HHS Public Access

Author manuscript

IEEE Trans Biomed Eng. Author manuscript; available in PMC 2018 January 01.

Published in final edited form as:

IEEE Trans Biomed Eng. 2018 January ; 65(1): 43–51. doi:10.1109/TBME.2017.2693157.

EEG-based Affect and Workload Recognition in a Virtual Driving Environment for ASD Intervention

Jing Fan [Student Member, IEEE],
Electrical Engineering and Computer Science Department, Vanderbilt University, Nashville, TN, USA

Joshua W. Wade,
Electrical Engineering and Computer Science Department, Vanderbilt University, Nashville, TN, USA

Alexandra P. Key,
Vanderbilt Kennedy Center for Research on Human Development and Department of Hearing and Speech Sciences, Vanderbilt University, Nashville, TN, USA

Zachary E. Warren, and
Treatment and Research Institute for Autism Spectrum Disorders, Pediatrics, Psychiatry and Special Education, Vanderbilt Kennedy Center, Nashville, TN, USA

Nilanjan Sarkar [Senior Member, IEEE]
Mechanical Engineering Department, Electrical Engineering and Computer Science Department, Vanderbilt University, Nashville, TN, USA

Abstract

objective—To build group-level classification models capable of recognizing affective states and mental workload of individuals with autism spectrum disorder (ASD) during driving skill training.

Methods—Twenty adolescents with ASD participated in a six-session virtual reality driving simulator based experiment, during which their electroencephalogram (EEG) data were recorded alongside driving events and a therapist's rating of their affective states and mental workload. Five feature generation approaches including statistical features, fractal dimension features, higher order crossings (HOC)-based features, power features from frequency bands, and power features from bins (Δf = 2 Hz) were applied to extract relevant features. Individual differences were removed with a two-step feature calibration method. Finally, binary classification results based on the k-nearest neighbors algorithm and univariate feature selection method were evaluated by leave-one-subject-out nested cross-validation to compare feature types and identify discriminative features.

Results—The best classification results were achieved using power features from bins for engagement (0.95) and boredom (0.78), and HOC-based features for enjoyment (0.90), frustration (0.88), and workload (0.86).

Correspondence to: Jing Fan.

Conclusion—Offline EEG-based group-level classification models are feasible for recognizing binary low and high intensity of affect and workload of individuals with ASD in the context of driving. However, while promising the applicability of the models in an online adaptive driving task requires further development.

Significance—The developed models provide a basis for an EEG-based passive brain computer interface system that has the potential to benefit individuals with ASD with an affect- and workload-based individualized driving skill training intervention.

Index Terms

Virtual reality-based driving simulator; affective computing; mental workload recognition; electroencephalogram

I. Introduction

AUTISM spectrum disorder (ASD) is a neurodevelopmental syndrome that affects an estimated 1 in 68 children in the US [1] and is the fastest-growing developmental disability. Primary symptoms of ASD include deficits in social interaction, language and communication skills, and restricted, repetitive behaviors [2]. In addition to these core deficit areas, recent evidence suggests that adolescents and young adults with ASD have difficulty in learning safe driving skills [3–5]. In particular, compared with their typically developed peers, individuals with ASD demonstrated unsafe gaze patterns and higher levels of anxiety when operating a driving simulator [6], [7], responded slower during steering, identified fewer social hazards, and showed problematic multi-tasking ability [8], [9]. In the US, driving plays a critical role in everyday life and is essential for achieving adult independence. Given the heterogeneity and developmental nature of ASD [10], effective driving interventions tailored to specific individuals are needed for this population.

While virtual reality (VR)-based intervention for teaching social skills to children with ASD has been investigated in recent years due to various advantages of VR [11], [12], exploration into VR-based driving skill training for adolescents with ASD is only beginning to emerge [7], [13]. VR provides a safe driving skill training environment that can be designed to optimally engage individuals with ASD. Studies of game-based learning environments have argued for the importance of combining game design with flow theory to achieve optimal experience and enhance learning [14], [15]. Flow theory asserts that optimal experience is gained when the challenge level matches the skill level of a player [16]. Hence, by measuring the affective states and the mental workload of individuals with ASD and purposefully adapting the VR-based driving intervention system to keep them in the flow, the VR-based driving intervention may be optimally impactful. As a first step to designing such an individualized intervention system, in this study, an electroencephalogram (EEG) sensory modality was integrated into a VR-based driving simulator to build models for recognizing several affective states and the mental workload of individuals with ASD when they performed driving tasks. The goal of this research is to demonstrate that EEG-based affect and mental workload recognition is feasible during driving in a VR-based simulator so that in the future such an ability can help individualize the training.

In recent years, there has been an increasing interest in developing EEG-based passive brain computer interface (BCI) applications to enrich human-machine interaction. Kohlmorgen et al. trained an individualized mental workload detector using EEG in a real world driving scenario. The workload detector was then applied in real time to switch off the secondary task in the case of high workload [17]. Wang et al. proposed an online closed-loop lapse detection and mitigation system that continuously monitored a driver's EEG signature of fatigue based on EEG spectra, and delivered warnings accordingly during an event-related lane-keeping task using a VR-based driving simulator [18]. Dijksterhuis et al. classified the mental workload of drivers with varying speed- and lane-keeping demand by applying common spatial pattern and a linear discriminant analysis algorithm, again with a driving simulator [19]. Compared with mental workload, affective states are less studied in driving because affective states are not directly related to the safety-critical aspect of driving. Instead, most studies focused on drivers' states such as fatigue, drowsiness, stress, and alertness. Nonetheless, in learning and intelligent systems, EEG-based engagement indices and emotional states recognition have been evaluated and studied by many researchers [20–24]. Various feature generation approaches as well as machine learning algorithms have been applied to improve the reliability of EEG-based affective states recognition [25], [26].

In addition to a paucity of research on EEG-based affective states recognition for driving, the research to date has tended to focus on healthy adults rather than on individuals with greater potential for unsafe driving. Differences in EEG activity between individuals with ASD and their typically developed peers have been well documented by researchers [27], [28]. Given that driving is a necessary skill for independent living in the US and that individuals with ASD demonstrate a pattern of unsafe driving habits, there is a need to understand how driving skill training can be imparted to these individuals in a safe and flexible environment such as in a VR-based simulator. We believe that such training will be more effective if the system can tailor individual learning experiences based on their affective states and mental workload to accommodate for individual differences inherent in this spectrum disorder.

The primary contributions of this paper are: a) an experimental design to generate EEG data from adolescents with ASD during realistic VR-based driving tasks; b) development of a two-step feature calibration method to allow group-level training. This will dramatically reduce the training sessions needed compared to individualized model training; c) systematic evaluation of feature generation approaches to demonstrate the possibility of group-level affect and workload recognition based on EEG data; and d) systematic evaluation of feature and electrode usage to identify discriminative features. Together they provide a proof of concept that such EEG-based recognition could be useful to individualize ASD intervention. Although existing feature generation methods were applied in the current work, the analyses on EEG data collected from real world tasks with ASD population were not reported in the literature. Such analysis is needed prior to designing an EEG-based BCI for individualized ASD intervention. This paper substantially extends our earlier short conference paper [29] by incorporating rigorous methodology, additional data, and extended results and discussion.

The paper is organized as follows. Section II describes the VR-based driving simulator and data acquisition modules. Section III presents the methodology used to systematically

analyze the EEG signals. Classification and feature selection results are reported in Section IV. In the remaining Sections, we discuss the results and summarize the major findings and significance of the work.

