Improving Intention Detection in Single-Trial Classification Through Fusion of EEG and Eye-Tracker Data

Xianliang Ge, Yunxian Pan, Sujie Wang, Linze Qian, Jingjia Yuan, Jie Xu, Nitish Thakor, Fellow, IEEE, and Yu Sun, Senior Member, IEEE

Abstract—Intention decoding is an indispensable procedure in hands-free human–computer interaction (HCI). A conventional eye-tracker system using a single-model fixation duration may issue commands that ignore users’ real expectations. Here, an eye-brain hybrid brain–computer interface (BCI) interaction system was introduced for intention detection through the fusion of multimodal eye-tracker and event-related potential (ERP) [a measurement derived from electroencephalography (EEG)] features. Eye-tracking and ERP data were recorded from 64 healthy participants as they performed a 40-min customized free search task of a fixed target icon among 25 icons. The corresponding fixation duration of eye tracking and ERP were extracted. Five previously-validated linear discriminant analysis (LDA)-based classifiers [including regularized LDA, stepwise LDA, Bayesian LDA, shrinkage linear discriminant analysis (SKLDA), and spatial-temporal discriminant analysis] and the widely-used convolutional neural network (CNN) method were adopted to verify the efficacy of feature fusion from both offline and pseudo-online analysis, and the optimal approach was evaluated by modulating the training set and system response duration. Our study demonstrated that the input of multimodal eye tracking and ERP features achieved a superior performance of intention detection in the single-trial classification of active search tasks. Compared with the single-model ERP feature, this new strategy also induced congruent accuracy across classifiers. Moreover, in comparison with other classification methods, we found that SKLDA exhibited a superior performance when fusing features in offline tests (ACC = 0.8783, AUC = 0.9004) and online simulations with various sample amounts and duration lengths. In summary, this study revealed a novel and effective approach for intention classification using an eye-brain hybrid BCI and further supported the real-life application of hands-free HCI in a more precise and stable manner.

Index Terms—Electroencephalography (EEG), event-related potential (ERP), eye–brain–computer interface, eye-tracker, single-trial classification.

Manuscript received 11 July 2021; revised 4 May 2022 and 30 September 2022; accepted 21 November 2022. Date of publication 12 December 2022; date of current version 16 January 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 82172056, Grant 81801785, and Grant T2192931, in part by the National Science Foundation for Distinguished Young Scholars of Zhejiang Province, in part by the National Key Research and Development Program of China under Grant 2021YD020408, in part by the Key Research and Development Program of Zhejiang Province under Grant 2022C03036, in part by the Hundred Talents Program of Zhejiang University, in part by the Zhejiang University Global Partnership Fund under Grant 100000-11320, in part by the Zhejiang Lab under Grant 2019KNOAD01, in part by the Space Medical Experiment Project of China Manned Space Program under Grant HYZHXM03001, and in part by the Science and Technology Special Project of the Institute of Wenzhou, Zhejiang University under Grant XMGL-KJZX-202203. This article was recommended by Associate Editor M. Lau. (Corresponding authors: Xianliang Ge; Yu Sun.)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board of Zhejiang University under application no. 2019-001 and performed in line with the Declaration of Helsinki. Xianliang Ge, Yunxian Pan, and Jie Xu are with the Center for Psychological Sciences, Zhejiang University, Hangzhou 310027, China (e-mail: 0918082@zju.edu.cn; yxpan94@zju.edu.cn; xujie0987@zju.edu.cn).

Sujie Wang, Linze Qian, and Jingjia Yuan are with the Key Laboratory for Biomedical Engineering of Ministry of Education of China, Department of Biomedical Engineering, Zhejiang University, Zhejiang 310027, China (e-mail: sujie0987@zju.edu.cn; qiaqiayuan@gmail.com).

Nitish Thakor is with the Department of Biomedical Engineering, Johns Hopkins University School of Medicine, Baltimore, MD 21218 USA, and also with the Department of Biomedical Engineering, National University of Singapore, Singapore 117456 (e-mail: elenu@nus.edu.sg).

Yu Sun is with the Key Laboratory for Biomedical Engineering of Ministry of Education of China, Department of Biomedical Engineering, Zhejiang University, Hangzhou 310027, China, also with the Department of Neurology, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou 310027, China, and also with the Zhejiang Lab, Hangzhou 310027, China (e-mail: yusun@zju.edu.cn).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/THMS.2022.3225633.

Digital Object Identifier 10.1109/THMS.2022.3225633
SWLDA  Stepwise linear discriminant analysis.
VEOG  Vertical electrooculogram.

I. INTRODUCTION

HANDS-FREE human–computer interaction (HCI) has drawn growing attention due to its convenience in various environments. The purpose of hands-free interaction is to free the hands to prevent overload, for example, by undertaking multiple tasks while driving [1], and avoiding contaminating the environment during surgery [2]. To realize hands-free interaction, the most important thing is that the existing equipment can accurately identify the user’s behavioral intention. As an efficient input modality in HCI, eye-gaze systems have been used to predict intentions without requiring individuals to verbally communicate those intentions [3], [4], [5], [6]. For example, Zelinsky et al. [7] predicted a specific search target by analyzing eye-gaze patterns. Sattar et al. [8] further predicted the categories of search targets from eye-gaze data. In addition to predicting search intent, Hu et al. [9] proposed a method for predicting user interface interaction intent based on eye-gaze patterns. During the experiment, gaze data from participants under various interface conditions were analyzed to predict the type of user interaction. Although the eye-gaze system is intuitive in HCI, it also has limitations. First, the quality of eye-gaze systems greatly depends on many factors, such as material and brightness. Second, eye-gaze systems lack natural activation commands, such as clicking on a target. When using an eye-tracker, selection is typically achieved through gazing on the target item for a specific period [10], [11], [12]. However, unexpected commands may be issued when using a standalone eye-gaze system, leading to the Midas-touch problem [13], [14]. Therefore, to enhance the fluency and robustness of these systems, a natural and intuitive input and decoding approach is necessary.

