Human-Aided Computing: Utilizing Implicit Human Processing to Classify Images

Pradeep Shenoy
University of Washington
Box 352350, Seattle, WA 98195
pshenoy@cs.washington.edu

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

ABSTRACT

In this paper, we present Human-Aided Computing, an approach that uses an electroencephalograph (EEG) device to measure the presence and outcomes of implicit cognitive processing, processing that users perform automatically and may not even be aware of. We describe a classification system and present results from two experiments as proof-of-concept. Results from the first experiment showed that our system could classify whether a user was looking at an image of a face or not, even when the user was not explicitly trying to make this determination. Results from the second experiment extended this to animals and inanimate object categories as well, suggesting generality beyond face recognition. We further show that we can improve classification accuracies if we show images multiple times, potentially to multiple people, attaining well above 90% classification accuracies with even just ten presentations.

Author Keywords: Brain-Computer Interface (BCI), human cognition, implicit processing, visual attention, image classification, Electroencephalography (EEG).

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

INTRODUCTION

Distributed computing, which divides complex computing tasks and performs them using processing cycles on multiple computers, has captured the imagination of researchers (eg. [10]). Some of these researchers have set out to design systems that distribute their tasks onto computers with unused processing cycles. For example, SETI@home (setiathome.berkeley.edu) makes use of available processing cycles to analyze radio-telescope data in search of intelligent extraterrestrial signals [2], without impacting the user and their tasks. Similarly, Folding@home (folding.stanford.edu) focuses on simulations of protein folding to find cures for diseases [17], World Community Grid (www.worldcommunitygrid.org) aims to create the world's largest public computing grid, and distributed.net focuses on using these cycles to break cryptographic ciphers.

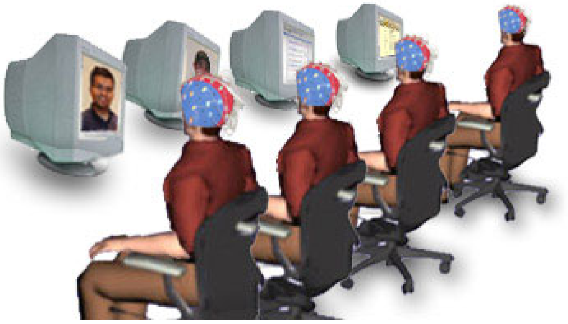

Figure 1: Artist's depiction of multiple people connected to electroencephalograph devices, implicitly classifying images they see.

We believe that there exists a parallel opportunity for utilizing processing cycles in human brains. In his work on Human Computation, von Ahn has built systems that motivate people to perform decision making tasks, mostly within game environments [24]. With clever design, he is able to redirect these conscious human decisions to perform secondary tasks such as labeling images, which would otherwise be tedious for humans and difficult for computers [25]. Amazon.com's Mechanical Turk system operates on a similar principle, except that users are rewarded for their work with small monetary payments [1].

While current work in Human Computation requires conscious attention and explicit intent to perform the specific task, we assert that there is a set of tasks that can be usefully performed by humans even when they are not explicitly trying to perform them. We recognize that the human brain implicitly processes a large amount of environmental information [23]. For example, the brain constantly performs recognition of objects in the environment. In fact, it has been asserted that humans cannot help but process some of these tasks, even when they are actively trying not to. Furthermore, they may not always be aware of the result. Recent advances in neuroscience and brain sensing technologies provide us with the unprecedented ability to interface directly with activity in the brain and measure some of

CHI 2008, April 5–10, 2008, Florence, Italy.
Copyright 2008 ACM 978-1-60558-011-1/08/04…$5.00

the presence and output of this implicit processing. We believe these results can complement existing computational methods such as using machine learning techniques to perform image classification (e.g. [19]).

The contributions of this paper are threefold. First we introduce the novel concept we call *Human-Aided Computing*, which proposes using brain sensing technologies to extract results from the implicit processing that the human brain already performs. We envision that using multiple people to redundantly process information, just as distributed computing technologies use multiple machines, would make the process more robust to individual differences and for complex tasks. See Figure 1 for an artist's depiction of this idea.

Second, we provide background on electroencephalograph (EEG) technology, and review work relevant to Human-Aided Computing. We hope this will lower the barriers to entry for researchers with classical human-computer interaction training, and that these first two contributions will spur interest and inspire creativity within the community.

Third, we present a classification system and results from two experiments that serve as initial proof-of-concept for these ideas. In the first experiment, we show that we can use EEG technology to reliably sense when a user has seen a face or not in a picture that they are trying to memorize. In the second experiment, we extend this result to more categories than just faces, suggesting generality in the approach. We also demonstrate improved classification accuracies if we present images multiple times. In fact, we attain well above 90% classification accuracies with even just ten presentations to single users or distributed to multiple users.

BACKGROUND

Implicit vs. Explicit Processing

The literature dealing with the distinction between implicit and explicit processing is not only extensive and complex, but is also fraught with divergent hypotheses and theories. However, there does seem to be consensus that awareness of a stimulus is preceded by subconscious, or implicit, information processing. The physical features of verbal or visual stimuli, for example, are thought to be implicitly analyzed within the first 250ms or so or presentation [21].

While the implicit processing system seems to be able to simultaneously analyze the physical properties of multiple stimuli, Broadbent provides evidence showing that the channel used for explicit analysis of meaning has limited parallel processing capacity [3]. He asserts that this causes a "bottleneck" in the human information processing system. Consequently, only some of the implicitly processed information can be selected for explicit processing. Recent advances in brain sensing technologies and processing techniques allow us to extract some of the outcomes of these implicit processing activities.