II. System Description and Data Acquisition

The VR-based driving system was comprised of a VR driving module and four data acquisition modules, which were used to record EEG, peripheral physiology, eye gaze, and observer rating data (Fig. 1). The VR driving module consisted of two components: a virtual driving environment rendered from the viewpoint of the driver's perspective and a Logitech G27 driving controller for intuitive control of the virtual vehicle via a steering wheel and a pedal board.

A. Virtual Driving Environment

The virtual driving environment was created using CityEngine and Autodesk Maya modeling software. The model offered roughly 120 square miles of diverse terrain and provided foundation for enough unique roadways to design hours of driving tasks. Trees, houses, and skyscrapers populated the virtual world, and traffic lights, pedestrians, and various automobiles were used to simulate a bustling city. The roadways included one- and two-way streets as well as an eight lane highway encircling the entire city. Driving tasks, or *trials*, were designed to utilize each of these particular aspects of the city. Four categories of trials were implemented: turning, merging, speed-maintenance, and laws. Turning trials involved either a left or right turn at an intersection; merging trials included lane changes, overtaking vehicles, and exiting/entering highways; speed-maintenance simply dealt with adjusting speed as appropriate to the specific situation (e.g., highway or school zone speed changes); and laws trials included scenarios in which the driver must obey important road laws, for example, yielding to pedestrians and stopping at stop signs. In all, 144 trials were created.

Eight trials were grouped together in a sequence to create an *assignment*. Subjects attempted to complete an assignment with as few errors as possible. A sufficient number of errors meant the failure and termination of the assignment. Trial errors were monitored by the system for a wide variety of possible offenses. These included – to list only a few – running red lights or stop signs, wrong turns, excessive speed, vehicle collisions, driving in the wrong lane, and failing to move out of the way of emergency vehicles. Errors detected by the system resulted in the subject's virtual vehicle being reset to the start of the unsuccessful trial. Resets were accompanied by matching audio and text feedback indicating what went wrong and how to avoid the error moving forward. During trials, the system's built-in navigation system directed drivers – visually and with audio (e.g., "left turn ahead" and "turn left") – towards their destinations. In addition to the navigation system, the subject's perspective included a first-person view of the road, speedometer, turn signal indicators, steering wheel, side view mirrors, and a score shown as text to indicate progress.

In order to modulate the level of challenge faced by the subjects, a range of difficulty levels were introduced. Six different difficulty levels were designed. Several parameters were used to configure the level of difficulty presented by the system: number of vehicles on the road,

aggressiveness of other drivers, visibility, weather conditions, and responsiveness of the input device. The easiest level of difficulty was intended to be effortless for most drivers, whereas the hardest level of difficulty was intended to be overly challenging; the other difficulty levels were interpolations between these extremes. These difficulty levels were approved by trained ASD clinicians.

B. Data Acquisition

We used a 14-channel wireless Emotiv EPOC neuroheadset (www.emotiv.com) to record EEG signals from locations AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, and AF4, defined by the 10–20 system of electrode placement [30]. The reference sensors were placed at locations P3 and P4. The bandwidth of the headset was 0.2–45 Hz and the sampling rate was 128 Hz. We modified an existing EEG data acquisition application to log EEG signals, sensor contact quality, and driving event messages received from the VR driving module. Driving event messages – assignment start/end messages in particular – were used to align EEG signals with driving tasks and observer rating data. Subjects' eye gaze data and physiological data were collected using a Tobii X120 eye tracker (www.tobii.com) and a Biopac MP150 physiological data acquisition system (www.biopac.com), respectively.

We relied on an experienced therapist to report the ground truth of subjects' affective states as well as mental workload during driving. The rating categories employed were engagement, enjoyment, frustration, boredom, and task difficulty. In the context of computer-based learning environments, affective states of engagement, enjoyment, frustration, and boredom have been identified to capture useful learning experience across different learning situations and learners [31]. Therefore, we chose to build a model of affect that was able to recognize these four affective states. The last rating category, task difficulty, was adopted to represent subjects' mental workload. Mental workload characterizes the demands imposed on human's working memory by tasks. It has three attributes, which are mental load, mental effort, and task performance [32], [33]. Task difficulty contributes to mental load. More difficult tasks impose higher demands on the limited mental resources for information processing. Several works have established correlation between perceived task difficulty and mental workload [23], [34]. Mental effort is related to the characteristics of subjects, such as their driving experience. Since the mental effort varies over time and is different among subjects, perceived task difficulty, as rated by an experienced therapist, was adopted as the ground truth for subjects' mental workload. In the discussion section, we demonstrate a strong linear correlation between perceived task difficulty rating and task performance, which justified our method of acquiring subjects' mental workload.

Observer rating data were logged based on a 0–9 continuous rating scale, where larger ratings indicated a higher intensity. At the end of an assignment, which usually took five minutes, the therapist was prompted to provide a summary rating for each of the categories on subjects' overall states.

C. Subjects

This study was conducted with the approval from the Vanderbilt University Institutional Review Board. Twenty subjects (19 males, 1 female, mean age: 15.29 years) with a clinical diagnosis of ASD took part in the study. One subject had a driver's license and three subjects had driver's permits. Their ASD assessment results and IQs are reported in Table I. Subjects attended six sessions on different days. During each visit, spanning approximately 60 minutes, they completed three preselected assignments. We measured their EEG response before the session for a three minutes baseline period and during the session.

III. Methods

The continuous rating data were transformed into binary classes, low intensity class and high intensity class, for building models of affect and workload. As a first step we chose to develop binary classifiers. With more data and experience, multiclass classifiers will be developed in the future. The thresholds used for the transformation were chosen by the therapist before conducting the driving experiment. For each category, if the rating score was less than the threshold, the corresponding assignment was labeled as low intensity class, otherwise it was labeled as high intensity class. The thresholds for engagement, enjoyment, boredom, frustration, and difficulty were 6, 6, 2, 2, and 5, respectively.

A. Signal Preprocessing

Out of the 120 sessions (20 subjects × 6 visits), raw EEG data from 111 sessions were processed. Inevitable and unforeseen events, such as subjects selecting the wrong assignments, or the system being restarted due to eye tracker failure, led to the loss of some data. The spikes in EEG data were first removed by slew rate limiting. Then a 0.2–45 Hz bandpass filter was applied. Filtered data from baseline and each assignment were segmented into one-second epochs with 50% overlap. Corrupt epochs were identified and removed, and eye movement and muscular artifacts were corrected automatically. A detailed description of signal preprocessing procedure can be found in our previous paper [29].

For both baseline and assignment, EEG data that contained less than 60 artifact-free epochs were discarded. In the end, a total of 269 assignments were used for affect and workload recognition. The mean and standard deviation of the number of artifact-free epochs for those assignments were 296.42 and 150.40, respectively, whereas for the corresponding baseline EEG data, the number of useful epochs had a mean value of 170.12 and standard deviation of 52.61.

B. Feature Generation

A recent literature review on EEG features for emotion recognition quantitatively evaluated and compared different feature types [26]. The results showed that the first difference feature in statistical features, fractal dimension (FD) features, and higher order crossings (HOC)-based features performed well and were frequently selected by feature selection methods. Other feature types, such as higher order spectra (HOS) bicoherence features and Hilbert-Huang spectrum (HHS) features, contained valuable features as well. As a first step to explore EEG-based affect and workload recognition, statistical features, FD features, and

HOC-based features were selected because of reported strong performances in other studies [24], [26], [35] and relatively simple implementation and less computational demand compared to HOS bicoherence and HHS features. Power features are the most popular features in EEG studies and therefore were included in the analysis.