Another technology that is widely considered in hands-free interaction is the brain–computer interface (BCI). BCI establishes a direct communication channel from brain signals, which can be used to detect ongoing cognition, such as mental fatigue [15], emotional state [16], and intention [17], [18], [19]. This technology has been proposed to decode cognitive information implicitly from users’ minds, without interruption of the primary task [20]. For example, Kang et al. [21] proposed a human implicit search intent recognition system based on electroencephalography (EEG) signals. They collected EEG data from participants viewing images on ten web pages and categorized search intent. Kim et al. [22] classified two implicit intentions of nuclear operators’ agreement and disagreement by machine learning-based algorithms. This classification estimates the operator’s implicit intentions during nuclear power operations, based on EEG data collected during the performance of simulated operational tasks. For the free visual search task, Kaunitz et al. [23] found that when subjects detected a target among distractors, a robust sensory component of fixation event-related potentials (ERPs) emerged, and a single-trial analysis could differentiate the type of stimulus based on EEG signals. Devillez et al. [24] also observed a P300 component for fixation of the target natural scene compared with free viewing without any target.

In summary, the constituents of ERP contain abundant information toward personal intention and accompany gaze-based control intuitively in free search tasks, which can serve as the feature for distinguishing the intended selection from involuntary fixation. As a result, a combination of ERP and the eye-gaze input system could be complementary and provide a more robust interaction experience. Accumulating evidence indicates that this hybrid BCI can satisfy the need for speed and accuracy simultaneously, overcoming the Midas-Touch problem of eye-tracker and interperson variability of the BCI protocol. For example, Kalika et al. [25] fused eye-gaze data into a P300 speller pipeline and reported an improved classification accuracy and decreased flash number for character spelling. Choi et al. [26] utilized the gaze position to contract a 12 × 12 character matrix into a 3 × 3 one, and highlighted this smaller navigation area to enhance the decoding performance of the P300 speller. However, to the best of authors’ knowledge, most hybrid BCI studies on the free visual search task focus on only the decoding of EEG signals, or use the EEG and eye-tracker for separate control purposes. Few of them have attempted to analyze these two modalities in parallel and fuse them as input for intention classification.

This research gap inspired our study to determine whether the alliance of input data from eye-tracker and BCI could facilitate the performance of intention detection in single-trial classification for active search tasks. Specifically, a self-designed HCI paradigm was proposed in which participants were required to search and identify a target icon among a total of 25 icons for each trial. We analyzed the fixation duration of each stimuli and corresponding ERP components, and these two inputs served as features for target/nontarget classification. In previous studies, linear discriminant analysis (LDA) was extensively used for ERP detection owing to its satisfactory performance and simplicity [27], [28], while it was also accompanied by the disadvantages of high noise sensitivity, poor interperson generalization, and the need for a large training sample [29]. Therefore, multiple improved LDA classifiers [including regularized linear discriminant analysis (RLDA), stepwise linear discriminant analysis (SWLDA), Bayesian linear discriminant analysis (BLDA), shrinkage linear discriminant analysis (SKLDA), and spatial-temporal discriminant analysis (STDA)] with divergent edges were adopted to evaluate their performance in this task. CNN, the most prevalent deep learning framework in the study of BCI [30], was also used for comparison. Our study demonstrated that the fusion of concurrent ERP and fixation duration induced superior performance over single feature in target intention decoding among all classifiers. Pseudo-online validation was further conducted to explore the proper amount of training set, response time of the system, and optimal classification approach to provide additional support for practical application in various real-life HCI scenarios.

II. METHODS AND MATERIALS

A. Subjects

The study sample consisted of 70 university students (male/female = 35/35) from Zhejiang University, China. All participants were aged between 17 and 29 years (mean
Fig. 1. Schematic diagram of the experimental protocol. (a) Setup of the experiment. EEG data were obtained from a 64-channel BP system, and eye-tracker data were collected by using the EyeLink 1000 Plus system. (b) Each participant performed a 40 min target identification experiment, where the participant was asked to search for the target icon as quickly and accurately as possible and memorize and identify its location in the response period.

Fig. 2. Sample screenshots of the visual target identification interface for two random trials. The icons surrounding the target icon, as highlighted with a red box, were enlarged to facilitate identification.

The experimental protocol was a typical target identification task that was customized using the C programming language (see Fig. 1). Specifically, the participants were requested to search for a target icon among multiple nontarget icons as quickly and accurately as possible. A total of 25 icons were presented on a screen (1920 × 1080 pixels) with the background color set at [R, G, B] = [192, 192, 192], which were arranged in 5 × 5 blocks. The experimental interface is shown in Fig. 2. The size of each icon was set at 24 × 24 pixels, corresponding to a field of view (FOV) of 0.67° × 0.67°. The horizontal/vertical distance between each pair of adjacent icons was set at 100 pixels. If an icon was highlighted, its size would be enlarged to 1.5 times the original size (i.e., 36 × 36 pixels, FOV = 1° × 1°). During the target searching, a certain icon that the participants gazed at and the surrounding eight icons were highlighted (see Fig. 2).

In the experiment, a predefined target icon (60 × 60 pixels, FOV = 1.67° × 1.67°) was initially presented for 3 s. A black fixation cross was presented for 1 s, indicating the start of each experimental trial. Then, after 0.8 s of a blank screen, the search interface was presented with a randomly assigned 5 × 5 icon pattern and a timer was started. The cursor was hidden at the moment. The participants were asked to search for the target icon within 5 s using the search interface. The mouse cursor appeared on the screen after finishing the 5 s search period. Meanwhile, all the icons on the display were masked with a dotted line. The participants were requested to move the cursor to the location of the masked target icon and click to confirm the selection. The program would proceed to the next trial starting with a black fixation cross. A short break period (i.e., 5 s) was introduced after 15 trials were completed, while a long break...
period (i.e., 10 s) was introduced after completing 30 trials. For each participant, a total of 240 trials were administered and the entire experiment lasted approximately 40 min.