Attentive vs. Explicit Processing

Many psychologists have traditionally assumed that implicit and pre-attentive processing are identical, and so are explicit and focal-attentive processing (e.g. [19]). However, recent work has shown that it is possible to manipulate the two factors independently. For example Koch and Tsuchiya review the growing body of evidence showing that visual processing can occur in the absence of conscious perception [16]. Conversely, people can make certain visual judgments about objects in the attentional periphery.

In our work, we carefully consider the cross between these two factors. While there has been quite a bit of work done in using brain sensing technologies to measure focal-attentive explicit processing (e.g. [7, 9]), much less has been done to explore the implicit and pre-attentive components of cognition. The work reported in this paper represents a move towards focal-attentive implicit processing, that is, the user is paying specific attention to the stimuli, but they are not explicitly trying to perform the task we care about and measure.

Electroencephalography (EEG) Primer

In this paper, we use an Electroencephalograph, a sensing technology that uses electrodes placed on the scalp to measure electrical potentials related to brain activity (see Figure 1). Each electrode consists of a wire leading to a conductive disk that is electrically connected to the scalp using conductive paste or gel. The EEG device records the voltage at each of these electrodes relative to a reference point (often another electrode on the scalp). Electrode placements on the scalp are typically defined by the International 10-20 electrode placement standard [14]. Because EEG is a non-invasive, passive measuring device, it is safe for extended and repeated use, a characteristic crucial for adoption in HCI research. Additionally, EEG does not require a highly skilled operator or medical procedure to use. Recently, Lee and Tan have shown that even low-cost versions of such devices can be used for task classification in HCI research [18]. For more information about electrical signals generated by the brain as well as EEG see [6].

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 coordinated firing of large groups of neurons near the surface of the brain, and then only when they are directed along a vector perpendicular 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. One way to analyze EEG data is to look at the spectral power of the signal in a set of frequency bands, which have been observed to correspond with certain types of neural activity [6]. These frequency bands are commonly defined as 1-4 Hz (delta), 4-8 Hz (theta), 8-12 Hz (alpha), 12-20 Hz (beta-low), 20-30 Hz (beta-high), and >30 Hz (gamma).

Another way the EEG signal can be analyzed is by inspecting the Event-related Potential (ERP), the spatiotemporal pattern of EEG displayed in response to discrete visual or auditory stimuli. The idea is that different kinds of discrete stimuli evoke distinct, characteristic ERPs, which can be detected in the shape of the raw data.

Since EEG is a noisy signal, researchers typically average ERP brain responses across a large number of stimulus presentations and users, resulting in a *grand average* ERP. These grand average responses can be compared across different stimulus classes in order to make statistical statements about the signals. We show an example of an average ERP response to images containing faces in a later section. Unfortunately, showing that a signal is statistically different from another does not trivially allow one to conclude if we can classify such signals, especially when we only have a small amount of data available to us. In our work, we aim to explore whether we can classify such signals for the domain of tasks in which we are interested.

Using EEG and ERPs to Measure Cognitive Processing

ERPs are often used as a tool for exploring the mechanisms and timings of processing in the brain. For example, Johnson and Olshausen argue that the temporal features of the ERP in an image categorization task show two distinct phases [15]. The *early* features, which show up between 100 and 220 ms after presentation of the stimulus in the response, are associated with visual processing of stimuli, whereas the late features, which show up 350 to 550 ms later, are indicative of *post-sensory processing*. Post sensory processing refers to the phase in which the user makes a conscious decision about the perceived stimulus. The study considered two different scenarios. In the first, users were shown a category word (e.g., “face”, “animal”, etc.) and an image, in that order, and asked to decide if the two stimuli represented the same category. In the second, images were shown before the accompanying category word. For both scenarios, the grand average ERPs for the *second* stimulus was compared between cases when the stimulus pairs matched and cases where they did not. They found that the match vs. non-match differences for both images and words were similar in the post-sensory range, indicating that the difference must be the outcome of the decision process in the categorization task.

Other work [4, 9, 13] has shown that different classes of stimuli (for example, faces, cars, animals, mushrooms, chairs, etc.) evoke spatially and temporally different responses. This is true especially of faces, which are possibly the most well-studied class of visual stimuli in the neuroscience community. Faces have received particular attention not only because they are an ecologically important class of stimuli for humans, but also because behavioral and brain-sensing data indicate that faces may be processed in a manner different from other stimuli.

Numerous functional Magnetic Resonance Imaging (fMRI) studies show characteristic activation of certain regions of the brain, popularly known as the *fusiform face area* (this is summarized, e.g, in [9]). In terms of EEG, there is significant evidence that a feature of the ERP response is highly sensitive to faces [22]. This feature is referred to as the N170, named for the negative peak in the raw data seen about 170 ms following stimulus presentation. In particular, manipulation of stimuli that show behavioral differences in face recognition also modulate the N170 feature. Because of how well documented this is, we use this as our starting point and grow our work from there.

Various researchers have shown this even when categorization was not an explicit requirement of the task (e.g. [13]). Unfortunately, most of this work shows statistical differences in grand averages of large amounts of data, and do not show that the signal is at all classifiable, which is a necessary component for real-world use of such a system.

Single Trial ERP Classification

From the perspective of psychology and neuroscience, statistical differences in grand averages are useful tools in informing our understanding of the brain. However, to extract useful information about a particular stimulus, we need the ability to classify *single-trial event-related responses* as belonging to one of a small set of classes. Ideally, we need to be able to do this with a small number of trials and in as little time as possible (i.e. without excessive repeated presentation or large amounts of processing time).

There is quite a bit of work on classifying single-trial ERPs. For example, the brain-computer interface community has extensively used a single-trial P300 recognition response as a direct communication scheme. The P300 is named for the consistent positive peak seen in the ERP approximately 300ms following the presentation of a sought-after stimulus. By locating the stimuli that elicit the P300 response, one can decode the stimuli to which the user is attending. This forms the basis of several brain-operated communication schemes (e.g. a keyboard built on this principle [5]).