We calculated five sets of features for EEG signals recorded from each electrode, summarized in Table II. Time domain feature types captured the statistical measures (statistical features), signal complexity (FD features), and signal oscillation patterns (HOC-based features), whereas the frequency domain feature types characterized the strength of EEG oscillations at a given frequency range (bands and bins). Features extracted from artifact-free epochs belonging to the same baseline/assignment were averaged to obtain the feature vector for the corresponding baseline/assignment.

C. Feature Calibration

In order to train group-level models using EEG data collected from different subjects and different visits, it is necessary to remove feature variations resulting from time and individual differences. We developed a two-step feature calibration method, baseline feature subtraction followed by individualized feature normalization, to prepare the EEG features for group-level affect and workload recognition. The normalization step, as shown in (1), rescaled the range of each subject's features based on the means of his/her features that belong to the low and high intensity classes. This requires each subject to provide data from both classes, e.g., examples from low engagement and high engagement. As a consequence, after individualized feature normalization, the number of examples reduced to 82 (6 subjects) for engagement, 184 (13 subjects) for enjoyment, 146 (10 subjects) for boredom, 248 (18 subjects) for frustration, and 244 (18 subjects) for workload. The effect of feature calibration is illustrated in Fig. 2 using engagement examples. As can be seen, the examples are more separable after feature calibration.

$$f' = \left(f - \overline{\boldsymbol{f}_{\text{low}}}\right) / \left(\overline{\boldsymbol{f}_{\text{high}}} - \overline{\boldsymbol{f}_{\text{low}}}\right) \quad (1)$$

D. Feature Selection and Classification

For the purpose of identifying discriminative features, we ranked the features based on univariate statistical tests. The one-way analysis of variance (ANOVA) test was used to rank features in descending order of the F-values. In other words, a larger F-value indicates the feature has greater discriminative power. Then, model training and evaluation was conducted on a subset of top-ranked discriminative features, where the number of features increased from 3 to 45 by iteratively adding 3 features based on F-value ranking. For FD features, the maximum number of features was 14.

The k-nearest neighbors (kNN) method was used for model training and evaluation because it performed the best in our preliminary study [29]. Three key hyper-parameters (16 combinations) were tuned, including the number of nearest neighbors (1, 3, 9, or 27), distance metric (Manhattan or Euclidean), and weighting scheme (uniform-weighted or distance-weighted).

Nested cross-validation (CV) was used for model selection. The outer loop was leave-one-subject-out CV whereas the inner loop was stratified ten-fold CV with randomization. The inner CV compared the classification performance of kNN models using each of the 16 combinations of hyper-parameters. The best performing model was then evaluated with the test set in the outer CV. The macro-averaged F_1 score of two classes was used as the scoring function to compare hyper-parameters. For imbalanced datasets, the macro- and micro-averaged methods are more suitable than classification accuracy for representation of the results. However, in the case of binary classification, the micro-averaged method is the same as accuracy measure. Therefore, we selected the macro-averaged method in this study. The classification result of the outer CV is the macro-averaged F_1 score of the combined results of all the subjects. Since randomization was used in the inner CV, we performed 50 repetitions of nested CV to acquire more robust classification results. Standardization was used to preprocess the features.

IV. RESULTS

A. Classification Results

Fig. 3 summarizes the classification results of affect and workload recognition with respect to feature types and the number of features. For engagement recognition, power features from bins performed the best with 18 selected features. The performances of statistical features, HOC-based features, and power features from bands were less accurate (about 0.04). FD features scored significantly lower (by at least 0.2) than the other feature types. With only 3 features, power features from bands and bins reached a high macro-averaged F_1 score of 0.90. In the case of enjoyment recognition, HOC-based features outperformed the other feature types significantly with a score of 0.88. The second best feature type was power features from bins with a score of 0.79, closely followed by statistical features and power features from bands. FD features performed the worst in enjoyment recognition as well. Similarly, HOC-based features achieved a much higher score for recognizing frustration. In terms of boredom recognition, power features from bins and HOC-based features performed the best with power features from bins performed slightly better. As far as workload recognition is concerned, the performances of HOC-based features, power features from bands and bins were comparable to each other.

On average, except engagement recognition, HOC-based features were superior among the five feature types, especially in recognizing enjoyment and frustration. For engagement and boredom, power features from bins achieved the highest accuracy. The top feature types and the numbers of features selected by kNN were 18 power features from bins for engagement, 30 HOC-based features for enjoyment, 24 power features from bins for boredom, 45 HOC-based features for frustration, and 30 HOC-based features for workload. Given these feature types and the numbers of features, the most commonly selected hyper-parameters were 1 nearest neighbor, uniform weight, and Euclidean distance metric for engagement; 27 nearest neighbors, distance weight, and Manhattan distance metric for enjoyment; 3 nearest neighbors, uniform weight, and Euclidean distance metric for boredom; and 27 nearest neighbors, uniform weight, and Manhattan distance metric for frustration and workload. The final classification results using the best features with the best performing hyper-parameters

are shown in Table III. We list the precision, recall, and F_1 scores of the leave-one-subject-out CV for both low and high intensity classes. High precision score, or positive predictive value, indicates that the classifier is returning accurate results. High recall score, or true positive rate, shows that the classifier is returning a majority of all positive results. These results imply that models based on EEG activation are able to detect with high accuracy subjects' states of low engagement, low enjoyment, high frustration, and high workload. Boredom recognition had relatively low accuracy. Because the number of examples in the low intensity class was four times larger than that of the high intensity class for boredom, the performance of boredom recognition might improve with more examples of high intensity boredom.

B. Discriminative Features

To investigate which features and which electrodes provide the most information, we computed the relative frequency of each feature subtype and electrode. For each feature type and rating category, given the selected features that yield the best classification results, we counted the feature occurrence and electrode occurrence. Then, the occurrence counts were normalized by the total number of selected features and weighted by classification results. This step is important because it ensures that the relative frequencies accounted for the performances of the discriminative features, and it enables comparison of discriminative features across rating categories. The results are shown in Fig. 4. The feature usage is represented as histograms and the electrode usage is represented as topographies. The classification results, macro-averaged F_1 scores, of the discriminative features are labeled as well.

Because backward models do not allow for a definitive physiologically-based interpretation [37], we did not attempt to link each of the identified features to the underlying neural sources. The discriminative features illustrated in Fig. 4 were chosen jointly by the univariate feature selection method and kNN method to minimize the effect of noise on feature weights. For engagement, the majority of the discriminative features were located in the left hemisphere for power features from bands and bins, however, it was the opposite for HOC-based features. Power features were mostly selected from electrodes F7, F3, and T7, whereas HOC-based features were mostly drawn upon from electrodes FC6, F8, T8, and O2 with some features from electrodes F7, F3, and FC5. In terms of enjoyment, the important electrodes were AF3, FC5, FC6, F8, O2, and P8. In addition, electrodes F3 and O1 were frequently used by HOC-based features, and electrodes F7, AF4, and T7 were used often by statistical features. Power features from bands and bins were drawn upon from all the electrode locations. The discriminative features for boredom were mostly selected from electrodes FC5, F4, and O2. Additionally, electrode F7 was prominent for power features from bins and electrodes F8 and AF4 were important for HOC-based features. The prominent electrodes for frustration recognition were less clear. Locations O1, FC6, and F8 were mostly selected for statistical features. For HOC-based features, electrodes AF3, F7, F3, F4, P7, and O1 were frequently used. Power features from bands preferred electrodes F4 and F8, and power features from bins preferred electrodes T7, P8, and F4. In the case of workload recognition, for statistical features and power features the prominent electrodes were F7 and T8 with some importance given to F8 and P8. HOC-based features were mostly

selected from electrodes AF4, F4, FC5, and T7. In general, power features from bins and bands were similar in electrode usage. However, power features from bins performed better than power features from bands. According to the feature usage histogram of power features from bins, features related to β and γ bands seem more valuable in recognizing affect and workload.

