 93.1% when identifying whether the participant was performing the *Solo* or *Play* task during a given window.

It is also worth mentioning that for this particular experiment, it would have been trivial to create a sensor that would have achieved essentially perfect classification by either hooking into the game state or by monitoring keyboard and mouse activity. Our goal in this experiment was not to illustrate the best method of discriminating these tasks but to demonstrate the impact of non-cognitive artifacts on EEG-based task classification in a realistic computing scenario. We believe that, given these two studies, EEG shows interesting potential as a general physiological input sensor for distinguishing between tasks in a wide variety of computing applications without requiring detailed prior knowledge of the tasks.

DISCUSSION AND FUTURE WORK

While the electroencephalograph was invented nearly a century ago, it is only recently that researchers have begun to apply it to problems outside the medical and neuroscience domains. We believe that success in applying this relatively mature technology to new domains is dependent not only on understanding the device and current usage paradigms, but also on creatively challenging traditional assumptions to create innovative solutions such as embracing, rather than rejecting, motion artifacts to improve classification accuracies. We believe this work represents an initial step in exploring how these brain sensing technologies can be applied in a relevant manner to contemporary HCI research problems.

In describing our two studies, we provide guidance for HCI researchers unfamiliar with the technology, but who would like to use it in their work. We uncover some of the intricacies of experimental design, cautioning researchers to be aware of artifacts such as temporal EEG drift proposing the use of a blocked design that minimizes inadvertently training on signal features orthogonal to the tasks. We also discuss electrode placement and present one such layout used in our experiments. Once data is collected, we describe a process for performing signal processing, feature generation and selection, machine learning, as well as temporal averaging methods to improve classifier performance. This is a general procedure that can be applied to a wide range of task classification problems within HCI. We feel that this process is flexible and robust enough that it can be customized for specialized problems. This work also illustrates that relatively high classification accuracies can be accomplished using off-the-shelf EEG equipment comparable in cost to a typical laptop computer and without extensive medical expertise.

Higher end EEG systems, however, do offer a potential benefit in having a greater number of electrode channels. This improves the chances researchers have in detecting stimulus related brain signals. Extremely high-end EEG systems containing 256 channels or more have begun exploring the idea of using complex electrical models of the human head to perform source localization in an effort to pin-point the location of a signals origin within the brain. However, this work has only had moderate success and still remains very experimental. Obtaining as much information as possible regarding brain activity is important for neuroscience researchers who ultimately wish to make claims regarding the neurological behavior of the brain. However, a much coarser level of detail can be used for simply performing task classification and detection.

The data processing and machine learning procedure described in this paper required approximately 15 minutes on a modern desktop computer for a given subject. However, once the classifier had been trained, classification of new test data occurred nearly instantly allowing it to be used in a real-time implementation. The classifier model could then be updated periodically in the background given the availability of new training data. The effective data rate of such a system would be approximately 10-30 bits/min depending on the degree of averaging, which corresponds to the level of noise filtering in the data stream. This is similar in performance to recent BCI work utilizing high-end EEG systems [4]. With such a slow and relatively noisy signal, EEG-based input is certainly not going to replace the keyboard or mouse anytime soon. However, we believe that such a signal is well suited for applications such as evaluation tools or context sensitive computing where a secondary input stream can be used to supplement the primary input or to augment system behavior.

While the work presented in this paper focuses primarily on EEG technology, we are very interested in exploring how this process can be expanded to other brain and physiological sensing technologies for the purposes of task classification and identification. This work represents a starting point for a wide range of research work exploring how computers can tune into the activity within our minds to help us perform the tasks of our everyday lives. We hope this work will inspire and encourage other researchers in the HCI community to explore these technologies in their own research.

ACKNOWLEDGMENTS

We thank Sumit Basu, Ed Cutrell, Mary Czerwinski, James Fogarty, Eric Horvitz, Scott Hudson, Nuria Oliver, Dan Olsen, John Platt, Pradeep Shenoy, and members of the Visualization and Interaction Group at Microsoft Research for assistance throughout the course of this work.

REFERENCES

1. Anderson, C.W., & Sijerčić, Z. (1996). Classification of EEG Signals from Four Subjects During Five Mental Tasks. *Proceedings of the Conference on Engineering Applications in Neural Networks*, 407-414.
2. Brainmaster. http://www.brainmaster.com.
3. 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.
4. Coyle, S., Ward, T., & Markham, C. (2003). Brain-computer interfaces: A review. *Interdisciplinary Science Reviews, 28*(2), 112-118.
5. Cutmore, T.R.H, & James, D.A. (1999). Identifying and Reducing Noise in Physiological Recordings. *International Journal of Psychophysiology, 32*, 129-150
6. Fayyad, U. M., & Irani, K. B. (1992). On the handling of continuous-valued attributes in decision tree generation. *Machine Learning, 8*, 87-102.
7. Fisch, B.J. (2005). *Fisch & Spehlmann's EEG primer: Basic principles of digital and analog EEG*. Amsterdam: Elsevier.
8. Fitzgibbon, S.P., Pope, K.J., Mackenzie, L., Clark, C.R., & Willoughby, J.O. (2004). Cognitive tasks augment gamma EEG power. *Clinical Neurophysiology, 115*, 1802-1809.
9. Fogarty, J., Ko, A.J., Aung, H.H., Golden, E., Tang, K.P., & Hudson, S.E. (2005). Examining task engagement in sensor-based statistical models of human interruptibility. *Proceedings of SIGCHI 2005 Conference on Human Factors in Computing Systems*, 331-340.
10. Gevins, A., Leong, H., Du, R., Smith, M.E., Le, J., DuRousseau, D., Zhang, J., & Libove, J. (1995). Towards measurement of brain function in operational environments. *Biological Physiology, 40*, 169-186.
11. Gevins, A.S., & Remond, A. (1987). *Handbook of Electroencephalography and Clinical Neurophysiology: Methods of analysis of brain electrical and magnetic signals*. Amsterdam: Elsevier.
12. Gevins, A.S., Zeitlin, J.C., Doyle, J.C., Schaffer, R.E., & Callaway, E. (1979). EEG patterns during 'cognitive' tasks. II. Methodology and analysis of complex behaviors. *Electroencephalography and Clinical Neurophysiology, 47*, 704-710.
13. Keirn, Z.A., & Aunon, J.I. (1990). A new mode of communication between man and his surroundings. *IEEE Transactions on Biomedical Engineering, 37*(12), 1209-1214.
14. 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.
15. Kramer, A.F. (1991). Physiological metrics of mental workload: A review of recent progress. In *Multiple Task Performance* (ed. Damos, D.L.), 279-328.
16. Mason, S.G., & Birch, G.E. (2003). A general framework for brain-computer interface design. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 11*(1), 70-85.
17. Millan, J. Adaptive brain interfaces. *Communications of the ACM, 46*(3), 74-80.
18. OpenEEG Project. http://openeeg.sourceforge.net.
19. 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.
20. 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.
21. Smith, R.C. (2004). Electroencaphalograph based brain computer interfaces. *Masters Dissertation, University College Dublin*.
22. 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
23. van Boxtel, G.J.M. (1998). Computational and statistical methods for analyzing event-related potential data. *Behavior Research Methods, Instruments, & Computers, 30*(1), 87-102.
24. Witten, I.H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd Edition), San Francisco: Morgan Kaufmann.
25. Wolpaw, J.R., Birbaumer, N., McFarland, D.J., Pfurtscheller, G., & Vaughn, T.M. (2002). Brain-computer interfaces for communication and control. *Clinical Neurophysiology, 113*, 767-791.