Gerson et al. propose cortically-coupled computer vision [7], which exploits the explicit P300 response to speed up human labeling of images. The authors combine sophisticated detection algorithms and a rapid serial presentation scenario to label images at a rate that is significantly faster than manual labeling of images. This work, along with the brain-computer interface work described previously, requires explicit user cooperation and awareness of the task at hand (e.g., categorizing images as target/non-target). In our work, we explore whether the ERP features that encode object category can be used to label images on a single-trial basis, without explicitly requiring users to consciously categorize the images.

OBJECT CLASSIFICATON SYSTEM

Given our experiences with pilot experiments, we built a prototype system in MATLAB that can classify the category of the image at which the user is looking. This system takes as input a stream of EEG data, measured while the

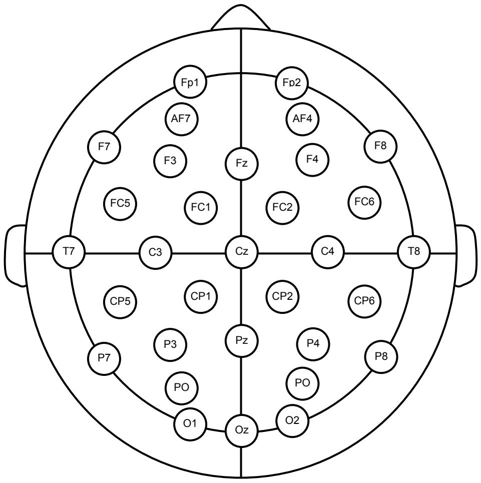

Figure 2: Layout of the channels according to the internationally accepted 10-20 system of EEG electrode placement.

user is viewing images. It does not actually assume a fixed number of EEG channels. In our experiments we used the following channel set: T7, T8, P3, PZ, P4, P7, P8, PO3, PO4, O1, Oz, O2. See Figure 2 for a map of the EEG channels and their locations on the user's head.

In order to process this data, we first downsample the data from each channel to 100 Hz in order to reduce the raw amount of data. This 100 Hz signal is more than sufficient for our analysis since we then bandpass it to only retain the frequency band in which most useful EEG information is thought to reside, between 0.15 and 30 Hz. We do this using Finite Impulse Response (FIR) filters because they are inherently stable and computationally efficient.

Figure 3 shows an example of the response in one of our users with face and non-face stimuli. The response for face images, a strong N170 face-specific response is seen in the left image, data measured after a user has seen a face, but not in the right, in which they see non-faces. This is the purple line that protrudes out the bottom of the series at about 170 ms after stimulus presentation.

In order to exploit this response and others like it, the system uses a time window that is 100-300ms following stimulus presentation from some set of EEG channels. The system utilizes a recently developed spatial projection algorithm [12] designed for processing ERPs. This algorithm projects the response sequences from the multiple channels onto three maximally discriminative time series. We then use Regularized Linear Discriminant Analysis (RLDA), a supervised machine learning method, to classify the resulting features into mutually exclusive and exhaustive groups, namely the categories of interest [11].

While we show in this paper that this technique works relatively well, implementing and testing other machine learning techniques for such problems remains future work. Also, while we batch processed experimental results, the system is able to classify the signal in real-time once the model is built.

EXPERIMENT ONE

We conducted the first experiment to explore whether people were implicitly processing faces when they viewed a set of images, and whether we could use our system to detect this without them being aware. We limited this initial exploration to faces so the problem would be tractable.

Task

The task that users explicitly performed was simple. They viewed a series of images on the display and tried to memorize them. Each experimental block contained sequences of

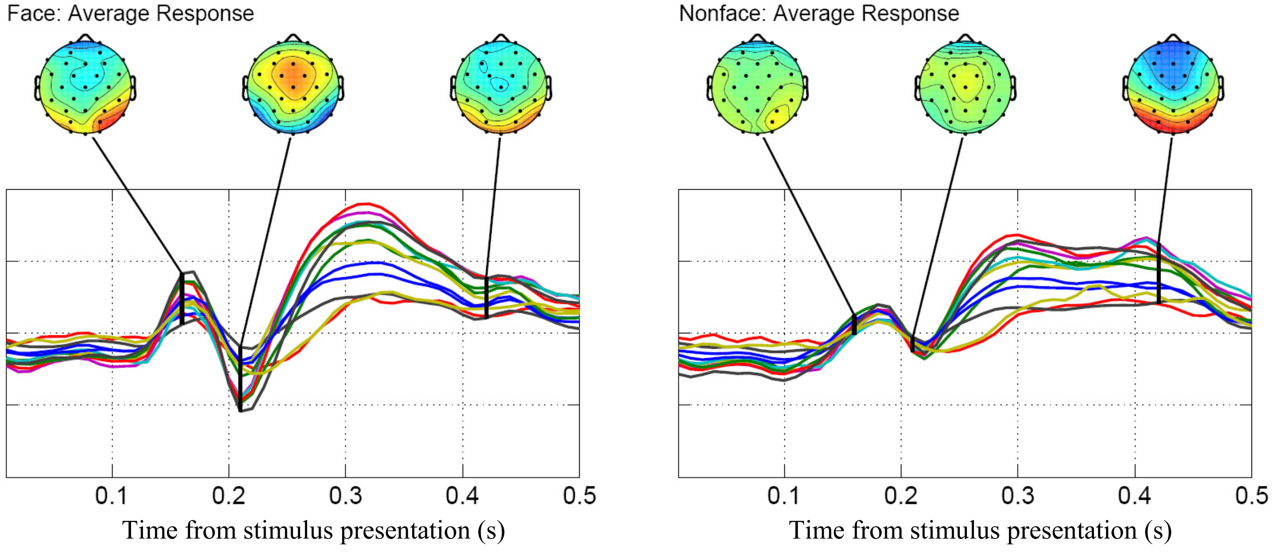

Figure 3: Average ERP responses to viewing faces (left) and non-faces (right) for one user with the controlled images. The time series for each channel are shown in multiple colored lines within each graph and the accompanying scalp plots show the spatial distribution of these signals at snapshots in time. In these plots red signifies higher activity. The N170 (at ~170 ms) face specific negative peak is evident in the face response and not the non-face one.