V. Discussion

A. Validation of Therapist's Measures of Mental States

The results presented in the previous section suggest that EEG-based affect and workload recognition is possible for adolescents with ASD and thus can be used to individualize driving training. However, the results need to be interpreted cautiously. It is unclear whether the overall rating data provided by a therapist can accurately capture subjects' affective states and mental workload. While expert rating is a widely used method [31] that we have used in this work, obtaining the ground truth of implicit user states is still a hard problem [38]. To examine the validity of the overall rating data and the thresholding values, we related the overall rating data to the performance data. The results are illustrated in Fig. 5. The x axes are the performance data in terms of the number of errors that occurred in each assignment. The means and standard deviations of all the rating categories with respect to error count are shown as line plots. From the line plots, positive correlations between error count and three other variables: frustration, boredom, and difficulty, can be observed. We further evaluated the correlations based on the Pearson product-moment correlation coefficient. Strong positive correlations were found between frustration and error count ($r = 0.57$, $p < 0.001$) and between difficulty and error count ($r = 0.60$, $p < 0.001$). Moderate positive correlation between boredom and error count ($r = 0.31$, $p < 0.001$) were found. In the cases of engagement ($r = -0.29$, $p < 0.001$) and enjoyment ($r = -0.25$, $p < 0.001$), weak negative correlations were found. Strong correlations between frustration and error count, and between difficulty and error count indicate that the overall rating data reflect the affective states and mental workload of individuals with ASD. Regarding the other three affective states, no conclusion could be drawn with respect to the validity of the overall rating data. In fact, their relationships with performance data are not entirely clear. In terms of the chosen thresholding values, it can be seen from the line plots in Fig. 5 that the predetermined thresholding values approximately separate data into two camps, before 3 errors and after 4 errors. This observation is salient, especially for frustration and difficulty. In terms of boredom, the overall rating score decreased slightly when error count was 5 and 6, whereas for engagement and enjoyment the scores increased slightly when error count was 5 and 6. Overall, the choices of the thresholding values are consistent with performance data.

Learning effect is one factor that would bias subjects' performance, workload, and affective states. We designed the experiment so that the first session and the last session consisted of the same assignments, one assignment from difficulty level two and two assignments from difficulty level five. Fig. 6 shows the averaged error counts and therapist's ratings for session one and session six. Subjects' performance improved over the course of the six sessions due in part to the learning effect. For affect and workload recognition, we used the therapist's

rating data as labels. This information is irrelevant to subjects' driving skills. As can be seen in Fig. 6, subjects' frustration and workload levels were lower in session six than that in session one. Therefore, the learning effect biased the performance, but it should not bias the affective states and workload recognition. The learning effect is also one reason why we did not use the designed difficulty levels as labels for mental workload. In addition, we list subjects' error counts for all six levels of difficulty in Table IV. More difficult driving tasks did not necessarily indicate higher error rates. In fact, the correlation between levels of difficulty and error counts was $r = -0.03$ ($N = 360$). Because one attribute of mental workload, mental effort, varies due to learning effect and is different among subjects, it is not surprising that designed levels of difficulty does not correlate well with subjects' performance data.

B. Comparison with Related Works

Direct comparisons of classification accuracies between studies are difficult due to different experimental designs, subjects' characteristics, data preprocessing procedures, EEG recording devices, etc. The classification accuracies of a workload detector was more than 70% for a subset of subjects in [17]. Dijksterhuis et al. reported averaged accuracies up to 75–80% from lower EEG frequency ranges for workload classifications [19]. With respect to affective states recognition, [39] achieved an accuracy of 86.52% for distinguishing music likability. The best accuracy achieved in identifying emotional states was up to 77.78% over 5 subjects and 56.1% over 15 subjects in [40]. In another study, emotion recognition accuracy reached up to 83.33% for distinguishing 6 emotions and 100% for distinguishing fewer emotions [24]. Lan et al. [41] combined features with high intra-class correlation and improved accuracy to 73.1% for detecting positive negative emotions. Similarly, with combination of features, Liu et al. [42] achieved 87.02% accuracy in recognizing 2 emotions. Based on Table III, our classification results scored 0.95 over 6 subjects for engagement, 0.895 over 13 subjects for enjoyment, 0.78 over 10 subjects for boredom, 0.875 over 18 subjects for frustration, and 0.855 over 18 subjects for workload. Overall, these results are comparable to the extant literature.

As far as the feature generation approaches are concerned, HOC-based features were shown to be superior in detecting emotional states. HOC-based features outperformed statistical features in [24] and power features from bands and bins in [26] for emotion recognition. Power features from bands and bins performed well for engagement and workload recognition. This is in line with other studies that used power features to monitor task engagement [20], [21], and analyzed correlations between power features and workload and engagement levels [23]. In addition, our results indicate that power features from bins are more valuable for engagement recognition compared to the rest of the feature types. Power features from bins outperformed power features from frequency bands in general. This trend is in accordance with [26].

The driving tasks were designed to resemble real world scenarios. The complexity of the task requires working memory and long-term memory, visuospatial processing, visual and auditory processing, attention and emotion regulation, and decision making. According to the electrode usage results, features derived from frontal electrodes were the most

discriminative for affect and workload recognition. Temporal electrodes were also frequently used for engagement and workload recognition. Additionally, occipital electrodes were selected often for engagement, enjoyment, boredom, and frustration recognition. The discriminative features for enjoyment and frustration were mostly drawn upon from parietal electrodes as well. In terms of feature usage, features related to β and γ bands seem more valuable. These results are consistent with previous studies. Dijksterhuis et al. found that driver's workload classifications were most accurate when based on high frequencies and the frontal electrodes [19]. Frontal and parietal electrodes were found by Lin et al. [22] to be most informative for classifying emotional states. In [39], frontal, prefrontal, temporal, and occipital electrodes correlated significantly with ratings of music likability. Compared to other power features, features from β and γ bands were more discriminative in [39] and [26].

C. Limitations and Future Works

The main limitation of the current work comes from the small sample size. Unlike studies that could record large amount of samples per session, at most three samples were available per session in this study. Extracted EEG features were used to recognize the affective states and workload in each assignment. We did not attempt to detect the affective and workload changes based on epochs due to the requirement of human therapist's ground truth rating. It is not feasible to ask a therapist to provide rating data every few seconds. A secondary reason is that the individualized adaptation should not occur so quickly. For performance-based system adaptation, we do not reduce task difficulty whenever a driving error is detected. Similarly, for affect- and workload-based system adaptation, we will select the next driving task based on the overall states during the entire assignment instead of the states detected during the last few seconds of data.

There are several other limitations to address. First, a large portion of EEG data were removed in the process of artifact removal. It is worthwhile to explore how eye movements and muscular artifacts influence the classification results. They may improve the affective states and mental workload recognition [15], [19]. Second, different affective states may have an influence on mental workload recognition and vice versa. That is to say, in this study affective states are likely to co-vary with mental workload. Whether this confounding factor inflates the results, or could be harnessed to improve the classification accuracies, requires further study. Third, the current work is limited to offline analysis. The future system needs to close the loop between the VR-based simulator and subjects using EEG signals to achieve individualized intervention. In addition, different strategies for VR-based driving system adaptation will be explored and the results will be subjected to comparison with performance-based system adaptation. Majority voting is one method to combine affect and workload prediction results for adaptive automation.

