Feature Selection for Gaze, Pupillary, and EEG Signals Evoked in a 3D Environment

[Extended Abstract]

David C. Jangraw
Columbia University
530 West 120th Street
New York, NY 10027
dcj2110@columbia.edu

Paul Sajda
Columbia University
530 West 120th Street
New York, NY 10027
psajda@columbia.edu

ABSTRACT
As we navigate our environment, we are constantly assessing the objects we encounter and deciding on their subjective interest to us. In this study, we investigate the neural and oculair correlates of this assessment as a step towards their potential use in a mobile human-computer interface (HCI). Past research has shown that multiple physiological signals are evoked by objects of interest during visual search in the laboratory, including gaze, pupil dilation, and neural activity; these have been exploited for use in various HCIs. We use a virtual environment to explore which of these signals are also evoked during exploration of a dynamic, free-viewing 3D environment. Using a hierarchical classifier and sequential forward floating selection (SFFS), we identify a small set of features that can be used to distinguish targets from distractors in the virtual environment. The identification of these features may serve as an important factor in the design of mobile HCIs.

Categories and Subject Descriptors
H.1.2 [Information Systems]: User/Machine Systems; I.5.2 [Pattern Recognition]: Design Methodology

General Terms
Human Factors

Keywords
HCi, gaze control, pupil dilation, EEG, multimodal classifier, feature selection, virtual environment

1. INTRODUCTION
As mobile technology becomes more and more prevalent, physiological signals are becoming easier to record in a mobile setting. This includes wearable electroencephalography (EEG) headsets [17], eye trackers [22], and heart monitors [14]. These mobile capabilities raise the intriguing possibility of using a human-computer interface (HCI) to identify objects that are of subjective interest to a user based on the physiological signals she produces when she encounters them in her everyday life (e.g., while walking or driving). Such "naturally evoked" signals could be combined across modalities with a multimodal or "hybrid" HCI classifier [30, 23, 2] and used to label objects of interest without any conscious effort from the user. These labels could then be used to provide the user with effortless assistance, such as triggering a video recording of subjectively interesting events or recording a map of subjectively interesting places.

With the proliferation of research in hybrid HCI has come an abundance of potentially useful features. Studies have found multimodal feature sets that provide complementary information to certain HCIs (e.g., those classifying a user's affect or intent to manipulate a 3D display), and they have established strategies for fusing multimodal information [30, 12]. Recent brain-computer interface (BCI) research has also considered multimodal feature sets, combining EEG with heart rate to classify motor imagery [23]. Other studies have simply identified neural and oculair features that have been available for non-invasive monitoring, the number of features that can be extracted from these modalities is limitless.

But as some researchers have noted [30, 7], including too many features in a hybrid classifier could result in high computational costs or require an impractical quantity of training data. These constraints are especially relevant in a mobile setting, where training data may contain artifacts and battery life and processor speeds are limited. The success of a hybrid mobile HCI, then, depends heavily on the problem of feature selection. But the fusion of neural and oculair data naturally evoked by objects of interest, and the selection of a robust multimodal feature set, are so far largely unexplored.

In this paper, we extract numerous candidate features from the EEG, eye position, and pupil size data recorded from subjects as they search for objects of interest in a dynamic 3D environment under free-viewing conditions. We then employ sequential forward floating selection (SFFS) [26] to efficiently select features useful for classification. In this way, we identify a small set of features that can be integrated to effectively classify naturally evoked signals in a realistic environment. This represents an important step towards the development of a high-functioning mobile HCI.
2. EXPERIMENT

2.1 Virtual Environment

The objective of our analysis was to identify physiological signals useful for classification in a realistic, free-viewing scenario, taking a step away from a typical laboratory setting and towards a mobile one. Our experimental paradigm was therefore designed to present randomized stimuli in a realistic but consistent setting, keeping stimuli relatively controlled while preserving most aspects of natural viewing.