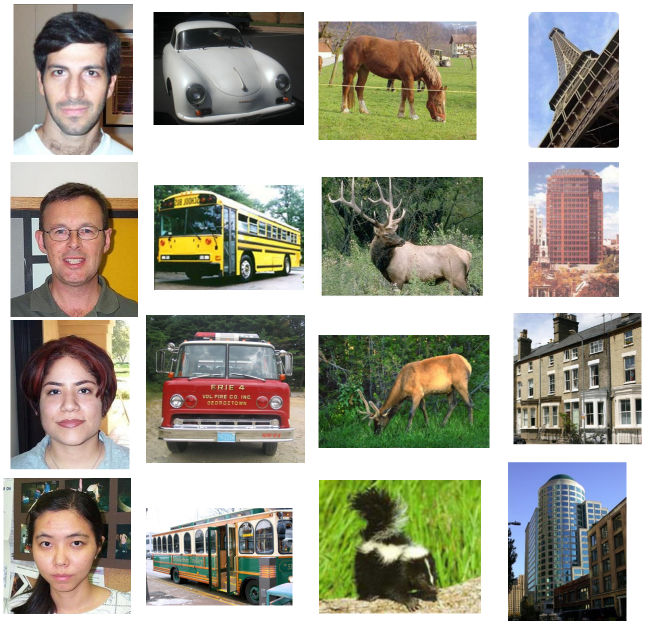

Figure 4: Controlled images taken from the Caltech-256 Object Category Dataset. The columns represent the face, vehicle, animal, and structure classes.

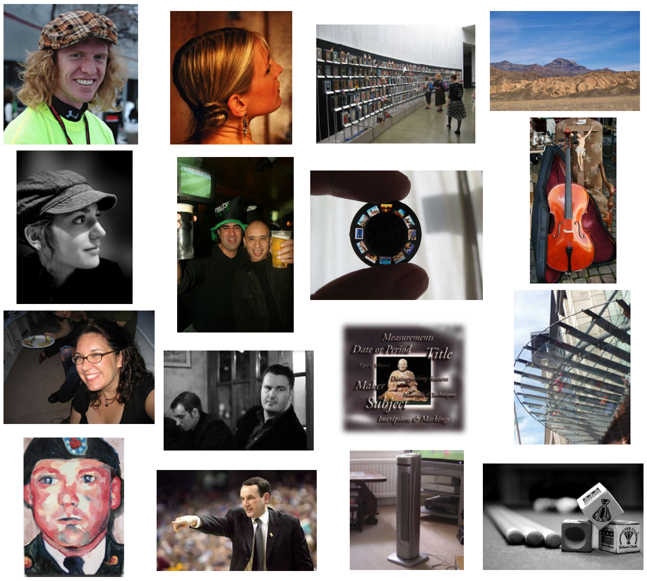

Figure 5: Ecologically valid real-world images downloaded from the web. Left two columns represent face images, right two represent non-faces.

50 images. At the end of each block, users saw six more images one at a time and had to decide if they had seen the image in that block. Three of these images were taken from the set the user had seen within that block and three were brand new images the user had never seen.

This task was really a distracter task and merely a method of getting users to look at the display. It specifically did not require users to explicitly decide whether or not each image was a face or not. To this end, we were also careful not to tell users anything about the face classification task, and users assumed we were studying human memory performance. We briefed them of the purpose of the experiment only at the very end.

Design and Materials

We were interested in exploring two basic conditions. First, we wanted to know how well we could classify implicit cognitive responses with carefully controlled stimuli. Each of the stimuli within the classes had similar image properties such as composition and color histograms. Second, we wanted to know how well the results would extend to less controlled and more ecologically valid stimuli. We used images taken randomly from the web.

Additionally, we wanted to know how quickly people could implicitly process these images and whether or not presentation time would affect our classification accuracy. This has direct implications on the efficacy of any system that is using these techniques to classify images. Each image was flashed for 150ms and images were presented one after another with some delay in between. We call this the inter-stimulus delay. We refer to the total time the image is displayed and the inter-stimulus time as the Presentation time. We picked three presentation times (500, 750, and 1000ms) based on prior work on using EEG to detect conscious object detection, suggesting that people could process these images in 200 to 600ms. This equated to inter-stimulus delays of 350, 600, and 850ms. All images were presented in the middle of the display in a box that was 400 pixels, or about 3 degrees of visual angle, large.

While we would have liked to perform a full factorial design, crossing stimulus-type with presentation time, we also needed to balance this with possible fatigue from staring at multiple flashing pictures for too long. Hence we chose four of the six possible conditions to test in this experiment. The four conditions we chose were the controlled stimuli with the three possible presentation times, and the ecologically valid stimuli with the middle 750 ms presentation time. We ran the experiment as a within-subjects design, with each user performing all four conditions.

We selected images for the controlled stimuli conditions from the Caltech-256 Object Category Dataset [8]. This data set includes 30607 images manually categorized into 256 object categories and is a benchmark test set for computer vision based object recognition algorithms. These images tended to be fairly homogeneous within categories and we thought would allow us to test the upper bounds of our ability to classify objects. Based on pilot studies, we selected several categories from this test set: faces (our object of interest), vehicles, animals, and structures. To broaden the non-face categories, we selected images for the latter three of these categories from multiple sub-categories within the test set. For example, animals came from the categories for dogs, cats, lions, and so on. For an example of the controlled images, see Figure 4.