VI. Conclusion

We integrated an EEG input modality into a novel VR-based driving simulator which was developed for ASD intervention. EEG data as well as a therapist's overall rating data on five categories (engagement, enjoyment, boredom, frustration, and difficulty) were collected

from 20 subjects diagnosed with ASD over a total of 120 sessions. Models of affect and workload were trained to provide the basis for a future EEG-based passive BCI system, which has the potential to tailor the driving skill training for specific individuals with ASD based on their affective states and mental workload. We systematically evaluated and compared five feature generation approaches with univariate feature selection method and the kNN algorithm. The classification results imply that models based on EEG activations are able to detect with high accuracy the states of low engagement, low enjoyment, high frustration, and high workload for ASD population. Boredom recognition had relatively low accuracy. In the end, classification models were built using power features from bins for engagement and boredom, and using HOC-based features for the rest of the states. The most discriminative features for affect and workload recognition were extracted from frontal electrodes. The analyses on EEG data collected from real world tasks with ASD population demonstrated the feasibility of EEG-based ASD intervention individualization and provided insight on the performance of several different feature types in this context. However, despite all its promise the current work is limited to binary classification and offline analysis after extensive artifact rejection. Thus while the current work is the first step towards an adaptive driving simulator for ASD intervention, the true potential of the developed models to measure the flow states of individuals with ASD based on online predictions of their affective states and mental workload requires further exploration in the future.

Acknowledgments

This work was supported in part by the National Institute of Health Grant 1R01MH091102-01A1, the National Science Foundation Grant 0967170, the National Institute of Health Grant 1R21AG050483-01A1, and the Hobbs Society Grant from the Vanderbilt Kennedy Center.

References

1. Wingate M, et al. Prevalence of autism spectrum disorder among children aged 8 years-autism and developmental disabilities monitoring network, 11 sites, United States, 2010. MMWR SURVEILLANCE SUMMARIES. 2014; 63(2)
2. Lord C, et al. Autism spectrum disorders. Neuron. 2000; 28(2):355–363. [PubMed: 11144346]
3. Cox SM, et al. Driving simulator performance in novice drivers with autism spectrum disorder: the role of executive functions and basic motor skills. Journal of autism and developmental disorders. 2015:1–13. [PubMed: 25428291]
4. Classen S, et al. Driving characteristics of teens with attention deficit hyperactivity and autism spectrum disorder. American journal of occupational therapy. 2013; 67(6):664–673. [PubMed: 24195900]
5. Daly BP, et al. Driving behaviors in adults with Autism Spectrum Disorders. Journal of autism and developmental disorders. 2014; 44(12):3119–3128. [PubMed: 24925544]
6. Reimer B, et al. Brief report: Examining driving behavior in young adults with high functioning autism spectrum disorders: A pilot study using a driving simulation paradigm. Journal of autism and developmental disorders. 2013; 43(9):2211–2217. [PubMed: 23338532]
7. Wade, J., et al. Universal Access in Human-Computer Interaction. Universal Access to Information and Knowledge. Springer; 2014. Design of a virtual reality driving environment to assess performance of teenagers with ASD; p. 466-474.
8. Sheppard E, et al. Brief report: Driving hazard perception in autism. Journal of autism and developmental disorders. 2010; 40(4):504–508. [PubMed: 19890705]
9. Cox NB, et al. Brief Report: Driving and young adults with ASD: Parents' experiences. Journal of autism and developmental disorders. 2012; 42(10):2257–2262. [PubMed: 22359179]

10. Stahmer AC, et al. Toward a technology of treatment individualization for young children with autism spectrum disorders. Brain research. 2011; 1380:229–239. [PubMed: 20858466]
11. Kandalaft MR, et al. Virtual reality social cognition training for young adults with high-functioning autism. Journal of autism and developmental disorders. 2013; 43(1):34–44. [PubMed: 22570145]
12. Bekele E, et al. Understanding how adolescents with autism respond to facial expressions in virtual reality environments. Visualization and Computer Graphics, IEEE Transactions on. 2013; 19(4): 711–720.
13. Wade J, et al. A Gaze-Contingent Adaptive Virtual Reality Driving Environment for Intervention in Individuals with Autism Spectrum Disorders. ACM Transactions on Interactive Intelligent Systems (TiiS). 2016; 6(1):3.
14. Kiili K. Digital game-based learning: Towards an experiential gaming model. The Internet and higher education. 2005; 8(1):13–24.
15. Nijholt A, et al. Turning shortcomings into challenges: Brain–computer interfaces for games. Entertainment Computing. 2009; 1(2):85–94.
16. Csikszentmihalyi, M. Flow: The Psychology of Optimal Experience. Harper & Row: 1990.
17. Kohlmorgen J, et al. Improving human performance in a real operating environment through real-time mental workload detection. Toward Brain-Computer Interfacing. 2007:409–422.
18. Wang Y-T, et al. Developing an EEG-based on-line closed-loop lapse detection and mitigation system. Using Neurophysiological Signals that Reflect Cognitive or Affective State. 2015:85.
19. Dijksterhuis C, et al. Classifying visuomotor workload in a driving simulator using subject specific spatial brain patterns. Using Neurophysiological Signals that Reflect Cognitive or Affective State. 2015:202.
20. Freeman FG, et al. Evaluation of an adaptive automation system using three EEG indices with a visual tracking task. Biological psychology. 1999; 50(1):61–76. [PubMed: 10378439]
21. Szafir, D., Mutlu, B. 30th ACM Conference on Human Factors in Computing Systems, CHI 2012, May 5, 2012 – May 10, 2012. Austin, TX, United States: Association for Computing Machinery; 2012. Pay attention! Designing adaptive agents that monitor and improve user engagement; p. 11-20.
22. Lin Y-P, et al. EEG-based emotion recognition in music listening. Biomedical Engineering, IEEE Transactions on. 2010; 57(7):1798–1806.
23. Berka C, et al. EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. Aviation, space, and environmental medicine. 2007; 78(Supplement 1):B231–B244.
24. Petrantonakis PC, Hadjileontiadis LJ. Emotion Recognition From EEG Using Higher Order Crossings. Information Technology in Biomedicine, IEEE Transactions on. 2010; 14(2):186–197.
25. Kim M-K, et al. A review on the computational methods for emotional state estimation from the human EEG. Computational and mathematical methods in medicine. 2013; 2013
26. Jenke R, et al. Feature extraction and selection for emotion recognition from EEG. Affective Computing, IEEE Transactions on. 2014; 5(3):327–339.
27. Oberman LM, et al. EEG evidence for mirror neuron dysfunction in autism spectrum disorders. Cognitive Brain Research. 2005; 24(2):190–198. [PubMed: 15993757]
28. Murias M, et al. Resting state cortical connectivity reflected in EEG coherence in individuals with autism. Biological psychiatry. 2007; 62(3):270–273. [PubMed: 17336944]
29. Fan, J., et al. presented at the 37th Annual International Conference of the IEEE Engineering in Medicine and Biological Society (EMBS). in press; 2015. A Step Towards EEG-based Brain Computer Interface for Autism Intervention.
30. Jasper HH. The ten twenty electrode system of the international federation. Electroencephalography and Clinical Neurophysiology. 1958; 10:371–375. 1958.
31. Baker RS, et al. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. International Journal of Human-Computer Studies. 2010; 68(4):223–241.
32. Moray, N. Mental workload: Its theory and measurement. Springer Science & Business Media; 2013.
33. Cain B. A review of the mental workload literature. DTIC Document. 2007

34. Wilson GF, Russell CA. Performance enhancement in an uninhabited air vehicle task using psychophysiologically determined adaptive aiding. Human Factors: The Journal of the Human Factors and Ergonomics Society. 2007; 49(6):1005–1018.