C. Data Acquisition and Preprocessing

The EyeLink 1000 Plus eye-tracker system (model: SR Research, Ottawa, Canada) was used to record the eye-tracker data. The sample rate was set at 1000 Hz. The participants were seated 60 cm from the monitor with an FOV of 13.86° × 13.86° using a chin support. Prior to the experiment, the eye-tracker system was calibrated for each participant. EEG data were recorded from a 64-channel EEG system (model: BrainAmp DC, Brain Products Gmbh, Gilching, Germany) according to the international 10–20 system. In addition, horizontal and vertical electrooculograms (EOG) were recorded lateral to the outer canthi (HEOG) as well as above and below (VEOG) the right eye. Electrode impedance was kept below 10 kΩ throughout the experiment. Antialiasing was achieved with a bandpass filter (0.5–100 Hz), and a 50 Hz notch filter was applied to avoid main interferences. Raw EEG and EOG signals were digitized at a sampling rate of 500 Hz using FCz as the reference. Two subjects were excluded due to data recording issues.

In the analysis of eye-tracker data, we used the 100 × 100 px area centered on each icon as the area of interest (AOI) of the icon. When a fixation fell in an AOI including nontarget icon, this fixation was classified as the fixation of the nontarget icon, and when a fixation fell in an AOI including a target icon, this fixation was classified as the fixation of the target icon. The fixation duration was calculated by summing the duration of all fixations in the AOI.

A standard EEG preprocessing pipeline was adopted here, which included FIR bandpass filtering (1–40 Hz), re-referencing to the average of all electrodes and ocular artifacts removal by removing the most correlated components to the EOG signals through independent component analysis [31]. All preprocessing steps were performed using customized codes and the EEGLAB toolbox [32] in MATLAB 2017b (The MathWorks Inc, US). More details of the preprocessing steps can be found in our previous studies [16].

The channel selection was based upon the results in [27], which showed the optimal balance between classification performance and the lowest number of electrodes. Here, Fz, Cz, Pz, Oz, P3, P4, PO7, and PO8 electrodes were included for the following classification. To extract features for target and nontarget responses, the continuous EEG signals were segmented to 500 ms epochs time-locked to fixation onsets (i.e., the time at which the fixation fell in the AOI) with baseline correction by 100 ms intervals before the fixation onsets. Consequently, a 5 s visual search process might contain multiple target and nontarget EEG epochs. Afterward, each epoch was down-sampled to 32 Hz, that is, 16 points for each channel and 128 points in total for all 8 channels.

D. Classification Algorithms

Several widely-used algorithms that were popular in studies of ERP-BCI were adopted in this work to assess the classification performance [33], [34], including RLDA [35], SWLDA [36], BLDA [37], SKLDA [38], STDa [39], and CNN [40]. These algorithms were selected to cover the common categories of method for ERP-BCI, that is, the concatenation of temporal points and spatial channels (RLDA, SWLDA, BLDA, SKLDA), the adoption of spatial-temporal samples (STDA), and the deep learning approach (CNN).

1) Regularized LDA: RLDA, a regularized version of LDA, is a popular technique for dimensionality reduction and feature extraction. It was originally introduced to solve the small sample size problem. The performance of RLDA technique depends upon the choice of the regularization parameter. In this work, the regularization parameter was estimated using a deterministic approach according to [41]. This approach avoids the use of the heuristic cross-validation procedure for parameter estimation and improves the computational efficiency. Here, the amount of regularization was set as λ = 0.01.

2) Stepwise LDA: SWLDA, another regularized version of LDA, has been shown to be superior in the case of a small sample size due to its implementation of combined forward and backward stepwise analysis to select suitable features in the discriminant model. Briefly, model estimation for SWLDA is conducted in a greedy manner by iteratively inserting and removing features from the model based upon statistical tests until the maximal number of active variables is reached or no additional features satisfy the entry/removal criteria. Here, the criteria were set as pm = 0.1 and pem = 0.15, as recommended in [42].

3) Bayesian LDA: BLDA is a probabilistic method based upon Bayesian regression and has been shown to outperform the original LDA method when only a small number of training sets were obtained or there is strong noise contamination in the data [43]. According to [37], the neurophysiological and experimental priors are employed explicitly by modeling the trial-level covariance and the weight vector covariance of LDA explicitly as linearly separable components with the relative contribution of each component is controlled by the hyperparameters that could be estimated via restricted maximum likelihood.

4) Shrinkage LDA: By adjusting the extreme eigenvalues of the covariance matrix toward the average eigenvalue, SKLDA improves the traditional LAD when using insufficient training samples. For high-dimensional data with only a few data points given (i.e., EEG data), the estimation for a covariance matrix may become imprecise, which may lead to a systematic error: estimations of the large eigenvalues of the original covariance matrix are too large, and those of small eigenvalues are too small. Of note, shrinkage is a common remedy for compensating the systematic bias of the estimated covariance matrices and shrinkage parameter for high-dimensional feature spaces. In this work, the shrinkage parameter was set at 0.1 according to [44]. Details of SKLDA and its interpretation can be found in [34].

5) Spatial-Temporal Discriminant Analysis: STDA is a multivariate extension of LDA that tries to maximize the discriminant information between target and nontarget classes by finding two projection matrices from spatial and temporal dimensions collaboratively. Unlike the abovementioned versions of LDA method where data were concatenated as input, by incorporating
the spatial and temporal information, STDA reduces the feature dimensionality in the discriminant analysis and decreases the number of required training samples [39].

6) Convolutional Neural Network: CNN was initially used in computer vision and has drawn substantial interest in BCI most recently for its superior performance. In this research, a five-layer CNN was developed for EEG pattern detection. The input of the network was a 2-D space-time EEG signal with a size of \(8 \times 16\). It was followed by two paired layers, with each pair comprising a convolutional layer with batch normalization and a max-pooling layer. In the first convolutional layer, we utilized 32 kernels with a size of \(1 \times 5\) for time domain convolution. The second was used for spatial domain convolution, containing 32 kernels with a size of \(8 \times 1\), which equaled the number of EEG electrodes. After each convolution process, an rectified linear units (ReLU) function was employed for nonlinearization. For max-pooling layers, they both utilized a pooling filter size of \(1 \times 2\) to reduce computational complexity. After the dropout process with a dropout rate of 0.3, the output of the max-pooling layer was applied to two fully connected layers comprising 64 and 2 neurons, respectively. In the decision step, the classification probability is determined by softmax function.