For each block in the controlled stimuli conditions, we manually selected 200 images from this dataset, 50 from each category of interest. We produced stimuli sets for four blocks (800 images) in each condition. This led to a total of 2400 images in these conditions.

To generate images for the ecologically valid condition, we took the top results from a web-based image search using images.google.com with the search terms “faces” and “objects”. Three independent people labeled these results on a scale of 1 to 5, classifying the degree to which the image represented a face. A rating of 1 signified that the image was clearly not a face, nor did it contain anything that could be interpreted as such, and a rating of 5 indicated that the image was definitely a face. Before they labeled the images, they were shown example images in each of the categories so that they could label them somewhat consistently. We discarded any image that received user ratings deviating by more than a 1 point on this scale, and treated the 2, 3, and 4 ratings as distracter stimuli in our analysis. For an example of the stimuli shown in the ecologically valid conditions, see Figure 5. Note how the pictures are much less well defined than the controlled images in Figure 4.

For each block in the ecologically valid stimuli condition, we randomly chose 200 images from the sorted web images, 40 from each rating level. This led to 40 faces and 160 non-faces for each block. We generated enough stimuli for four blocks, but ran each of these blocks twice because we wanted to know if having a user look at a given image multiple times could boost classification performance.

As users performed the tasks, they were encouraged to take small breaks in between each block and a larger break between conditions.

Equipment

We collected EEG data using a Biosemi ActiveTwo system (www.biosemi.com), sampling at a rate of 2048 Hz. As seen in Figure 2, we placed electrodes at approximately evenly spaced locations on the scalp using the internationally accepted 10-20 system [14]. We did not control or eliminate any of the traditionally considered noise elements (e.g. 60 Hz power hum, etc.) found in the experimental environment, a research lab with running computers and other electrical equipment. This was because we wanted to simulate an environment that was likely both in an HCI lab, as well as in a real-world setting. Before beginning, the experimenter explained the EEG device and requested that users try to reduce unnecessary physical movements during the testing phases of the experiment. This, however, was not enforced.

Participants performed the tasks on a Compaq EVO N800c laptop with 2 GB RAM and an attached 18" NEC Multisync 1880SX LCD monitor running at a 1280×1024 pixel resolution. Each user sat at a comfortable distance from the monitor (about 30" or 76cm) and provided answers with the left and right arrow keys on the keyboard.

The task users consciously performed was mostly a distracter task to ensure that they were looking at the display. We collected the accuracy of answers and response times as dependent behavioral measures just to verify that they were indeed looking at the images.

Participants

Eight (3 female) student interns at an industrial research lab volunteered to participate in the experiment. The average age of users was 21, ranging from 20 to 22 years of age. None of the users reported any known neurological disorders, and 7 of the 8 users were right-handed. We also screened users for color blindness and required normal or corrected-to-normal eyesight. The experiment took about an hour and users were given a small gratuity for their time.

Evaluation Methodology

We ran the collected EEG data through our classifier system described above. We used 10-fold cross-validation accuracies as a measure of classifier performance. We randomly partitioned the labeled data into 10 groups and used a partition as the testing data for a classifier trained on the remaining partitions. The cross-validation procedure was repeated 10 times and the average of the runs is presented as the performance measure, or classification accuracy.

We also explored combining multiple user responses to a particular image in order to more accurately label it. We built a separate user-specific classifier based on individual users' responses, and used another RLDA classifier to combine each classifier's prediction. This creates a voting mechanism of sorts, allowing each viewing of the stimulus to contribute to an overall classification answer. We evaluated the performance of this compound-classifier in a method similar to the user-specific one.

Results

We present the results in several sub-sections. First, we present behavioral data, showing that users paid attention to the task and that performance did not deteriorate with time. Second, we describe the results when training our classification on the controlled stimuli and cross-validating with hold out sets also of controlled stimuli. Third, we describe results when training on the controlled stimuli and testing on the ecologically valid stimuli. This tests how well the controlled stimulus set was able to train the model to discriminate the factor of interest, faces. Fourth, we describe results when training and testing on ecologically valid stimuli, which we assert is the most likely case for real-world systems. We had initially projected that this would be noisier and hence less accurate than the controlled set, but we show that this is not really the case. Finally, we show preliminary results suggesting that classification might be improved if we show the same stimulus multiple times, either to the same user or to different users.

Behavioral Data

The stated task was for the user to decide whether a given image had already been shown in the immediately preceding block. We examined behavioral user responses to the questions following each block for correctness. There were twenty blocks per user, for a total of 120 questions. Of these, half had been previously seen and the other half were new. Users were able to respond correctly on average 81.1% of the time. This performance is significantly above chance and indicates that they were paying attention to the presented images. The average response time was 1.3s for correct responses, and 1.09s for incorrect responses.

Training and Testing on Controlled Stimuli

The first analysis that we ran with the EEG data was one in which we trained the system using our controlled stimuli and tested on hold out sets also taken from this set of stimuli. From three of our conditions, we looked at the three presentation times of 500, 750, and 1000ms. For these three times, the average classification accuracies were 69.2%, 72.5%, and 66.6%, with standard deviations across users of 7.1%, 5.0%, and 8.4% respectively.

These results suggest that it is possible to train a system to classify whether a user is looking at images of faces vs. non-faces even when the user is not trying to do that task and even with just a single presentation of the stimulus. Furthermore, the differences across presentation rates were small, suggesting that we can present many images in rapid succession. In future work we will attempt to push the lower bounds of this presentation rate to find out how fast we can present these images before the classification starts to deteriorate significantly.