A virtual environment was created using Unity 3D software (Unity Technologies, San Francisco, CA) consisting of a grid of city streets. The subject’s viewpoint was automatically navigated up and down the streets as if riding in a car. Each block had an alley on either side, and a billboard object was placed in one of the alleys so that it gradually became visible as the subject passed it. Each billboard contained a randomly selected image from one of four categories from the CalTech-101 database [8]. The subject was instructed to count objects of one category (“targets”) and ignore the others (“distractors”) but make no physical response. Each subject viewed 20 objects per session and performed 13-15 sessions.

To keep the subjects focused and default their gaze to the center of the screen, they were asked to press a button when the “leading car” in front of them put on its brakes. The subjects were allowed to move their eyes freely during the task; thus, any changes in viewing patterns between targets and distractors were naturally evoked by the objects and not produced for the purposes of interacting with the HCI. A screenshot of the virtual environment is shown in Figure 1.

2.2 Data Collection

10 subjects participated in the experiment (ages 19-42, 3 female, 1 left-handed). All subjects reported normal or corrected-to-normal vision. Informed consent was obtained in writing from all participants in accordance with the guidelines and approval of the Columbia University Institutional Review Board.

Data were collected in a tightly controlled laboratory setting, representing not the current state of mobile recording, but our optimistic prediction of future signal quality. Monocular eye position and pupil area data were collected at 1000 Hz using an EyeLink 1000 eye tracker (SR Research, Ontario, Canada). A 9-point eye tracker calibration was performed before the first session. A validation was performed before each subsequent session, and if the match was unsatisfactory the eye tracker was re-calibrated. EEG data were collected at 1000 Hz from 77 electrodes using a Sensorium DBPA-1 amplifier (Sensorium Inc., Charlotte, VT). Eye tracker and EEG data were synchronized by sending a parallel port pulse from the EyeLink computer to the EEG system every 2 seconds. The records of the times at which each pulse was sent and received were used to realign the data during analysis. More information on the collection of data in the virtual environment can be found in [13].

3. FEATURE EXTRACTION

3.1 Eye Position Features

The EyeLink online parser was used to identify the times and positions at which saccades and fixations started and ended. The standard “cognitive configuration” was used, which is robust to noise but ignores small saccades. All further analysis was performed using MATLAB software (The MathWorks, Inc., Natick, MA). To combat drift across sessions, a post-hoc drift correction was performed so that the subject’s median eye position during each session fell on the center of the screen. Fixations falling within 100 pixels of the object’s on-screen bounding box were considered fixations on the object (this cutoff was chosen based on a histogram of all fixations’ distances to the object). The subject’s first fixation on an object (which we call “fixFirst”) was used to time-lock the EEG and pupil size data for that trial.

Various features were extracted from the eye position data that have been identified as possibly indicative of interest or mental effort during reading or visual search tasks [27, 11, 1]. To do this, we first identified various eye position events: the last fixation before fixating the object (“fixBefore”), the first and last fixations on the object (“fixFirst” and “fixLast”), and the first fixation after leaving the object (“fixAfter”). The durations of these fixations (denoted by the prefix “dur_”) and their distances from the object’s center on the screen (“dist_”) were used as candidate features. Next, we identified the saccades to and away from the object (“sacTo” and “sacAway”). The sizes of these saccades in pixels (“size_”) were also used as features. We also found the mean duration of all fixations on the object (“dur_fixMean”), the number of fixations on the object (“nFix”), the time between the object’s appearance on the screen and the start of fixFirst (“latencyFixFirst”), and the total time the subject spent fixating the object (“dwellTime”); and we included these as features. Data from a toy trial illustrating these candidate features is shown in Figure 2(a).

