Implicit Search Intent Recognition using EEG and Eye Tracking: Novel Dataset and Cross-User Prediction

Mansi Sharma
DFKI, Saarland Informatics Campus
Saarbrücken, Germany
mansi.sharma@dfki.de

Shuang Chen
DFKI, Saarland Informatics Campus
Saarbrücken, Germany
shuang.chen@dfki.de

Philipp Müller
DFKI, Saarland Informatics Campus
Saarbrücken, Germany
philipp.mueller@dfki.de

Maurice Rekrut
DFKI, Saarland Informatics Campus
Saarbrücken, Germany
maurice.rekrut@dfki.de

Antonio Krüger
DFKI, Saarland Informatics Campus
Saarbrücken, Germany
antonio.krueger@dfki.de

ABSTRACT

For machines to effectively assist humans in challenging visual search tasks, they must differentiate whether a human is simply glancing into a scene (*navigational intent*) or searching for a target object (*informational intent*). Previous research proposed combining electroencephalography (EEG) and eye-tracking measurements to recognize such search intents implicitly, i.e., without explicit user input. However, the applicability of these approaches to real-world scenarios suffers from two key limitations. First, previous work used fixed search times in the informational intent condition - a stark contrast to visual search, which naturally terminates when the target is found. Second, methods incorporating EEG measurements addressed prediction scenarios that require ground truth training data from the target user, which is impractical in many use cases. We address these limitations by making the first publicly available EEG and eye-tracking dataset for navigational vs. informational intent recognition, where the user determines search times. We present the first method for cross-user prediction of search intents from EEG and eye-tracking recordings and reach 84.5% accuracy in leave-one-user-out evaluations - comparable to within-user prediction accuracy (85.5%) but offering much greater flexibility.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**.

KEYWORDS

Intent recognition; Dataset; Multimodal fusion; Eye-Tracking; EEG

ACM Reference Format:
Mansi Sharma, Shuang Chen, Philipp Müller, Maurice Rekrut, and Antonio Krüger. 2023. Implicit Search Intent Recognition using EEG and Eye Tracking: Novel Dataset and Cross-User Prediction. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23), October 09–13, 2023, Paris, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3577190.3614166

ICMI '23, October 09–13, 2023, Paris, France

ACM ISBN 979-8-4007-0055-2/23/10.
https://doi.org/10.1145/3577190.3614166

Figure 1: Examples of the visual scenes we created for our study. In total, we created 120 unique scenes.

1 INTRODUCTION

Visual search is ubiquitous in daily life. For example, searching for a desired chocolate bar on a supermarket shelf or the wrench in a cluttered workshop. Approaches that can automatically - and without explicit user input - infer humans' search targets have the potential to assist humans and avoid unnecessary frustration or delays [4, 32]. For an assistive system to effectively help humans in visual search tasks, it has to solve one key problem before even considering which target the user might be searching for: the system needs to be able to recognize that the user is searching at all (i.e., having an *informational intent*) and not simply looking at the scene with no such purpose in mind (called *navigational intent* in the literature) [17, 18, 33]. Previous studies on the recognition of navigational versus informational intent from EEG and eye-tracking data proved the general feasibility of this task [17–19, 30]. However, they suffer from limitations concerning the study design and the prediction scenarios that reduce their applicability to the real world.

Using fixed time limits (usually 5s) in visual search tasks is a common practice in previous studies [19, 30], as it allows for better control over the experiment and ensures that all users complete the task in a consistent amount of time. However, this approach does not allow for the evaluation of unrestricted search duration, where user has no time constraints. This type of design allows for a more naturalistic approach to studying visual search behavior. It can provide insight into how participants prioritize search strategies and allocate their attention over time.

Concerning the prediction scenario, previous studies on multi-modal navigational vs. informational intent recognition from EEG and eye-tracking are limited to cross-user scenarios where data from a given user can appear both in the test- and training sets [30]. In contrast, practical scenarios often require a model to apply to unseen target users without needing to collect training data. A final but crucial limitation to achieve progress in search intent recognition based on EEG and eye-tracking is the need for a publicly available dataset that will provide us valuable insights into real-world search behavior and help us generalize our intent recognition models across a larger group of users.

Our work addresses these shortcomings by proposing a novel EEG- and eye-tracking dataset for navigational vs. informational intent classification where the search duration depends entirely on the time it takes to find the target. We intentionally select the industrial workplace scenarios where visual search often helps workers quickly and accurately identify equipment, components, and tools, which can improve efficiency, safety, and productivity. Furthermore, we propose the first multi-modal approach to the best of our knowledge for navigational vs. informational intent prediction in a strict leave-one-user-out evaluation scenario. Our experiments reveal that feature selection is crucial in achieving high performance across users, unlike within-user prediction. Our specific contributions are threefold:

(1) We present MindGaze[1], the first publicly available dataset for informational vs. navigational intent prediction with a large variety of different workplaces in Unity [15] (see Figure 1 for examples). The dataset consists of EEG and eye-tracking recordings of 15 participants, performing 3600 trials.
(2) We present the first approach for cross-user (leave-one-user-out) prediction of navigational vs. informational intent from EEG and eye-tracking, revealing that appropriate choice of features plays a crucial role in cross-user prediction.
(3) We conduct extensive evaluations for within-user and cross-user scenarios and compare different multi-modal fusion strategies (early, late, and hybrid fusion). Subsequently, we perform experiments with the smaller windows of 0.5s, 1s, 1.5s, and 2s to show the potential for near real-time intent prediction applications.