35. Chai, TY., et al. International. Vol. 1. Penerbit UTHM; 2010. Classification of human emotions from EEG signals using statistical features and neural network; p. 1-6.

36. Accardo A, et al. Use of the fractal dimension for the analysis of electroencephalographic time series. Biological cybernetics. 1997; 77(5):339–350. [PubMed: 9418215]

37. Haufe S, et al. "On the interpretation of weight vectors of linear models in multivariate neuroimaging," (in eng). Neuroimage. Feb 15.2014 87:96–110. Research Support, Non-U.S. Gov't. [PubMed: 24239590]

38. Brouwer A-M, et al. Using neurophysiological signals that reflect cognitive or affective state: six recommendations to avoid common pitfalls. Frontiers in neuroscience. 2015; 9

39. Hadjidimitriou SK, Hadjileontiadis LJ. Toward an EEG-based recognition of music liking using time-frequency analysis. Biomedical Engineering, IEEE Transactions on. 2012; 59(12):3498–3510.

40. Sohaib, AT., et al. Foundations of Augmented Cognition. Springer; 2013. Evaluating classifiers for emotion recognition using EEG; p. 492-501.

41. Lan Z, et al. Real-time EEG-based emotion monitoring using stable features. The Visual Computer. 2016; 32(3):347–358.

42. Liu, Y., Sourina, O. Transactions on Computational Science XXIII. Springer; 2014. Real-time subject-dependent EEG-based emotion recognition algorithm; p. 199-223.

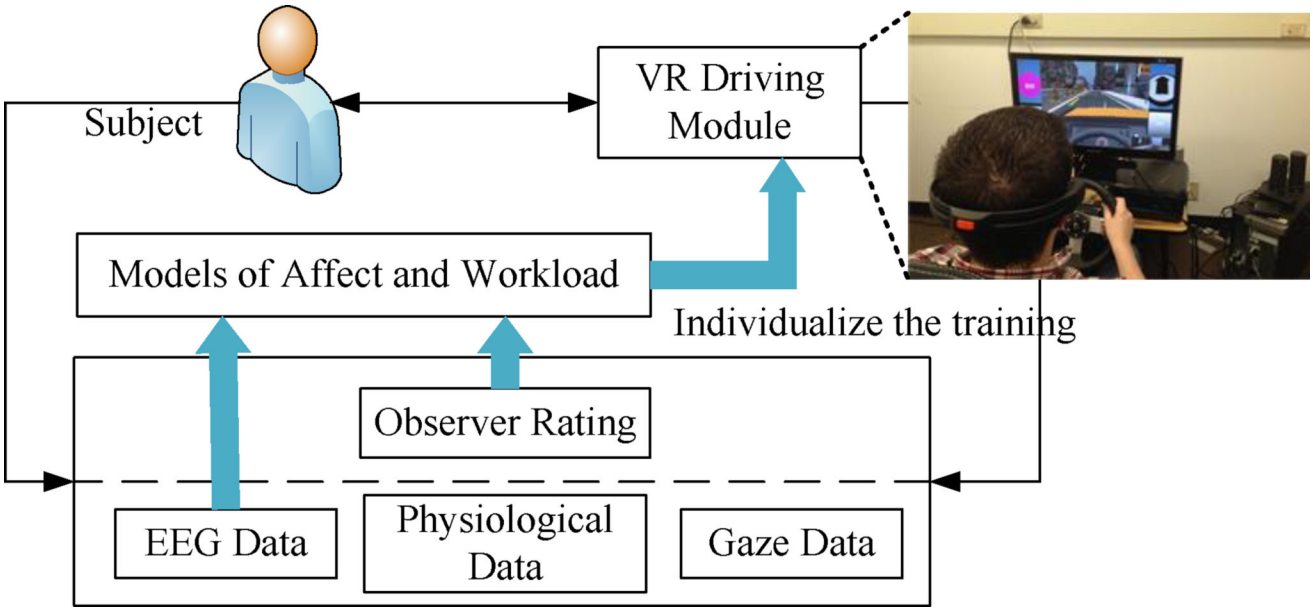

Figure 1.
Framework overview.

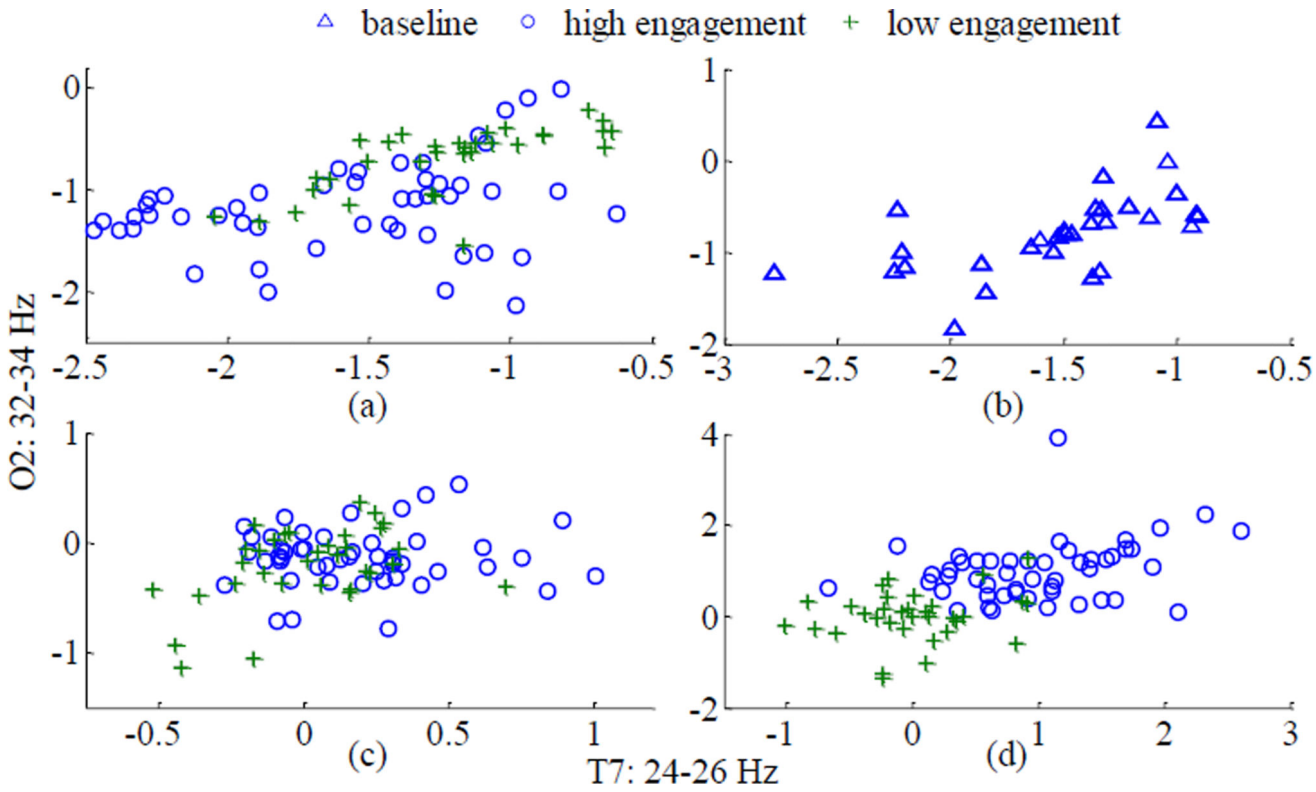

Figure 2.
Comparison of 2D feature scatter plot for engagement (a) before feature calibration, (b) baseline feature distribution, (c) after baseline removal, and (d) after feature normalization.