E. Offline Classification

To demonstrate that fusion of EEG and eye-tracker data would lead to superior performance to that of single EEG or eye-tracker modality, classification was initially performed using only EEG or eye-tracker data. Specifically, the fixation duration corresponding to the extracted epoch of one gaze was initially estimated and selected as input for LDA classifier as a benchmark. For EEG data, a 0–500 ms epoch after a gaze was cut out and selected as input for the offline classification. Of note, one trial (corresponding to a 5 s period of target search) might contain multiple target and nontarget samples with the number of nontarget samples larger than that of target samples. A cross-validation approach was initially employed to assess the performance of classifiers under different numbers of training samples using reformatted balance data. Specifically, the training set (Target:Nontarget = 1:1) was designed using sample number from 30 to 420 with a step of 30, while the testing set was randomly selected from the remaining samples and maintained a Target:Nontarget = 1:1 fashion with a maximum amount. Of note, the same training and testing samples were applied on all classification algorithms to allow for fairness comparison. This procedure was repeated 10 times, and the average area under the receiver operating characteristic curves (AUC) was computed for the quantitative comparisons. Then, a separate 10-fold cross-validation approach was applied to the real data (on average, Target:Nontarget \(\approx 1.23\)) to demonstrate that the fusion of multimodal features outperformed single feature.

F. Pseudo-Online Validation

1) Online Classification: A pseudo-online analysis was performed to validate the feasibility and practicability of implementing the decoding algorithm based upon our analysis framework. As at least 240 trials were performed for each subject in the experiment, the samples within the first 80 trials were utilized as the training set, whereas the remaining data were considered as the testing set for assessing the performance of online classification. To avoid the imbalance of the sample amount between two classes in the training set and for the convenience of result analysis in the testing set, the number of target classes was held equal to that of nontarget classes, respectively. Of note, a fixation duration longer than 500 ms was redefined as 500 ms to ensure the same duration of EEG data.

2) Epoch Threshold: As mentioned previously, gaze fixation duration was defined as the duration after a gaze for either a target or nontarget and EEG data between 0 and 500 ms were used as the threshold for data extraction and the following classification. To assess the influence of various thresholds on the classification, we also used 300 ms to 800 ms with a step of 50 ms as the threshold for epoch extraction. For instance, for a predefined threshold (e.g., 400 ms), a gaze fixation duration above the threshold would be redefined as the threshold value and EEG data between 0 and 400 ms would be used as input. In addition, the number of training trials was also considered as a factor contributing to online performance. In detail, the samples within the first 20 to 100 trials with a step of 10 trials were regarded as the training set, while the remaining samples were used for testing. The target to nontarget ratio was also rearranged to 1:1 in both the training and testing set. All software programs ran on a computer with an Intel Core i7 @ 3.6 GHz processor and 32 GB RAM.

III. RESULTS

A. Behavioral Performance

Data from four participants were excluded for signs of poor motivation on the task, likely due to boredom experienced during the target identification experiment. The threshold for signs of poor motivation on the task was set if the error rate of the participant was 1 S.D. lower than the group average. Our final dataset thus consisted of 64 participants (male/female = 31/33) and the following classification was conducted on these participants. Overall, the remaining participants performed the experiment well, as indicated by the relatively high detection rate (mean \(\pm\) S.D. = 99.49% \(\pm\) 0.93%). We performed additional statistical analysis to assess the gender effect and found no significant difference between males and females (\(t_{62} = -1.597, p = 0.115\)).

B. Characteristics of ERPs

The characteristics of ERPs were first analyzed and compared between target and nontarget stimuli. Fig. 3 shows the temporal and spatial differences between two kinds of ERPs for a randomly-selected subject. Specifically, the discriminant ERP features between targets and nontargets were restricted to the occipital areas poststimulus. Hence, these evident differences between targets and nontargets serve as salient underlying features for the following classification algorithms. Moreover, the observed posterior differences were in line with the findings in [27] and justify the selection of the EEG channels (i.e., Fz, Cz,
TABLE I
Offline Classification Performance Across Algorithms

| Algorithms | EEG | Fusion |
|------------|-----|--------|
|            | Accuracy | AUC   | Accuracy | AUC   |
| RLDA       | 0.7793 ± 0.0403 | 0.8007 ± 0.0599 | 0.8802 ± 0.0551 | 0.9104 ± 0.0547 |
| SWLDA      | 0.7754 ± 0.0382 | 0.7908 ± 0.0617 | 0.8761 ± 0.0534 | 0.9072 ± 0.0565 |
| BLDA       | 0.6716 ± 0.0504 | 0.7933 ± 0.0632 | 0.7954 ± 0.0910 | 0.9066 ± 0.0573 |
| SKLDA      | 0.7225 ± 0.0503 | 0.7838 ± 0.0615 | 0.8783 ± 0.0603 | 0.9004 ± 0.0608 |
| STDA       | 0.7652 ± 0.0367 | 0.7789 ± 0.0583 | 0.8740 ± 0.0544 | 0.9049 ± 0.0568 |
| CNN        | 0.7854 ± 0.0403 | 0.8053 ± 0.0622 | 0.8772 ± 0.0535 | 0.9028 ± 0.0512 |

Note: Values are presented as mean ± S.D., Fusion indicates features from EEG and eye-tracker were fused to obtain the accuracy and AUC.

Fig. 3. Distributions of different ERP characteristics of Target and Nontarget from a randomly-selected subject. Lines in red represent selected EEG channels, while lines in blue indicate unselected EEG channels.

Pz, Oz, P3, P4, PO7, and PO8 in this work) for the classification algorithms.