2 RELATED WORK

The implicit (i.e., without explicit user input) prediction of user intents are of great interest in human-machine interaction (HMI), as it can help to adapt the behavior of machines without the overhead and discomfort associated with explicit input. Implicit intent recognition systems often rely on challenging-to-interpret measurements like eye-tracking, pupil dilation, or EEG [25, 40]. Previous research worked on implicitly predicting web users' click intents [34], search targets [4, 32], improving the performance of Motor Imagery task[8] their next focus of attention [26, 36, 39], or choice of ingredients [16] when preparing a meal. In this paper, we focus on the task of distinguishing between navigational intent and informational intent, a pre-requisite to assist humans when searching for a target object in cluttered environments [19]. In the following, we discuss previous work on this task. Subsequently, we compare datasets for navigational vs. informational intent recognition recorded in previous work.

Table 1: Datasets for search intent recognition. Availability shows whether the dataset is publicly available or not. Size is a product of the total number of users, scenes shown to each user, and intents. NA indicates missing information.

Reference	Availability	Users	Modality	Size
Kang et al. [19]	✗	10	EEG	500
Jang et al. [18]	✗	52	Eye	668
Huang et al. [16]	✗	13	Eye	276
Liang et al. [23]	✗	18	Eye	NA
Jang et al. [17]	✗	100	Eye	2400
Park et al. [30]	✗	8	EEG + Eye	400
Ours [MindGaze]	✓	15	EEG + Eye	3600

2.1 Navigational versus Informational Intent Recognition

While predicting the target of visual search is an increasingly popular task in implicit intent recognition [4, 32, 35, 37], few works addressed the prerequisite of any practical support system for search target prediction, namely the recognition of navigational vs. informational intent [18, 19, 30].

Kang et al. [19] performed EEG-based classification of navigational and informational intents in everyday images. Their study setup presented an image in the navigational intent condition for 5s, followed by the informational intent condition (search task) on the same image for a fixed search time of 5s. The authors analyzed the differences in phase-locking value (PLV) to classify intents in a within-user prediction scenario. The results showed severe overfitting with a significant gap between the train (> 99%) and test accuracy (57.1% to 77.4%).

Using the same concept of sequential navigational and informational intent tasks, Jang et al. [18] classified human implicit navigational and informational intent based on the eyeball movement pattern and pupil size variation characteristics in a visual search task. For evaluation, authors used 25 users for training and other 27 users for testing, reaching a mean accuracy of 85.26% with an SVM classifier. In a follow-up work [17], the authors performed hierarchical classification to further differentiate states in task-oriented visual searches, such as intent generation, intent maintenance, and intent disappearance. Authors collected data from 100 users and used 40 random samples for training and 60 for testing, reaching

[1] https://doi.org/10.5281/zenodo.8239061

Figure 2: (Left) Navigational scan path, (Right) Informational scan path, search target: `Screwdriver`

90.36% with an SVM classifier. In both studies [17, 18], users had to search for different numbers of objects in indoor- and outdoor scenes, e.g., the cup and bottle in an indoor image or all humans in an outdoor image. Although [17, 18] did not explicitly mention fixed search times, they used the same experimental setup as [19], which has a fixed amount of time (5s) to perform the search task.

In a follow-up work to [18], Park et al. [30] proposed a multi-modal approach combining EEG and eye-tracking features while following the same experimental design principle as in [19], i.e., using 5s for both navigational and informational intent conditions. The authors neither followed a pure within-user nor a pure cross-user evaluation approach. They trained their model on several users, but samples from the same user could appear both in training- and the test set. With an early fusion approach, they improved classification accuracy by 5% over uni-modal baselines. While these results indicate the utility of a multi-modal approach, a comparison between different fusion methods (early, late, and hybrid fusion) was not presented. Furthermore, the applicability of their approach is limited by the fixed search times and by not employing a strict cross-user prediction scenario, i.e., where data from a single user can only be either in the train- or in the test set.

In contrast to previous work, search times in our study are entirely determined by the time it takes participants to find the target. Furthermore, we, for the first time, study multi-modal prediction of navigational vs. informational intents in a strict cross-user evaluation scenario.

2.2 Datasets for Navigational versus Informational Intent Recognition

We provide an overview of the datasets used by previous works on navigational vs. informational intent recognition in Table 1. Most datasets consist of eye-tracking recordings exclusively [16–18, 23], only two datasets involving EEG recordings were presented in previous work [19, 30], one of them also containing eye-tracking recordings [30]. The number of users in eye-tracking datasets is usually higher than those in datasets with EEG recordings (8-10 users), likely due to the time-consuming procedure required to set up EEG recordings. The number of trials varies significantly across previously recorded datasets, ranging from 276 to 2400 trials. With 3600 trials our novel EEG and eye-tracking recordings dataset has a much higher number of trials on informational vs. navigational