Testing Real-World Stimuli on Controlled Stimuli Models

Next, we examined whether classifiers built on the controlled images would be successful in labeling our real-world images collected from the internet. Hence we applied classifiers trained on controlled face vs. non-face image EEG responses to 40 each of the ecologically valid face and non-face categories. The average classification for single presentations of 750ms across all users was 66.4%.

We see from this slight drop in accuracy that while it is possible to generalize from controlled images to ecologically valid stimuli, there is some degradation in the process. We speculate that this is due to the fact that the controlled images are fairly homogenous within each class, and thus do not form a sufficiently representative set of stimuli for training our classifier. In other words, the EEG response might, in addition to the face-specific nature of the image, also have information regarding lighting, color, and size that would result in an over-fitted classifier. We explore a small portion of this in more detail in experiment two, where we present results suggesting that we were not actually latching on solely to color characteristics.

Training and Testing on Real-World Stimuli

To examine whether training on a more diverse and representative data set would increase classification performance, we measured cross-validation error on real-world stimuli, with 40 examples each of the face/non-face categories. Here, the training and test sets are constructed entirely from the ecologically valid images, presented at a rate of 750ms. We argue that this provides the most realistic test of real-world system performance.

In this test, the cross-validation accuracy remained relatively constant for the single presentation case, at about 66.5%. However, when we averaged the responses across two presentations of the stimuli, we saw a boost in classification accuracy, increasing to 76%. This strongly suggests real-world systems should train models on as diverse a representative stimuli set as possible and that there may be opportunities for drastically increasing accuracy by collecting EEG data from multiple presentations of the stimuli.

Combining Information Across Users

There are multiple ways to collect EEG response data for the same image. We have described in the results presented above that having a user view an image multiple times may remove extraneous noise associated with single-presentations, and increase classification accuracy. However, another approach is to show the stimulus to multiple users for classification. This approach has the potential benefit of utilizing inter-user differences in order to get a more robust result.

We explored the effects of multiple redundant presentations by picking data from random subsets of users, which we used as input to the classifier. For each user, we constructed a separate classifier with their training data, as we did before. We then combined the outputs from these classifiers as the input to a separate RLDA meta-classifier that predicted the image's category label using only the individual classifier outputs. We averaged results over eight runs with different subsets of users to eliminate selection bias.

As can be seen in Figure 6, the classification accuracies increase significantly as we add more and more users, with performance reaching near certainty (98.3%) by the time 8 users have seen the images twice each. In the next experiment, we explore opportunities for dynamically deciding whether an image should be shown again, as well as whether it would benefit from being shown to the same user or would attain a better label if shown to different users.

EXPERIMENT TWO

We conducted a second experiment to extend the results from the first experiment and to explore: (1) the performance of our classification method on a larger more diverse population than 20-22 year old students; (2) the generality of the approach in classifying more than just faces vs. non-faces; (3) the tradeoffs associated with the ability to present stimuli multiple times, potentially to multiple people; and (4) the effects of varying the amount training data used.

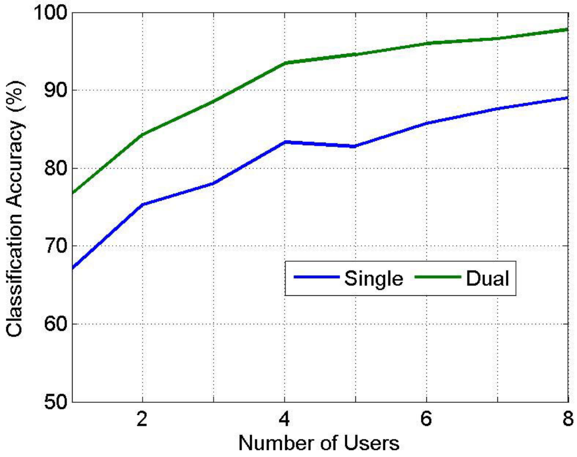

Figure 6: Classification accuracies showing the effects of varying number of users to which each image was shown. The two colored lines depict number of times the image was shown to each user.

Task

We used a similar task structure to that of the first experiment. However, instead of having users look at images and memorize them, we randomly inserted between three and ten images of butterflies within the sequences, and had users count the number of these seen within each block. We improved our distracter task because observations in the previous experiment suggested that the memory task was creating additional cognitive load as users continued to mentally rehearse seen images even as new ones came up.

Also, since users in the first experiment commented on the fatiguing nature of the rapidly flashing images, we decided to show the images for 250ms instead and to randomize the inter-stimulus delay between 500 and 750ms so as to eliminate predictable periodicity of the presentations.

Design and Materials

In this experiment we chose three different categories of images for which we hoped to discriminate: faces, taken from the set of images ranked 5 in the previous experiment, inanimate objects, previously ranked 1, and animals, randomly chosen from the Caltech 256 dataset.

We divided the experiment into two phases, a training phase and a testing phase. In the training phase, we presented users with three blocks of images, each containing 60 to 70 unique images from the three categories. In the testing phase, we chose 20 images from each of the three categories that did not already appear in the training set. While training images were presented exactly once during the experiment, we presented the test set of sixty images 10 times in ten different experimental blocks. Each presentation contained a random number of butterfly images and was presented in random order.

Equipment and Participants

We used the same equipment and setup as in the first experiment. We ran sixteen (5 female) volunteers who had not participated in the first experiment. The average age of users was 35, ranging from 20 to 58 years of age. None of the users reported any known neurological disorders or color blindness, and all had normal or corrected-to-normal eyesight. The experiment took about an hour and users were given a small gratuity for their time.

We excluded data from two users, one whose signal showed no discriminative power and another on whom the experimenter had trouble getting the EEG device to work and whose data we did not even begin to analyze. We report on the data collected from the remaining fourteen users.

Results

We utilized the same classification methods and evaluation methodology as in the first experiment. We built a classifier using the training data and tested with data from the testing phase. Behavioral results showed that user performed the butterfly counting task very accurately, miscounting only by an average of .16 per task, or approximately one miscount for every six blocks of images.