C. Offline Classification
In the offline classification, we first assessed the performance of classifiers under various number of training samples. In line with a previous study [33], we found that the classification performance monotonically increased with the number of training samples (see Fig. 4). We then assessed the classification performance when using features from eye-tracker and EEG data, respectively. When using the eye-tracker feature, we obtained a classification accuracy of $0.8734 \pm 0.0746$ that was served as benchmark. However, the classification performance across algorithms using only EEG data was significantly lower than the benchmark ($p < 0.001$ for all comparisons, paired $t$-test) probably due to the large interindividual differences in single-trial EEG characteristics (see Table I). Moreover, we found that by employing the features from both eye-tracker and EEG data, the classification performance was significantly improved ($p < 0.001$ for all comparisons, paired $t$-test) and exhibited performance superior to the benchmark for most of the subjects (see Fig. 5). Further investigation of the classification performance across six methods revealed that BLDA exhibited relatively low accuracy in the fusion manner. Hence, the remaining five methods (i.e., RLDA, SWLDA, SKLDA, STDA, and CNN) were selected for the following pseudo-online validation.

D. Pseudo-Online Classification
The performance of the pseudo-online classification is shown in (see Table II). Again, we first obtained the performance benchmark from eye-tracker data (accuracy = $0.8277 \pm 0.0862$ and
TABLE II

| Algorithms | Accuracy | AUC   | Time (ms) | Accuracy | AUC   | Time (ms) |
|------------|----------|-------|-----------|----------|-------|-----------|
| RLDA       | 0.6550 ± 0.0560 | 0.7125 ± 0.0692 | 3.3722 ± 0.0544 | 0.8419 ± 0.0879 | 0.8890 ± 0.0778 | 3.2570 ± 0.0464 |
| SWLDA      | 0.6466 ± 0.0535 | 0.7061 ± 0.0669 | 2.3832 ± 0.0243 | 0.8388 ± 0.0910 | 0.8858 ± 0.0820 | 2.3198 ± 0.0326 |
| SKLDA      | 0.6595 ± 0.0558 | 0.7624 ± 0.0513 | 2.3773 ± 0.0348 | 0.8454 ± 0.0868 | 0.9171 ± 0.0614 | 2.3343 ± 0.0306 |
| STDMA      | 0.6543 ± 0.0568 | 0.7123 ± 0.0674 | 3.2896 ± 0.0687 | 0.8410 ± 0.0855 | 0.8914 ± 0.0791 | 3.2217 ± 0.0287 |
| CNN        | 0.6592 ± 0.0557 | 0.7215 ± 0.0661 | 5.2227 ± 0.1007 | 0.7293 ± 0.0941 | 0.8426 ± 0.0790 | 6.2242 ± 0.0983 |

Note: Time indicates the duration for single-trial classification.

AUC = 0.9035 ± 0.0760). Similar to the offline results, the classification performance with single EEG feature was significantly lower than the benchmark for all methods (p < 0.01 for all comparisons, paired t-test). Although the performance was greatly improved by including eye-tracker features in classification, only SKLDA obtained both higher accuracy (0.8454 ± 0.0868) and higher AUC (0.9171 ± 0.0614) in a single-trial classification time of 2.3343 ± 0.0306 ms. Of note, the performance of CNN was significantly lower than that of the other four methods with a longer classification time (p < 0.001 for all comparisons, paired t-test). Subsequent validation was, therefore, performed only on the remaining four methods. Detailed individual results and statistical comparison results are presented in the supplementary materials.

To investigate how the number of training trials and the duration of fixation influenced the pseudo-online classification performance, the distribution of accuracy and AUC with different settings for these four selected classifiers is shown in Fig. 6. The rendered surface represents the classification performance obtained with fusion features, while the gray surface indicates the performance when using only eye-tracker data that was considered as the benchmark. Similar to the offline results, the performance of all methods improved monotonically with an increasing number of training trials. In contrast, the classification performance exhibited a complex dependent level for the setting of fixation duration, i.e., the best performance was not always obtained using a long fixation duration. Among the four methods, only SKLDA exhibited classification performance superior to the benchmark in most of the settings.

IV. DISCUSSION

In this study, we revealed an explicitly improved performance with the fusion of ERP and eye-tracker data in the single-trial classification of a free search task, both in offline and online analyzes. Our previous research proved the effectiveness of the block highlight display eye-controlled technique [45], which was further validated by the high target-detection rate in the performance of the active search paradigm in this study. The ERP components were also demonstrated in various spatial (occipital area) and temporal (100 and 300 ms) characteristics between target and nontarget detection. Six widely used classifiers were employed to verify the effectiveness of the hybrid BCI system. In the offline analysis, the classification approach with multimodal inputs of fixation duration and ERP significantly outperformed the method with unimodal brain/eye features. In addition, the multiple versions of LDA approach, except for BLDA, were proven effective in the single-trial classification, showing superiority in solving such problems with relatively simple inputs compared to a sophisticated neural network structure. For the online validation, we found that the fused feature still provided a
robust performance, while the various classification approaches exhibited divergent adaptability toward a limited number of samples and response time. SKLDA ranked top among all LDA classifiers in terms of both classification performance and application efficiency, that is, SKLDA obtained a high accuracy with a low computational complexity in a relatively short fixation duration.

In the offline analysis, we observed that the fused feature provided a much higher accuracy and AUC in the 10-fold classification. When ERP was taken as the only feature, the classification performance of the various classifiers was scattered. Although ERP signal could reflect cognitive processes with a high-temporal resolution, its low amplitude and sensitivity to various artifacts made it hard to extract stably and varied across subjects. This was consistent with previous findings that the prediction performance using ERP was mixed across studies. Specifically, SKLDA was proposed for single-trial classification of ERP-based BCI by Blankertz et al. [34], suggesting superior performance over ordinary LDA and SWLDA. Zhang developed STDA and demonstrated its superiority among several forms of LDA methods (LDA, SWLDA, and SKLDA) in ERP classification [39]. In our study, we found that among the LDA-based classifiers, RLDA outperformed other methods with an accuracy of 0.7793, and BLDA ranked last with an accuracy of 0.6716. As the most popular deep learning method in ERP-related studies [30], CNN also has an exceptional performance with the top accuracy of 0.7854. Briefly, the performance of classifier varied across studies and datasets for ERP-based classification.