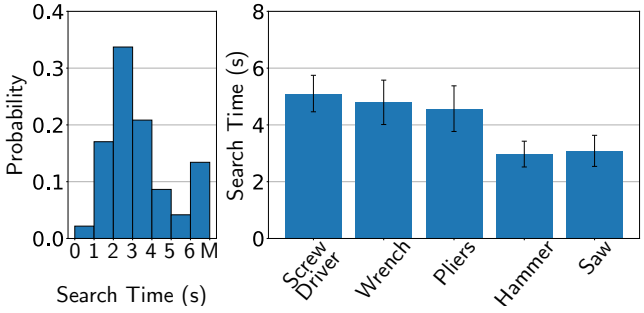

Figure 3: (Left) Distribution of search times for all users. All search times larger than 6s are accumulated in the rightmost bar, where M is the max duration. (Right) Search durations for individual tools.

recognition than any previous dataset. We recorded 15 users, making it the largest dataset for navigational vs. informational intent recognition with EEG recordings regarding the number of users. Most importantly, none of the previously recorded datasets for navigational vs. informational intent recognition is publicly available, severely limiting progress on this task.

3 DATASET

In the following, we describe the data recording and descriptive statistics on the dataset. We provide detailed statistics at the dataset website (see section 1).

3.1 Data Recording

Participants. We recruited 15 volunteers (5 female and 10 male) aged between 20 and 35 years old ($\mu = 27.46$, $\sigma = 4.22$). All participants had normal or corrected to normal vision, and none were exposed to the study design before. The study was approved by our institution's Ethics and Hygiene Board[2].

Hardware Setup. To display visual stimuli, we used a monitor with a resolution of 1920 x 1080 and screen brightness of 300 cd/m2. We used the LiveAmp 64 channel system[3] by Brain Products to

[2]https://erb.cs.uni-saarland.de/
[3]https://brainvision.com/products/liveamp-64/

Table 2: List of extracted features with PyEEG and added statistical features

Feature name	Description
Power spectral intensity	distribution of signal power over frequency bands: delta, theta, alpha, beta, and gamma
Petrosian Fractal Dimension	ratio of number of self-similar pieces versus magnification factor
Hjorth mobility and complexity	mobility represents the proportion of the standard deviation of the power spectrum Complexity represents the change in frequency
Higuchi Fractal Dimension	computes fractal dimension of a time series directly in the time domain
Detrended Fluctuation Analysis	designed to investigate the long-range correlation in non-stationary series
Skewness	measure of asymmetry of an EEG signal
Kurtosis	used to determine if the EEG data has peaked or flat with respect to the normal distribution
Minimum, Maximum, and Standard deviation	measure of variability of an EEG signal

record the EEG signals with a sampling frequency of 500 Hz. Electrodes were placed according to the 10-20 international electrode placement system [1]. To record eye-tracking data, we used wireless Tobii pro fusion[4] attached to the monitor's lower bezel with a sampling frequency of 250 Hz. The device was calibrated at the start of the experiment for each participant, using two coordinate systems. One is a 2D system that spans the monitor with $(0, 0)$ in the top right corner of the experiment setup monitor screen and $(1, 1)$ in the bottom left. The second is a 3D coordinate system for the experiment room, which measures the distance from the eye to the eye-tracker. EEG and eye-tracking data were synchronized via particular time-locked events, for example, the start and end of the navigational and informational stimulus.

Stimuli. We designed 120 realistic industrial scenes in Unity to simulate industrial environments, such as assembly units, manufacturing and production facilities, industrial labs, garage and repair workshop, and many more (see Figure 1 for examples). While crafting the scenes, we portrayed different clutter layers through the chaotic arrangement of parts of machines, tools, workbench, and others, as evident in Figure 1. We incorporated various tool locations (inside a cupboard, on the floor, and in some unexpected areas) and orientations in the industrial scenes to create different levels of complexity. We used five tools - Hammer, Pliers, Saw, Screwdriver, and Wrench - as our target stimulus.

Procedure. Participants were given a general introduction to the study, where we explained the experiment's purpose and procedure and discussed the anonymity and privacy of their collected data. Next, we gathered participants' consent and asked them to complete a demographic questionnaire. Participants were seated in a comfortable chair such that the distance between the user and the screen was 60cm. We mounted the EEG cap on their head, filling the electrodes with gel. Following common practice [13] for noise reduction, we kept electrode impedances below 25 kΩ throughout the experiment. Overall, the preparation time was about 30 min. Our experimental design followed previous research [17, 18, 30], which presents the scene first in the navigational intent condition, followed by the search task (informational intent condition). This approach mirrors a typical workplace scenario where individuals often start searching for items within a scene that is already familiar to them. In particular, a single recording of the experiment consisted of 3 steps. In Step 1, we presented the scene for 5s. The participant glanced over the scene without knowledge of the target tool. In Step 2, we show the target tool for 5s. In the final Step 3, the participant searches for the displayed target tool in the scene until the tool is found. Once the participant fixated on the target tool for more than 1s, a red highlight appeared around the tool, which changed to green in the case of prolonged fixation of 1s , indicating that the tool was successfully located. The 120 unique scenes were split randomly into four equal sessions, with breaks in-between sessions. The sequence of the industrial scenes and the target objects are randomized with an average experiment duration of 90 min per participant.