We present the results in several sections. First, we show the results of our classification across the various categories. Next, we describe how this was affected by repeated presentations, both within but also between users. Finally, we show how these results are minimally affected by the amount of training data collected and used.

Classifying Multiple Categories

We ran several analyses, performing each of the three 2-way classification as well as the 3-way one. As can be seen in the leftmost data point in Figure 7, the system was able to classify face vs. inanimate objects at a 75.3% accuracy, only slightly higher than that of the first experiment. We also show comparably high accuracies for classifying between the other classes, 71.6% for face vs. animals, 65.0% for animals vs. inanimate objects, and 55.3% for the 3-way classification, which was a more difficult classification task. The standard deviations for these classifications were 7.6%, 7.6%, 9.2%, 10.5%, respectively. This is encouraging as it implies that the mechanisms we are using are relatively robust to multiple classes of stimuli beyond just human faces, as shown in the first experiment. Exploring the full scope of this generality remains future work.

As a peripheral note, we also analyzed the effect that color might have on the classification accuracies. We found no significant performance differences across color and non-color images, suggesting that color histograms were not a distinguishing factor in this experiment.

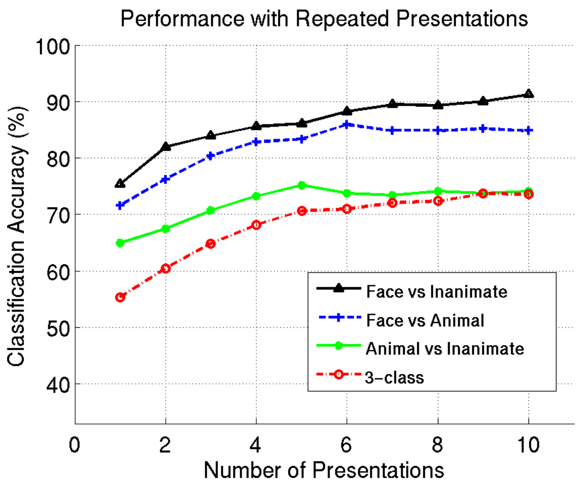

Figure 7: Relatively high accuracies for classification across multiple categories of images. This is especially true with repeated image presentation.

Effects of Repeated Presentations

In addition to the above analysis, we incrementally added data collected from repeated presentations of the same stimuli and combined their outputs to analyze how this would classification results. As can be seen in Figure 7, accuracies increase significantly as more repetitions are added, rising to 91.2%, 84.8%, and 74.1% for the 2-way classifications and 73.5% for the 3-way one, with just ten presentations. This suggests that multiple presentations, even to the same user, can reduce noise and add informational value.

As in the previous experiment, we also explored how repeated presentations to multiple users would affect classification accuracy. Results of this analysis for the multiple-user face vs. inanimate object classification can be seen in Figure 8. The other classifications yielded similar results and we omit discussion of them due to space constraints. As can be seen in this graph, adding data from multiple users clearly increases the performance of the classifier. While this might be partially attributed to repeated presentation of any form, analysis suggests that adding presentations to multiple users leads to better performance than adding the same number of presentations to one user. While our simple implementation works relatively well, building a classifier that is able to optimally combine the complementary data from various users remains future work.

Effects of Amount of Training Data Used

To evaluate the effect of amount of training data on classification performance, we used random subsets of training data to build the classifier and performed the same computations as above. We averaged results over ten runs with different subsets of training data to eliminate selection bias.

Results from this analysis, seen in Figure 9, suggest that reducing the amount of training data decreases performance. However, it is interesting to note that even with a very small training set such as 10 images, we are able to attain classification accuracies that seem to be above chance in all conditions tested. Loosely extrapolating from this graph, we also assert that we have not collected the amount of training data required for optimal performance and that designers of such systems should carefully consider the amount of training that they have users perform.

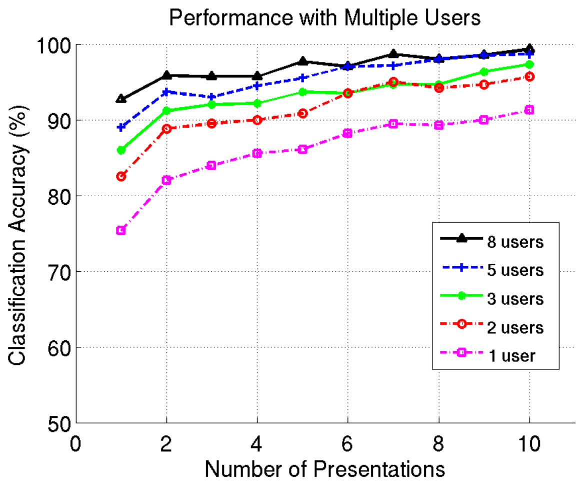

Figure 8: Accuracy increases with repeated presentations for face vs. inanimate object classification. The accuracies rise more when images are shown to multiple users as opposed to repeated to a single user.