However, when ERP and eye-tracker data were collectively adopted for classification input in this study, the performance of divergent classifiers was greatly improved compared with taking only single-model ERP or eye-tracker feature. Moreover, it was interesting to find that the accuracy and AUC of the various classification algorithms became similar. In our experimental protocol, the fixation duration tended to be higher when participants gazed on the target icon, consonant with the practical scenarios. The eye-tracker ensures a high temporal and spatial accuracy toward the gazing time and position, indicating a robust measure of underlying cognitive processes based on eye movement-related variables, such as fixation duration and saccade [46]. Therefore, compared with the single ERP signal, the fuse of fixation duration from the eye-tracker provided a relatively stable criterion toward cognitive states without large intersubject variability so that the input formulation was highly adaptable and the prediction performance tended to be similar across algorithms. However, compared with the single-fixation duration, the accuracy was also improved by feature fusing, which demonstrated that ERP was associated with the ongoing intention of target selection to enhance the recognition performance of traditional eye-tracker system. Among the different classification approaches, the performances of RLDA, SKLDA, SWLDA, and STDA were similar with the accuracies of 0.8802, 0.8783, 0.8761, and 0.8740, respectively. In addition, in the 10-fold offline analysis, CNN ranked second when fusing eye and brain features, possibly due to the abundant training sample and unlimited processing time. When the training set was massively reduced, most LDA-based classifiers outperformed CNN with fused features. Furthermore, in the profile of classification performance with increasing training sample, we found that the SKLDA, STDA, SWLDA, and RLDA provided an improving and robust accuracy for the fusion approach, and SKLDA was extraordinarily outstanding over other classifiers under various training set scales.

According to the review by Lotte et al. [29], LDA is one of the most prevalent forms of classifiers for EEG-based BCI, especially for online and real-time processing. The online analysis in this study further validated the improvement effect of feature fusion in the selected classification algorithms. Notably, we tested the practicability of the hybrid system by modulating the size of the training set and the length of every input sample (i.e., system response time). First, consistent with offline analysis, the accuracy and AUC remained at a high level when 80 trials were taken for training and decreased with the contraction of the training set. The size of the training sample determined the initial calibration time of the system and was directly related to the interaction convenience, but an extremely small training set, such as 20 trials, also immensely deteriorated the classification performance. In addition, to investigate the effect of system response time on the decoding of interaction intention, we truncated the fixation duration and corresponding ERP epoch of each sample, for the sake of simulating different response times for intention recognition of the hybrid system and evaluating the classification performance. The accuracy and AUC declined when response duration decreased, with a slight slope from 800 ms to 400 ms and a substantial drop below 400 ms. From the perspective of eye-tracker data, using a shorter recognition time made it more difficult to distinguish the intended selection from other conditions, such as long processing time of a stimulus for participants. A previous study showed that the dwell time of novices was typically between 450 and 1000 ms in gaze typing tasks, and decreased to 282 ms after repeated training [47]. In other eye selection-related ERP studies, the fixation duration usually had a long threshold, such as 1000 ms [48] or 2000 ms [49]. Considering that our participants underwent no pretraining, a proper recognition threshold above 400 ms could ensure a more accurate performance. From the aspect of ERP, it was obvious that a shorter EEG epoch after stimulus comprised fewer ERP components. Given that the predominate ERP was concentrated at 100 ms and 300 ms, as shown in Fig. 3, an epoch less than 400 ms might induce the loss of crucial information, especially when a latency shift occurs in P300 wave. However, the superiority of the fused feature compared with single-fixation duration was still observed across the whole range of system response times 300 to 800 ms as long as a considerable training set was implemented. A single eye-tracker was natural to use and could reach decent accuracy without much training, but only a few calibrations allowed the fusion of ERP and fixation duration to outperform the former. In addition, consistent with offline analysis, SKLDA still maintained the best performance for the fused feature among all classification methods, showing the largest area over single eye-tracker classification performance in Fig. 5, and a relatively lower requirement for the training set and response duration for the outperformance. A previous study proved that SKLDA remained effective when the training set was insufficient [34], so that it is practical
for use in real-time scenarios. For the computational time, we found that the single-trial process time was associated with the length of the input signal, especially in RLDA. The SKLDA and SWLDA classifiers required less computation cost in practical applications.

In this study, some factors should be considered when interpreting our results. First, the variance in classification performance indicated the individual differences in terms of our hybrid BCI system (as presented in Fig. 5), probably due to differences in background training and expertise [50]. To reduce assessment bias, a relatively large sample size (i.e., a total of 64 young and healthy university students) was involved to test our protocol. However, there is evidence suggesting that the P300 signal significantly differs between patients (e.g., schizophrenia [51] and amyotrophic lateral sclerosis [52]) and healthy participants, while the performance of LDA declined largely for the ERP classification of severely disabled individuals in real-life applications [28]. These findings lay the groundwork for developing a reasonable BCI that considers the cognitive characteristics of possible candidates for communication with this system.

Therefore, further study could optimize the algorithm model, stimulus presentation method, and feature extraction according to individual factors (healthy or patient) to meet the needs of individuals. Second, in the design of the present interaction interface, the icons were arranged in isolation and regularly with rigid distance as an array, to which the users would become accustomed as they progressed through the experiment. A real ecological application should display arbitrary targets, providing more oddball stimuli to the user and eliciting stable ERP responses. At the same time, the stimulus was expressed in images in this study, while other types of interaction items, such as words, are associated with different predominant ERP components [46]. Future studies could introduce other forms of selection targets to test the generalization of the hybrid system. In addition, this study observed a relatively limited increase in performance by fusing eye-tracker and ERP data compared with a single eye-tracker signal as the feature. Considering the computational complexity, only the time domain wave was selected targets to test the generalization of the hybrid system.