[4]https://www.tobii.com/products/eye-trackers/screen-based/tobii-pro-fusion

3.2 Descriptive Statistics

We collected the data from 15 users who viewed 120 scenes each. The data synchronization is performed based on the available timestamps. We pre-processed the collected dataset and filtered out scenes (an average of 18 scenes per user) due to technical issues with EEG and Eye-tracking data recording. Figure 2 showed the scan paths of an example participant. In the navigational condition, Figure 2 (Left), the users' attention is spread across the scene. In Figure 2 (Right), the user performs a more focused search for finding the target, which is a `Screwdriver`. Unlike existing studies [18, 19, 30], the search times in our informational intent condition were entirely determined by the time participants took to find the target. To illustrate the importance of this choice, Figure 3 (left) visualizes the distribution of search times. The search time follows a left-skewed Gaussian distribution with a peak between 2s and 3s. While the probability diminishes continuously for larger search times. Overall, it is evident that the variation in search times is large, supporting our approach of not setting a fixed search time in contrast to previous work [19]. Figure 3 (right) describes the search time with respect to five different target tools for all users. The search time appears to be connected with the size of the tool.

For example, `Screwdriver`, `Wrench`, and `Pliers` have comparable sizes and hence, comparable search times.

4 METHODS

In the following, we detail pre-processing and feature extraction pipelines for EEG and eye-tracking data and present our classification and multi-modal fusion approaches. Figure 4 showed an overview of data collection, signal processing, and classification.

4.1 Pre-processing

Eye data pre-processing. As humans extract visual information during fixations, visual search tasks are commonly analyzed based on fixations [4, 35]. We used the I-VT filter, a velocity-threshold fixation detection approach [29]. Overall, we follow the seven-step approach taken by [38]: *Gap fill-in* helps replace missing samples, which might occur due to unforeseen disturbances causing short gaps in the data. *Eye selection* averages the position data from the left and the right eye. We applied *Noise reduction* to smooth out the noise while preserving the features based on the moving median method. We then used the *Velocity calculator* to associate each gaze sample with a velocity. To classify the sample as either a part of fixation or not, we applied the *I-VT classifier*. Subsequently, we employed *Merge adjacent fixations* to correct erroneously split fixations due to noise. Lastly, using *Discard short fixations*, we discard fixations which are too short to be relevant in visual search.

EEG data pre-processing. We applied high-pass filters at a cutoff frequency of 1 Hz to clean the EEG data, followed by a notch filter between 48 Hz and 52 Hz [28] and a low-pass filter at a cutoff frequency of 40 Hz [20]. Then bad channels were removed and interpolated using signals in good locations based on the spherical interpolation method. In the next step, we referenced channels to a common average reference [24]. To reduce the correlation between electrodes, we executed independent component analysis using the second Order Blind Identification (SOBI) algorithm [12], followed by subsequent automated IC_Label rejection (muscle, heart, and eye components with a 95% threshold). Lastly, we extracted specific time windows from the continuous EEG signal [22], with reference to the stimulus onset in the pre-processed data. We took equal duration for Navigational and Informational Intent within each sample, a prerequisite of one of the feature extraction methods (Common Spatial Pattern). We followed the same approach for other feature extraction methods to ensure a fair comparison.

4.2 Feature Extraction

We used EEG and eye-tracking to extract feature sets from navigational and informational components of each trial, enabling us to tailor our approach for optimal results.

EEG based features. We used two different feature extraction methods for EEG signals. Firstly, we use PyEEG, an open-source Python module for EEG feature extraction [2] in the frequency and time domain. Additionally, we included statistical features, and in total, we extract 15 features per channel, resulting in 960 features (64 channels * 15 features). Table 2 shows the extracted features for each EEG channel. As the number of PyEEG features is much larger than our extracted gaze features, we applied principal component analysis (PCA) for dimensionality reduction, similar to [34], we select the principal components where the explained variance is 90%. Secondly, we used Common Spatial Pattern (CSP) to extract features from EEG data in a maximally discriminative manner [7, 11]. We used default parameters from the MNE toolbox [14], resulting in 4 spatial features.

Eye-tracking features. Table 3 shows the list of features adapted from the existing state-of-the-art [3, 6]. Features are based on fixation events, saccadic eye movements, and the scanned area. Some features are time-normalized by the total time covered by the provided gaze data. We compute total_time as the difference between the recording's last and first timestamp. After extracting the features, we follow a similar strategy as in [3] to reduce feature multi-collinearity by performing hierarchical clustering on the feature's Spearman rank-order correlations. We set the distance threshold to 0.2 and use one feature per cluster i.e., `fixn_dur_avg`, `fixn_dur_sd`, `scan_hv_ratio`, `fixn_dur_sum`, `scan_speed_h`, `fixns_per_box_area`, `avg_sacc_length`, `scan_speed_v`.

4.3 Classification and Fusion Techniques

In our study, we use three popular classification algorithms, including Support Vector Machine (SVM), Random Forest (RF), and Naïve Bayes (NB), which are widely used in EEG- and eye-tracking studies [18, 19, 30, 41]. Further, we explored three distinct fusion mechanisms: early, late, and hybrid strategies. For every fusion approach, we first performed a min-max scaling on the input features to harmonize the scale of features within- and across modalities.