CONCLUSION AND FUTURE WORK

In this paper, we introduced the concept of Human-Aided Computing, which proposes that we can use brain sensing technologies such as EEG in order to measure the outcomes of implicit processing done by the human brain. We have shown in two experiments that our classification system is able to classify between images of several classes with relatively high accuracy, even when the user is not aware of the task, and even with a single presentation of the image. We

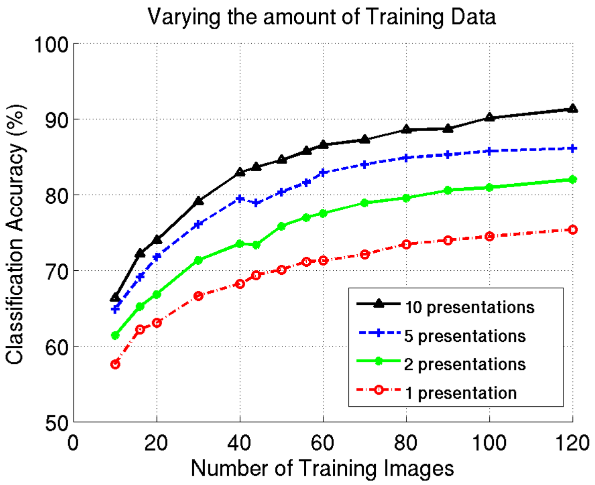

Figure 9: Accuracy falls as the amount of training data is reduced. However, even with only 10 images, classification for face vs inanimate objects remains above chance, especially with repeated presentations.

have further shown that we can drastically improve classification performance with multiple presentations of the same images, especially when we show them to multiple users.

This work represents only a first step towards our vision of a Human-Aided Computing system, and much future work remains. For example, we think it would be interesting to explore how the current results apply to pre-attentive implicit processing as well. In such a scenario, tasks could be placed in the attentive (e.g. visual or audio) periphery, not require explicit cognitive attention, allow the user to go about their primary tasks as they normally would, but get usefully processed by the brain and sensed by the system.

We would also like to expand the set of objects that we are able to classify, and in fact the set of tasks beyond image classification. We believe that this work will largely be driven by neuroscience findings, but hope that we can also contribute to fundamental understanding in that domain.

ACKNOWLEDGMENTS

We thank Ed Cutrell, Mary Czerwinski, Eric Horvitz, Rajesh Rao, and the Visualization and Interaction group at Microsoft Research for their support and discussions.

REFERENCES

1. Amazon.com. Mechanical Turk. http://www.mturk.com.
2. Anderson, D.P., Cobb, J., Korpela, E., Lebofsky, M., & Werthimer, D. (2002). SETI@home: An experiment in public-resource computing. *Communications of the ACM, 45(11)*, 56-61.
3. Broadbent, D.E. (1958). *Perception and communication*. Pergamon Press: London, UK.
4. Carmel, D., Bentin, S. (2002). Domain specificity versus expertise: factors influencing distinct processing of faces. *Journal of Cognition, 83*, 1-29.
5. Farwell, L.A., Donchin, E. (1988). Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography & Clinical Neurophysiology, 70(6)*, 510-23.
6. Fisch, B.J. (2005). *Fisch & Spehlmann's EEG primer: Basic principles of digital and analog EEG*. Elsevier: Amsterdam.
7. Gerson, A.D., Parra, L.C., & Sajda, P. (2006). Cortically-coupled computer vision for rapid image search. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 14(2)*, 174-179.
8. Griffin, G. Holub, A.D. Perona, P. (2007). The Caltech-256, *Caltech Technical Report*.
9. Grill-Spector, K. (2003). The neural basis of object perception. *Current opinion in neurobiology, 13*, 1-8.
10. Hariri, S., Parashar, M. eds. (2004). *Tools and environments for paralell and distributed computing*. Wiley Press: Hoboken, NJ.
11. Hastie, T., Tibshirani, R. & Friedman, J.H. (2001). The elements of statistical learning: Data mining, inference, and prediction. *Springer Series in Statistics.* Springer-Verlag: New York, New York.
12. Hoffman, U., Vesin, J-M, Ebrahimi, T. (2006). Spatial filters for the classification of event-related potentials. *In European Symposium on Artificial Neural Networks*, 47-52.
13. Itier, R.J., Taylor, M.J. (2004). N170 or N1? Spatiotemporal differences between object and face processing using ERPs. *Cerebral Cortex, 14*, 132-142.
14. Jasper, H.H. (1958). The ten-twenty electrode system of the international federation. *Electroencephalography and Clinical Neurophysiology, 10*, 371-375.
15. Johnson J.S., Olshausen, B.A. (2005). The earliest EEG signatures of object recognition in a cued-target task are postsensory. *Journal of Vision, 5(4)*, 299-312.
16. Koch, C., Tsuchiya, N. (2006). Attention and consciousness: Two distinct brain processes. *Trends in Cognitive Sciences, 11(1)*, 16-22.
17. Larson, S.M., Snow, C., & Pande, V.S. (2003). Folding@Home and Genome@Home: Using distributed computing to tackle previously intractable problems in computational biology. *Modern Methods in Computational Biology (ed. R. Grant)*. Horizon Press.
18. Lee, J.C., Tan, D.S. (2006). Using a low-cost electroencephalograph for task classification in HCI research. *In 19th ACM Symposium on User Interface Software and Technology*, 81-90.
19. Marszalek, Z.J.M., Lazebnik, S., & Schmid, C. (2006). Local features and kernels for classifcation of texture and object categories: A comprehensive study. *International Journal of Computer Vision, 73(2)*, 213-238.
20. Miller, G.A. (1987) *Psychology: the science of mental life*. Penguin: Harmondsworth, UK.
21. Posner, K.R. (1978). *Chronometric explorations of mind*. Lawrence Erlbaum Associates: Hillsdale, NJ.
22. Rossion, B., Gauthier, I., Delvenne, J.-F., Tarr, M.J., Bruyer, R., & Crommelinck, M. (1999). Does the N170 occipito-temporal component reflect a face-specific structural encoding stage? *Poster at Object Perception and Memory 1999.*
23. Velmans, M. (1991). Is human processing conscious? *Behavioral and Brain Sciences, 14*, 651-726.
24. von Ahn, L. (2006). Games with a purpose. *IEEE Computer Magazine, 39(6)*, 92-94.
25. von Ahn, L., Dabbish, L. (2004). Labeling images with a computer game. *In CHI 2004 Conference on Human Factors in Computing Systems*, 319-326.