V. CONCLUSION

In this study, we introduced an eye-brain hybrid BCI interaction system and assessed the performance in a customized free visual search paradigm. In comparison with the single-model EEG or eye-tracker features, the proposed hybrid BCI system achieved a better performance in both offline and online conditions. Furthermore, practical validation across six widely used classification methods showed that the SKLDA method could maintain superior performance under conditions of few training sets and fast response times. In summary, our study sheds new insights into the approach of hands-free HCI and provides novel and practical solutions to intention detection in real-world scenarios.

REFERENCES

[1] H. Grahn and T. Kujala, “Impacts of touch screen size, user interface design, and subtask boundaries on in-car task’s visual demand and driver distraction,” Int. J. Hum.-Comput. Stud., vol. 142, 2020, Art. no. 102467.

[2] M. Beckhacker, H. Syrek, M. Elster, S. Schmitt, and H. Roehl, “Evaluating usability of a touchless image viewer in the operating room,” Appl. Clin. Informat., vol. 11, pp. 088–094, 2020.

[3] J. Borji, A. Lennartz, and M. Pomplun, “What do eyes reveal about the mind?: Algorithmic inference of search targets from fixations,” Neurocomputing, vol. 149, pp. 788–799, 2015.

[4] F. Koochaki and L. Najafizadeh, “Eye gaze-based early intent prediction utilizing CNN-LSTM,” in Proc. Conf.: Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., 2019, vol. 2019, pp. 1310–1313.

[5] Y. Nie and X. Ma, “Gaze Based Implicit Intention Inference With Historical Information of Visual Attention for Human-Robot Interaction.” Berlin, Germany: Springer, 2021, pp. 293–303.

[6] Y. Pan et al., “Lane-change intention prediction using eye-tracking technology: A systematic review,” Appl. Ergonom., vol. 103, 2022, Art. no. 103775.

[7] G. Zelinsky, Y. Peng, and D. Samaras, “Eye can read your mind: Decoding gaze fixations to reveal categorical search targets,” J. Vis., vol. 13, no. 14, pp. 1–13, 2013.

[8] H. Sattar, M. Fritz, and A. Bulling, “Deep gaze pooling: Inferring and visually decoding search intents from human gaze fixations,” Neurocomputing, vol. 387, pp. 369–382, 2020.

[9] B. Hu, X. Liu, W. Wang, R. Cai, F. Li, and S. Yuan, “Prediction of interaction intention based on eye movement gaze feature,” in Proc. IEEE 8th Joint Int. Conf. Technol. Artif. Intell. Conf., 2019, pp. 378–383.

[10] A. Plopski, T. Hirzle, N. Norouzi, L. Qian, G. Bruder, and T. Langlotz, “The eye in extended reality: A survey on gaze interaction and eye tracking in head-worn extended reality,” ACM Comput. Surv., vol. 55, no. 3, pp. 1–39, 2022.

[11] P. Monteiro, G. Gonçalves, H. Coelho, M. Meiro, and M. Bessa, “Hands-free interaction in immersive virtual reality: A systematic view,” IEEE Trans. Vis. Comput. Graph., vol. 27, no. 5, pp. 2702–2713, May 2021.

[12] D. G. Zhao et al., “A passive BCI for monitoring the intentionality of the gaze-based moving object selection,” J. Neural Eng., vol. 18, no. 2, 2021, Art. no. 026001.

[13] C. Biele, Eye Movement. Cham, Switzerland: Springer International Publishing, 2022, pp. 23–37.

[14] L. Shi, C. Copot, and S. Vanlanduit, “Gazeend: Detecting visual intention in gaze-based human-robot interaction,” Robotics, vol. 10, no. 2, 2021, Art. no. 068.

[15] I. Stancin, M. Cifrek, and A. Jovic, “A review of EEG signal features and their application in driver drowsiness detection systems,” Sensors, vol. 21, 2021, Art. no. 3786.

[16] G. N. Dimitrakopoulos et al., “Functional connectivity analysis of mental fatigue reveals different network topological alterations between driving and vigilance tasks,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 26, no. 4, pp. 620–626, Apr. 2022.

[17] J. Atkinson and D. Campos, “Improving BCI-based emotion recognition by combining EEG feature selection and Kernel classifiers,” Expert Syst. With Appl., vol. 47, pp. 35–41, 2016.

[18] H. Qi, Y. Xue, L. Xu, Y. Cao, and X. Jiao, “A speedy calibration method using Riemannian geometry measurement and other-subject samples on a P300 speller,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 26, no. 3, pp. 602–608, Feb. 2018.

[19] Y. Si et al., “Predicting individual decision-making responses based on single-trial EEG,” NeuroImage, vol. 206, 2020, Art. no. 0116333.

[20] X. Gao, Y. Wang, X. Chen, and S. Gao, “Interface, interaction, and intelligence in generalized brain–computer interfaces,” Trends Cogn. Sci., vol. 25, no. 8, pp. 671–684, 2021.

[21] J.-S. Kang, U. Park, V. Gonuguntla, K. Veluvolu, and M. Lee, “Human implicit intent recognition based on the phase synchrony of EEG signals,” Pattern Recognit. Lett., vol. 66, pp. 144–152, 2015.

[22] J. Kim, C. Kim, and Y. Lee, “Electroencephalography-based intention monitoring to support nuclear operators’ communications for safety-relevant tasks,” Nucl. Technol., vol. 207, pp. 1753–1767, 2020.

[23] L. N. Kaunitz, J. E. Kamienowski, A. Varatharajah, M. Sigman, R. Q. Quiroga, and M. J. Ison, “Looking for a face in the crowd: Fixation-related potentials in an eye-movement visual search task,” NeuroImage, vol. 89, pp. 297–305, 2014.
[24] H. Devillez, N. Guyader, and A. Guérin-Dugué, “An eye fixation–related potentials analysis of the P300 potential for fixations onto a target object when exploring natural scenes,” J. Vis., vol. 15, no. 13, pp. 20-20, 2015.

[25] D. Kalika, L. Collins, K. Caves, and C. Throckmorton, “Fusion of P300 and eye-tracker data for spelling using BCI2000,” J. Neural Eng., vol. 14, no. 5, 2017, Art. no. 056010.