Early fusion. This is the simple concatenation of EEG- and eye feature representations and inputs them as a single vector into the classifier. It aims to exploit dependencies between features. The input format from both modalities must be temporally compatible so that it is possible to combine them. This fusion technique is common in previous studies on navigational vs. informational intent classification [30].

Late fusion. This strategy trains classifiers for individual modalities - i.e., an EEG classifier with EEG features as input, and an eye tracking classifier with eye tracking features as input. To fuse the outputs of the modality-specific classifiers, we trained an additional classifier producing the final decision.

Hybrid fusion. The fusion happens at the shared representation layer from late and early fusion. It was successfully used in emotion recognition [10, 27] and multimedia event detection [21], to name a few. In addition to the two uni-modal classifier outputs in late fusion it also utilizes the output of the early fusion classifier described above. It then trains an additional classifier that predicts the final decision based on these three classifier outputs (uni-modal EEG, uni-modal eye, multimodal early fusion).

For all fusion approaches and fusion stages, we evaluated different classifiers (SVM, NB, RF) in our experiments. SVM always performed best in these experiments, so we chose it as our classifier for all modalities and fusion stages.

5 EVALUATION

We present classification results for cross-user and user-specific prediction and evaluate different feature sets and multi-modal fusion methods. In the cross-user scenario, we perform leave-one-user-out cross-validation. In the within-user scenario, we report mean

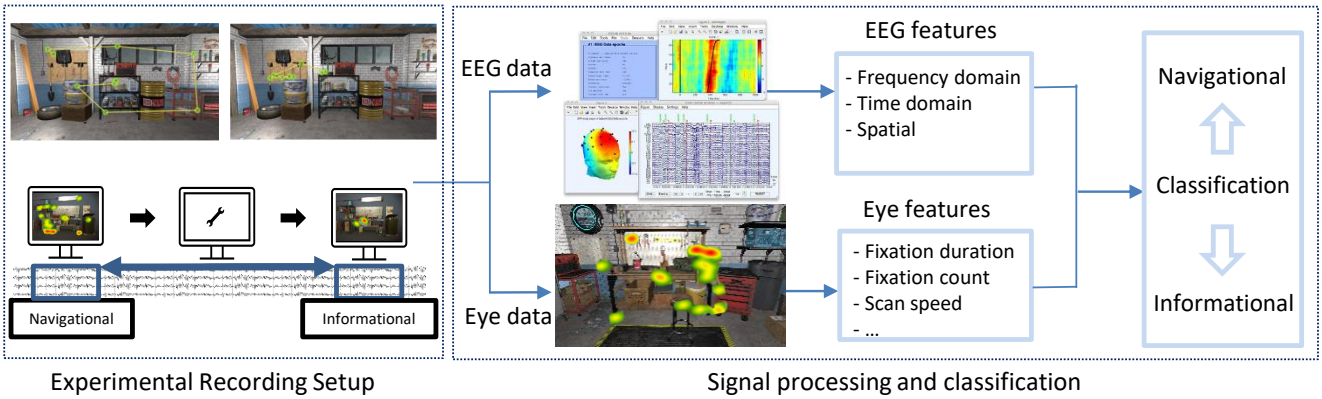

Figure 4: Overview of our approach to implicit classification of navigational versus informational intents.

Table 3: Overview of eye movement features based on fixation events, saccadic eye movements, and the scanned area

	Feature	Description
fixation-based	fix_n	Number of fixations
	fixn_dur_sum	Sum of fixation durations
	fixn_dur_avg	Mean of fixation durations
	fixn_dur_sd	Standard deviation of fixation durations
saccade-based	scan_dist_h	Sum of horizontal amplitudes of all saccades, normalized by a factor w, i.e., screen width
	scan_dist_v	Sum of vertical amplitudes of all saccades, normalized by a factor h, i.e., screen height
	scan_dist_euclid	Sum of Euclidean distances of normalized amplitudes of all saccade
	scan_hv_ratio	Ratio of horizontal to vertical amplitudes: scan_dist_ h/scan_dist_v
	avg_sacc_length	Average saccade amplitude: scan_dist_euclid/(fix_n - 1)
	scan_speed_h	Horizontal saccade velocity: scan_dist_h/scan_time
	scan_speed_v	Vertical saccade velocity: scan_dist_v/scan_time
	scan_speed	Saccade velocity: scan_dist_euclid/scan_time
area-based	box_area	Area spanned by summed saccade amplitudes: scan_dist_h × scan_dist_v
	box_area_per_time	The box_area normalized by the scan time: box_area/scan_time
	fixns_per_box_area	Number of fixations per scanned area: fixn_n/box_area
	hull_area_per_time	The hull area normalized by the scan time: hull_area/scan_time
	fixns_per_hull_area	Number of fixations per convex hull area: fixn_n/hull_area

accuracy averaged over 10 random train-test splits to estimate the generalization performance for each user. For each train-test split, we perform feature selection (see Section 4.2) and parameter tuning via grid search cross-validation only on the train data and use the test set only for prediction. In all evaluation scenarios, the random baseline is at 50% accuracy.