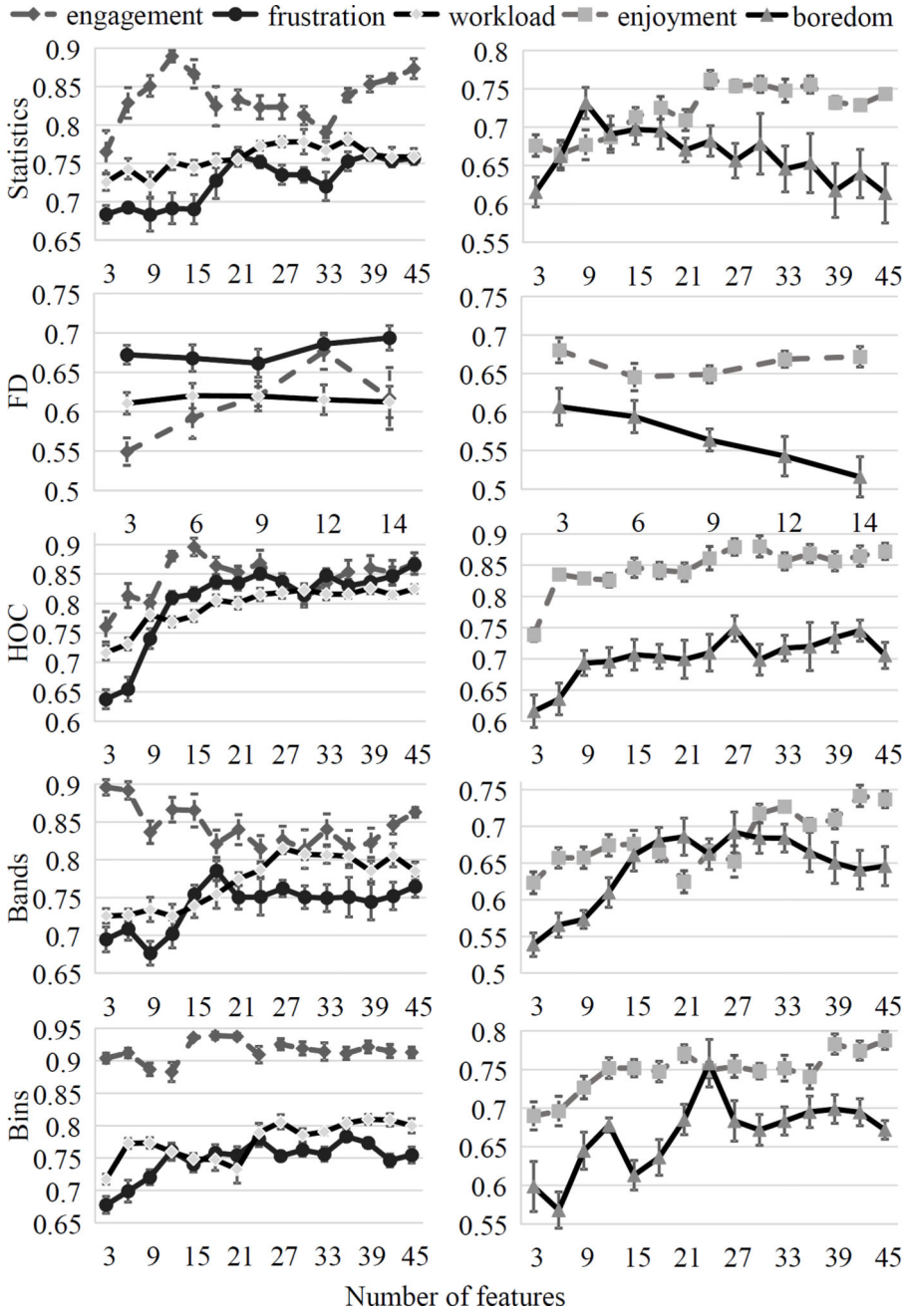

Figure 3.
Classification results evaluated based on macro-averaged F_1 score.

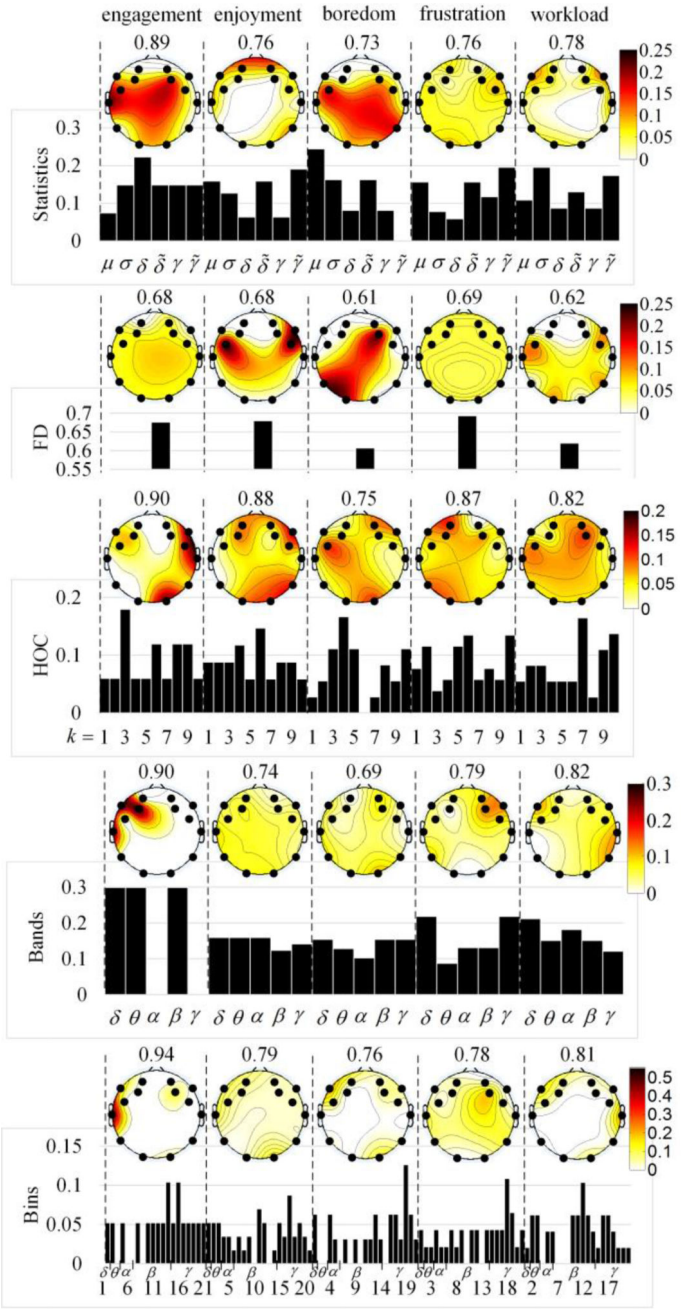

Figure 4.
Feature usage and electrode usage.