[26] J.-S. Choi, J. W. Bang, K. R. Park, and M. Whang, “Enhanced perception of user intention by combining EEG and gaze-tracking for brain–computer interfaces (BCIS),” Sensors, vol. 13, no. 3, pp. 3454–3472, 2013.

[27] D. J. Krusienski, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, “Toward enhanced P300 speller performance,” J. Neurosci. Methods, vol. 167, no. 1, pp. 15–21, 2008.

[28] V. Martínez-Cagial, J. Gomez-Pilar, D. Alvarez, and R. Hornero, “An asynchronous P300-based brain-computer interface web browser for severely disabled people,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 25, no. 8, pp. 1332–1342, Oct. 2016.

[29] F. Lotte et al., “A review of classification algorithms for eeg-based brain–computer interfaces: A 10 year update,” J. Neural Eng., vol. 15, no. 3, 2018, Art. no. 031005.

[30] A. Craik, Y. He, and J. L. Contreras-Vidal, “Deep learning for electroencephalogram (EEG) classification tasks: A review,” J. Neural Eng., vol. 16, no. 3, 2019, Art. no. 031001.

[31] A. Maddirala and K. Velavolu, “Eye-blink artifact removal from single channel EEG with k-means and SSA,” Sci. Rep., vol. 11, 2021, Art. no. 11043.

[32] C. Brunner, A. Delorme, and S. Makeig, “Eeglab—An open source Matlab toolbox for electrophysiological research,” Biomed. Eng./Biomedizinische Technik, vol. 58, no. SI-1-Track-G, 2013, Art. no. 000010151520134182.

[33] X. Xiao, M. Xu, J. Jin, Y. Wang, T.-P. Jung, and D. Ming, “Discriminative canonical pattern matching for single-trial classification of ERP components,” IEEE Trans. Biomed. Eng., vol. 67, no. 8, pp. 2266–2275, Dec. 2019.

[34] B. Blankertz, S. Lemm, M. Treder, S. Haufe, and K.-R. Müller, “Single-trial analysis and classification of ERP components—A tutorial,” NeuroImage, vol. 56, no. 2, pp. 814–825, 2011.

[35] F. Zha, J. Gao, J. Yang, and N. Ye, “Neighborhood linear discriminant analysis,” Pattern Recognit., vol. 123, 2022, Art. no. 108422.

[36] I. Pardoe, Applied Regression Modeling, vol. 326. Hoboken, NJ, USA: Wiley, 2020.

[37] X. Lei, P. Yang, and D. Yao, “An empirical Bayesian framework for brain–computer interfaces,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 17, no. 6, pp. 521-529, Dec. 2009.

[38] C. Vidaurre, N. Krämer, B. Blankertz, and A. Schlögl, “Time domain parameters as a feature for EEG-based brain–computer interfaces,” Neural Netw., vol. 22, no. 9, pp. 1313–1319, 2009.

[39] Y. Zhang, G. Zhou, Q. Zhao, J. Jin, X. Wang, and A. Cichocki, “Spatial-temporal discriminant analysis for ERP-based brain–computer interface,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 21, no. 2, pp. 233–243, Mar. 2013.

[40] H. Cecotti and A. Graser, “Convolutional neural networks for P300 detection with application to brain-computer interfaces,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 3, pp. 433–445, Mar. 2011.

[41] A. Sharma and K. K. Pulivail, “A deterministic approach to regularized linear discriminant analysis,” Neurocomputing, vol. 151, pp. 207–214, 2015.

[42] D. J. Krusienski et al., “A comparison of classification techniques for the P300 speller,” J. Neural Eng., vol. 3, no. 4, 2006, Art. no. 299.

[43] U. Hoffmann, J.-M. Vesin, T. Ebrahimi, and K. Diserens, “An efficient P300-based brain–computer interface for disabled subjects,” J. Neurosci. Methods, vol. 167, no. 1, pp. 115–125, 2008.

[44] J. Schäfer and K. Strimmer, “A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics,” Statist. Appl. Genet. Mol. Biol., vol. 4, no. 1, 2005, Art. no. 32.

[45] Y. Pan, X. Ge, L. Ge, and J. Xu, “Using eye-controlled highlighting techniques to support both serial and parallel processing in visual search,” Appl. Ergonom., vol. 97, 2021, Art. no. 103522.

[46] O. Dimigen and B. Ehinger, “Regression-based analysis of combined EEG and eye-tracking data: Theory and applications,” J. Vis., vol. 21, pp. 1–30, Jan. 2021.

[47] P. Majoranta, U.-K. Ahola, and O. Špakov, “Fast gaze typing with an adjustable dwell time,” in Proc. SIGCHI Conf. Hum. Factors Comput. Syst., 2009, pp. 357–360.

[48] J. Protzak, K. Ihme, and T. O. Zander, “A passive brain-computer interface for supporting gaze-based human-machine interaction,” in Proc. Int. Conf. Universal Access Hum.-Comput. Interact., 2013, pp. 662–671.

[49] T. O. Zander and C. Kothe, “Towards passive brain–computer interfaces: Applying brain-computer interface technology to human-machine systems in general,” J. Neural Eng., vol. 8, no. 2, 2011, Art. no. 025005.

[50] K. Dhindsa, D. Carone, and S. Becker, “An open-ended approach to BCI: Embracing individual differences by allowing for user-defined mental commands,” in Proc. Front. Comput. Neurosci. Conf. Abstr.: German-Japanese Adaptive BCI Workshop, 2013, doi: 10.3389/conf.fncosc.2015.56.00015.

[51] R. M. Reinhart, D. H. Mathalon, B. J. Roach, and J. M. Ford, “Relationships between pre-stimulus gamma power and subsequent P300 and reaction time breakdown in schizophrenia,” Int. J. Psychophysiol., vol. 79, no. 1, pp. 16–24, 2011.

[52] A. Riccio et al., “On the relationship between attention processing and P300-based brain computer interface control in amyotrophic lateral sclerosis,” Front. Hum. Neurosci., vol. 12, 2018, Art. no. 165.