5.1 Cross-user Prediction

We present cross-user prediction results for the leave-one-user-out evaluation scenario in Figure 5a. The best approach relies on PyEEG and eye-tracking features joined by early fusion, reaching 84.5% accuracy. This is a significant increase over eye-tracking features with 60.6% accuracy and a moderate increase over PyEEG-only features with 81.3% accuracy, documenting the effectiveness of a multi-modal approach. Interestingly, the CSP features performs far worse than the PyEEG dataset. Unlike PyEEG, CSP performs below eye-tracking features, even in each multi-modal scenario. There are no substantial differences between late and hybrid fusion strategies. Only early fusion showed an increase over the mono-modal EEG baseline. Figure 5a showed the error bars with 95% confidence interval. We performed a paired t-test, and the calculated p-value of $6.1e-07$ is less than the significance level $\alpha = 0.05$, showing that the choice of features (PyEEG vs. CSP) is significant. To compare the difference between the uni-modal Eye based classifier and multi-modal classifiers (Eye and PyEEG), we performed three pairwise t-test with the adjusted α-value of 0.0167 after Bonferroni correction. For eye, we obtained p-values of $4.8e-08$, $4.0e-07$, and $7.7e-08$ for early, late, and hybrid fusion, respectively, indicating significant improvements. Comparing uni-modal PyEEG features with multi-modal approaches (Eye and PyEEG), we obtained p-values of 0.02, 0.26, and 0.22 for early, late, and hybrid fusion, respectively, showing no significant effect. We computed three paired t-tests for comparing different fusion methods with the adjusted α value of 0.167 after Bonferroni correction. The results showed the difference between early and late fusion (p-value = $1.0e-03$), early and hybrid

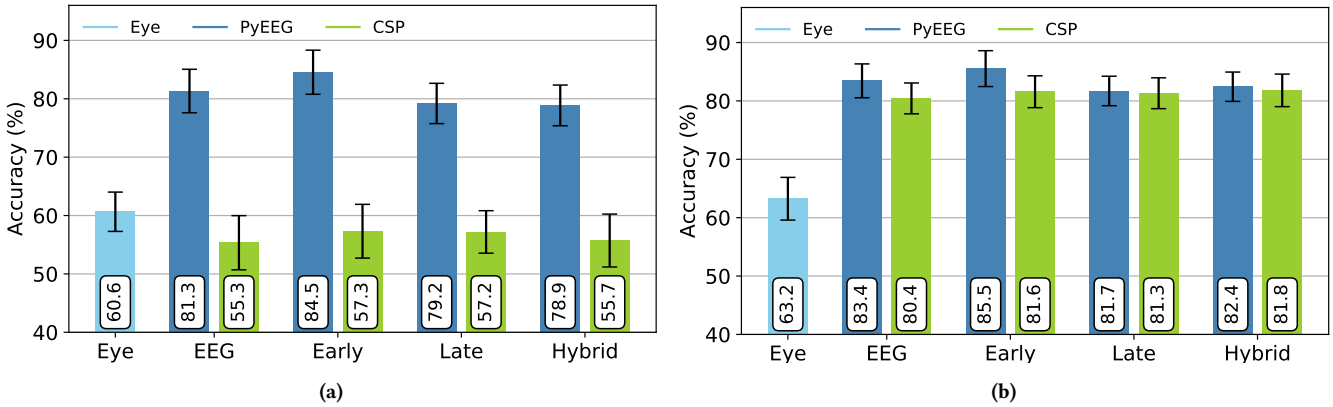

Figure 5: (a) Comparison of modalities and fusion strategies for cross-user conditions using the best overall classifier (SVM) (b) Comparison of modalities and fusion strategies for within-user conditions using the best overall classifier (SVM). Error bars indicate 95% confidence interval.

fusion (p-value = $6.6e-04$) is significant whereas late and hybrid fusion (p-value = $6.6e-1$) showed no significant effect.

5.2 Within-user Prediction

In Figure 5b, we present results for within-user prediction. In line with the cross-user prediction scenario, the best method for within-user prediction was an early fusion of eye-tracking and PyEEG features, reaching 85.5% accuracy. This is a substantial increase over eye-tracking features with 63.2% accuracy and a slight increase over PyEEG-only features with 83.4% accuracy. While the best method relies on joining eye-tracking- with PyEEG features, CSP features also perform well with an accuracy of 81.6% for early fusion with eye-tracking features. While there is an advantage for early fusion over late or hybrid fusion, the differences in fusion approaches are only moderate. Figure 5b showed the error bars with 95% confidence interval. We performed a paired t-test, and the calculated p-value of $2.0e-03$ is less than the significance level $\alpha = 0.05$. We have sufficient evidence that the choice of features (PyEEG vs. CSP) is significant. To compare the difference between uni-modality (Eye, EEG-PyEEG) and multi-modality fusions, we performed a pairwise t-test with the adjusted α-value of 0.0167 after Bonferroni correction. For eye, we obtained p-values of $1.8e-29$, $1.8e-24$, and $2.0e-27$ for early, late, and hybrid fusion, respectively, showing that the improvements are highly significant. For EEG-PyEEG, we obtained p-values of $5.0e-04$, $4.0e-03$, and 0.15 for early, late, and hybrid fusion, respectively, showing that the improvement for early and late fusion is highly significant. We computed three paired t-tests for comparing different fusion methods with the adjusted α value of 0.167 after Bonferroni correction. The results showed the difference between early and late fusion (p-value = $1.1e-06$), early and hybrid fusion (p-value = $6.3e-05$) is significant whereas late and hybrid fusion (p-value = $2.9e-1$) showed no significant effect.