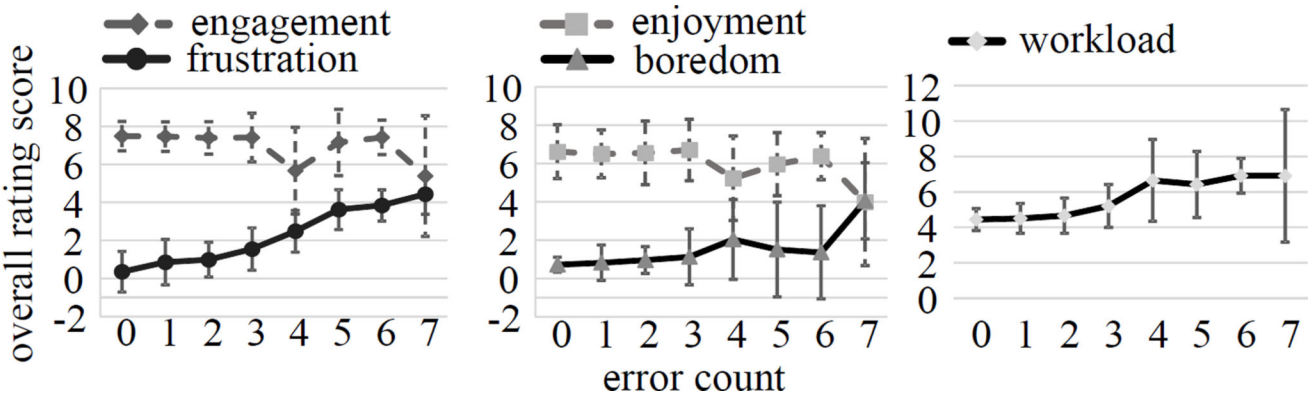

Figure 5.
Overall rating data as a function of participants' performance.

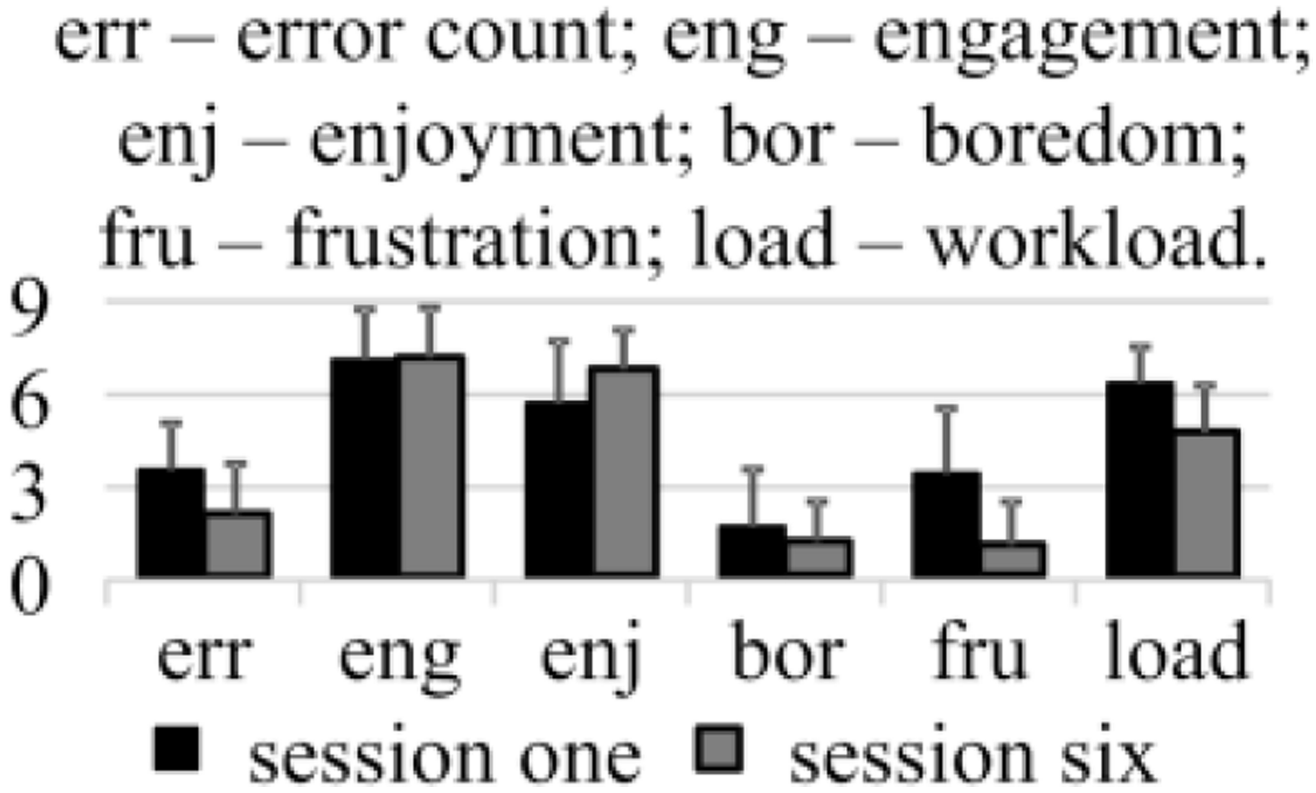

Figure 6.
Learning effect.

TABLE I

Subject Characteristics

	Sample size	Mean	Standard deviation
Chronological age in years	20	15.29	1.65
ADOS total raw score	16	13.56	3.67
ADOS severity score	16	7.81	1.42
SRS-2 total raw score	20	97.85	26.97
SRS-2 T-score	20	75.45	9.70
SCQ lifetime total score	19	20.84	9.48
IQ	15	108.93	17.47

ADOS = autism diagnostic observation schedule, SRS-2 = social responsiveness scale, second edition, SCQ = social communication questionnaire. Sample size varies due to missing data.

TABLE II

Extracted Features

Features	Method/Parameter	No.
Statistics	mean μ_x, standard deviation σ_x, first difference δ_x, standardized first difference $\tilde{\delta}_x$, second difference γ_x, standardized second difference $\tilde{\delta}_x$ [35]	84
FD	Higuchi algorithm, $k_{max} = 6$ [36]	14
HOC	backward difference operator, $k = 1,2,\ldots 10$ [24]	140
Bands	Hanning tapering function, PSD (Welch's method), δ (1–4 Hz), θ (4–8 Hz), α (8–13 Hz), β (13–30 Hz), γ (30–44 Hz), normalization with log transformation	70
Bins	Hanning tapering function, PSD (Welch's method), 2–44 Hz ($\Delta f = 2$Hz), normalization with log transformation	294

TABLE III

Classification Results Using the Best Performing Features and Hyper-Parameters

	Intensity	Examples	Precision	Recall	F1 score
Engagement	low	33	0.97	0.91	0.94
	high	49	0.94	0.98	0.96
Enjoyment	low	77	0.88	0.87	0.88
	high	107	0.91	0.92	0.91
Boredom	low	118	0.92	0.92	0.92
	high	28	0.64	0.64	0.64
Frustration	low	176	0.93	0.93	0.93
	high	72	0.82	0.82	0.82
Workload	low	122	0.83	0.89	0.86
	high	122	0.88	0.82	0.85

TABLE IV

Error Counts

Difficulty	Mean	STD	N
level 1	2.62	1.79	58
level 2	2.84	1.94	43
level 3	2.47	1.64	60
level 4	2.05	1.57	60
level 5	2.39	1.60	79
level 6	2.70	1.67	60