3.2 Pupil Dilation Features

Similarly, we extracted features from the pupil area data that have been identified as correlates of interest or task-relevance during reading, visual search and oddball tasks [25, 20]. We first normalized the data by dividing by the subject mean across all sessions. We then epoched the data by extracting the first 3000 ms of data after the onset of fixFirst (the shortest time between saccades to objects was 3272 ms) and subtracted a baseline of -1000 to 0 ms. The resulting pupil dilation was binned into six 500-ms bins, a
width chosen based on the slow modulations of pupil area observed in previous studies [3, 20]. The average pupil dilation and average derivative of pupil dilation in each bin were used as features, denoted by “PD(X)” and “PD’(X)”, respectively (where X is the time of the bin center, in ms). The max pupil dilation in the epoch (“PDmax”), and the time (relative to fixFirst) when this dilation was reached (“latency_PDmax”), were also used as features. Data from a toy trial illustrating these features is shown in Figure 2(b).

3.3 EEG Features
The EEG data was processed using EEGLAB toolbox [6]. It was band-pass filtered from 0.5 to 100 Hz, notch filtered at 60 Hz, and downsampled to 250 Hz. Noisy electrodes were removed by visual inspection. Epochs were extracted of the first 1000 ms after the onset of fixFirst, and a post-saccadic baseline of 0 to 100 ms was subtracted as in [10]. HEOG and blink components were removed using the maximum difference and maximum power methods described in [21]. Trials containing other large artifacts were detected and interpolated or removed as in [15]. To reduce the feature space and avoid rank deficiency issues during classification, Independent Component Analysis (ICA) was performed on the top 20 principal components of this data [29]. The activations of the resulting 20 ICs were used as features in the classifier.

4. CLASSIFICATION

4.1 Hierarchical Classifier
A hierarchical classifier was adapted from the hierarchical discriminant component analysis (HDCA) described in [9, 24] to include both EEG and ocular features. The EEG data from 100 to 1000 ms after the onset of fixFirst were separated into 9 100-ms bins. Fisher linear discriminant analysis (FLDA) was used to find a set of spatial weights for each bin. These weights were applied to a separate set of evaluation trials to determine a set of EEG features for the second-level classifier (one for each trial and time bin). The use of evaluation trials is essential to avoid overly weighting the EEG features: spatial weights learned on the training data may over-fit the training data, but not the evaluation data.

To scale the ocular features similarly to the EEG, FLDA was performed on each feature, and each feature was divided by its standard deviation across all training and evaluation trials. This process results in 37 scaled features: 9 EEG features (one for each time bin), 14 eye position features (those seen in Figure 2(a)), and 14 pupil dilation features (those seen in Figure 2(b)).

The second-level, “cross-feature” classifier used these scaled features as input to logistic regression, which assigned a set of cross-feature weights to maximize an objective function (including an L2 regularization term to combat over-fitting). These weights were applied to a separate set of testing trials to get a single “interest score” for each trial. A successful classifier will produce higher interest scores for targets than for distractors. The area under the ROC curve (AUC) was calculated as a threshold-free figure of merit, since approximately 25% of objects were targets. Training, evaluation, and testing trials were selected using nested 10-fold cross-validation to prevent over-fitting.

4.2 Feature Selection
To determine a set of features that yields good classification across subjects, we used a method known as sequential forward floating selection (SFFS) [26]. The algorithm searches for an optimal feature set by sequentially adding and removing the single feature that results in the highest classification score (in this case, the mean cross-validated AUC score across subjects). Our analysis uses the modified SFFS algorithm outlined in [31]. The algorithm produces a record of its predicted optimal feature set at each possible “set size” (the number of features included, from 1 to the total number of features).

To gauge the success of feature selection, we performed another 10-fold cross-validation. SFFS was used to select a set of features using 9/10 of each subject’s data. Using the set of features identified by SFFS at each set size, the hybrid classifier was trained on the same 9/10 of data and used to produce a set of cross-feature weights. These cross-feature weights were then used to classify the remaining 1/10 of each subject’s data.
the other bins for classification. Distance measures from the end of the fixation on the object (dist\textsubscript{fixLast}, sac\textsubscript{sizeAway}, dur\textsubscript{fixAfter}) were more useful than those from the beginning of object fixation. DwellTime was the most useful timing-based measure. Pupil dilation measures from the middle of the epoch (bins centered at 750, 1250, and 1750 ms) tended to be more useful than the others. The number of fixations and pupil dilation derivative features do not appear to be useful for classification.