5.3 Near Real-time Intent Prediction

We evaluate the influence of smaller search durations on user performance to understand better whether our approach can be used in near real-time intent prediction with the best configuration of EEG features and fusion methods. We select four different window sizes, from 0.5s to 2s, as post 2s duration, most of the users can locate the target, see Figure 3 (left). Monitoring the search performance and enabling proactive support to minimize search delays in real-time would be beneficial within these window sizes. To make a fair comparison across different window sizes, we take the same samples and therefore exclude samples where the search duration is less than 2s. Figure 6a showed the results for cross-user prediction, where 1.5s window achieves the best mean accuracy of 91.8%. Moreover, within-user follows a similar trend, as shown in Figure 6b, achieving the best mean accuracy of 90.1%. Interestingly, early fusion improves over mono-modal inputs in both prediction scenarios across all window sizes, with a much higher increase in within-user conditions. To compare the difference between uni-modality (Eye, EEG-PyEEG) and multi-modality fusions for the best-performing window size of 1.5s and for cross-user prediction, we performed a pairwise t-test with the α-value of 0.05. We obtained p-values of $1.0e-09$ and 0.18 for eye and EEG, respectively. Therefore, the improvement of early fusion over eye-tracking is highly significant. For similar analysis in the case of within-user prediction, we obtained p-values of $1.9e-31$ and $3.2e-10$ for eye and EEG, respectively. Therefore, the improvement of early fusion over eye-tracking and EEG is highly significant.

6 DISCUSSION

In the following section, we discuss the obtained results and focus on our method's performance and potential applications.

6.1 Achieved Performance

We presented the first method to the best of our knowledge for cross-user prediction of navigational vs. informational intent, reaching an accuracy of 84.5%. While this is comparable to the accuracy we reached for within-user prediction 85.5%, it is a highly promising result, allowing for much larger flexibility in application scenarios. In particular, when deploying an intent prediction system in an industrial context, collecting user-specific training data might not be feasible, highlighting the importance of effective cross-user

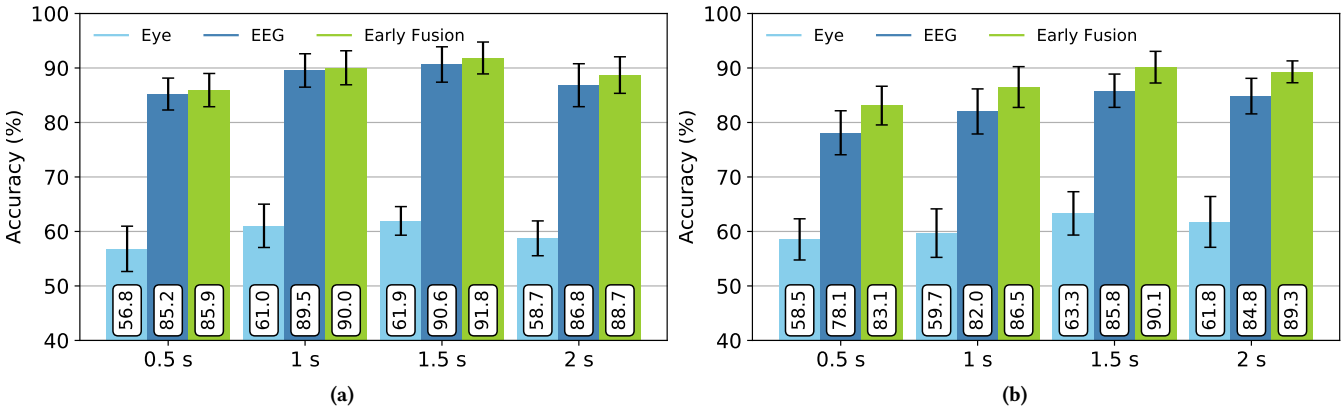

Figure 6: (a) Comparison with shorter time windows in cross-user condition with the best classifier (SVM), feature extraction (PyEEG), and fusion method (Early). (b) Comparison with shorter time windows in within-user condition with the best classifier (SVM), feature extraction (PyEEG), and fusion method (Early). Error bars indicate 95% confidence interval.

prediction. Using the CSP features extracted from EEG, we could not reach accuracies beyond 60%. With PyEEG features and feature selection, however, we reached a much higher accuracy of 84.5% when combined with eye-tracking features. There was no such substantial difference between EEG feature sets for within-user prediction. Thus, a key takeaway from our results is that cross-user prediction requires careful feature selection.

6.2 Applications

The ability to infer users' search intents by combining eye-tracking with the EEG data helps researchers study perceptual, attentional, or cognitive processes in more realistic situations as it aims to understand users' search intents without the need for them to communicate these intents verbally. By presenting the first method for cross-user prediction of navigational vs. informational intent, we pave the way for exciting new applications in other domains, such as VR-based gaming, where it can be used to provide a more immersive and interactive experience. For example, if the user searches for an object in a game, the system can adjust the game environment to make it more challenging based on the user's intent. The prime motivation of our work is applications in industrial work scenarios where workers are often faced with the task of finding a tool in a cluttered scene. Once informational intent is recognized for a user, a support system can help to find the desired tool. To know which tool the user is looking for, at least two approaches are conceivable. First, the support system could maintain a model of the user's current task and infer the tool the user most likely needs next to accomplish their task. Second, the prediction of informational intent could be followed by search target prediction [4, 32], informing the system about the tool the user is looking for. Apart from directly supporting humans in visual search, informational vs. navigational intent prediction could also be used to quantify how much time users spend in visual search. Repeated long search times may indicate that the work environment needs to be simplified or tidy for efficient work.