The set of features chosen at our optimal set size is of particular interest, as it suggests a small but robust feature set for use in future HCIs. The SFFS selection rates of the features at set size = 9 are shown in Figure 4(b). Seven features had 100% use rates at this set size: EEG(450), EEG(650), dur\textsubscript{fixLast}, size\textsubscript{sizeAway}, dist\textsubscript{fixAfter}, dwell\textsubscript{Time}, and PD(1750). Other features with use rates above zero (i.e., features included in some but not all folds) are dur\textsubscript{fixMean}, dur\textsubscript{fixLast}, PD(750), and PD(1250).

### 6. DISCUSSION

Our study indicates that an HCI system classifying physiological signals of interest evoked by a realistic environment should integrate features derived from EEG, eye position, and pupil dilation. SFFS results suggest that a successful hybrid classifier need only include a few important features: EEG features from 450-650 ms, pupil dilation features from 750-1750 ms, the subject’s dwell time on the object, and eye position features from the end of the fixation on the object. The plateau in both training and testing AUC scores at larger set sizes (seen in Figure 3) indicates that additional features do not tend to contribute significantly to classification across subjects.

The features selected by SFFS at our optimal set size are largely consistent with previous findings. The selected EEG time bins (450-650 ms) correspond with the peak of the P300, a correlate of rare, task-relevant stimuli reported in numerous studies [18, 5, 15]. Early features sometimes used in the HDCA [24] were generally not selected here, perhaps due to the post-saccadic baseline used in our free-viewing task. The selected pupil dilation time bins (750-1750 ms) correspond with the peak time of target-related pupil dilation in visual oddball tasks [20]. Dwell time is a strong correlate of interest that has been used successfully in gaze-controlled HCIs [28, 16].

The selection of three eye position features from the end of the fixation on the object may relate to the fact that relatively complex, natural images were used as stimuli (in contrast with most visual search studies [15, 4]), and the subject typically had to make multiple fixations on the object to decide on its identity. This would help explain why properties of the last, and not the first, fixation on the object were more likely to be influenced by the task-relevance of the object.

The fact that these eye position features are closely related to one another (see Figure 2(a)) did not disqualify them from selection by SFFS. This could indicate that a combination of these features is more robust to noise than any one alone, or that different features are effective for different subjects. To shed some light onto this question, we performed a separate SFFS analysis with leave-one-subject-out (LOSO) cross-validation. That is, a feature set was learned from all but one subject using SFFS, and then a set of subject-specific weights across those features was learned...
for the remaining subject. LOSO AUC values were similar (in fact, slightly higher on average across all set sizes), indicating that feature selection was relatively robust across subjects.

The study outlined here is, of course, not equivalent to the mobile setting we endeavor to move towards. In a real-world scenario, subjects might view multiple objects within the time span used by the classifier. Extracting and classifying such overlapping signals remains a challenge for future research. The use of a virtual environment allowed us to standardize presentation and lighting conditions, and to assume knowledge of each object’s location in the scene. In a mobile scenario, gathering the same prior information would require rapid image segmentation across a wide array of objects and outdoor scenes, an outstanding challenge in computer vision. Future work could endeavor to replace or aid such object detection using the user’s fixation patterns.

7. CONCLUSIONS
We have shown that signals naturally evoked by objects of interest can be integrated across modalities to effectively discriminate them from distractors in a free-viewing, realistic environment. We have developed a hybrid classifier capable of combining these signals in a principled way. In this study, we used SFFS to identify a small set of features that can be used to produce high levels of classification across subjects. Future work will explore the integration of these features into a fully functional mobile HCI system.

8. ACKNOWLEDGMENTS
The authors would like to thank Ansh Johri for his work refining the virtual environment.

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-10-2-0022. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

9. REFERENCES