6.3 Limitations

While our novel dataset and cross-user prediction approach represent a significant step towards recognizing navigational and informational intent in the real world, some limitations remain. While using virtually created work scenes allowed us to include a large variety of visual environments in our data collection, a laboratory experiment always implies a domain gap to the real world. Especially the analysis of EEG signals is challenging in real-world environments due to motion artifacts and noise [9]. However, researchers are developing wearable dry EEG electrodes devices to improve the overall reliability and applicability of EEG signal analysis in real-world scenarios and advanced signal processing techniques like continuous contact impedance monitoring [5] and Gaussian Elimination Canonical Correlation Analysis [31]. Although we utilized classical classification and multi-modal fusion techniques to distinguish between navigational and informational intent recognition, which were widely used in the current state-of-the-art, given the size of the dataset, it would be interesting to observe the performance of deep learning architecture on this dataset.

7 CONCLUSION AND FUTURE WORK

We proposed the first publicly available EEG- and eye-tracking dataset for informational vs. navigational intent recognition and a multi-modal approach for classifying intents for within-user and leave-one-user-out scenarios. Our dataset improves over previous recording procedures using a concrete application scenario with fully user-defined search times. We thoroughly analyzed the dataset by evaluating it for within-user and cross-user scenarios and comparing different shorter time windows to demonstrate the potential for near real-time intent recognition. The presented statistical analysis highlights the importance of selecting appropriate features and fusion methods. Future work should consider different scenarios to predict informational and navigational intent, including hospitals, retail, and even people's private spaces. Even after recognizing navigational vs. informational intent, the identity of the object the user searches for is not apparent. For such scenarios, future work should investigate how to integrate navigational and informational

intent classification with search target prediction to help the user and reduce search times most effectively.

ACKNOWLEDGMENTS

This work is funded by the German Ministry for Education and Research (BMBF), grant number 01IW20003. Philipp Müller is funded by the German Ministry for Education and Research (BMBF), grant number 01IS20075.

REFERENCES

[1] Jayant Acharya, Abeer Hani, Janna Cheek, Parthasarathy Thirumala, and Tammy Tsuchida. 2016. American Clinical Neurophysiology Society Guideline 2: Guidelines for Standard Electrode Position Nomenclature. *The Neurodiagnostic Journal* 56 (10 2016), 245–252. https://doi.org/10.1080/21646821.2016.1245558

[2] Forrest Sheng Bao, Xin Liu, and Christina Zhang. 2011. PyEEG: An Open Source Python Module for EEG/MEG Feature Extraction. *Comp. Int. and Neurosc.* (2011), 406391:1–406391:7. https://doi.org/10.1155/2011/406391

[3] Michael Barz, Omair Shahzad Bhatti, and Daniel Sonntag. 2022. Implicit Estimation of Paragraph Relevance From Eye Movements. *Frontiers in Computer Science* 3 (2022). https://doi.org/10.3389/fcomp.2021.808507

[4] Michael Barz, Sven Stauden, and Daniel Sonntag. 2020. Visual Search Target Inference in Natural Interaction Settings with Machine Learning. In *ACM Symposium on Eye Tracking Research and Applications (ETRA '20 Full Papers)*. Article 1, 8 pages. https://doi.org/10.1145/3379155.3391314

[5] Alexander Bertrand, Vojkan Mihajlović, Bernard Grundlehner, Chris Van Hoof, and Marc Moonen. 2013. Motion artifact reduction in EEG recordings using multi-channel contact impedance measurements. In *2013 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. 258–261. https://doi.org/10.1109/BioCAS.2013.6679688

[6] Nilavra Bhattacharya, Somnath Rakshit, Jacek Gwizdka, and Paul Kogut. 2020. Relevance Prediction from Eye-Movements Using Semi-Interpretable Convolutional Neural Networks. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 223–233. https://doi.org/10.1145/3343413.3377960

[7] Soumyadip Chatterjee, Saugat Bhattacharyya, Amit Konar, D. N. Tibarewala, Anwesha Khasnobish, and Ramadoss Janarthanan. 2013. Performance Analysis of Multiclass Common Spatial Patterns in Brain-Computer Interface. In *Pattern Recognition and Machine Intelligence - 5th International Conference, PReMI 2013*. 115–120. https://doi.org/10.1007/978-3-642-45062-4_15

[8] Shiwei Cheng, Jialing Wang, Lekai Zhang, and Qianjing Wei. 2020. Motion Imagery-BCI Based on EEG and Eye Movement Data Fusion. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* PP (12 2020), 1–1. https://doi.org/10.1109/TNSRE.2020.3048422

[9] Yu Mike Chi, Yu-Te Wang, Yijun Wang, Christoph Maier, Tzyy-Ping Jung, and Gert Cauwenberghs. 2012. Dry and Noncontact EEG Sensors for Mobile Brain–Computer Interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 20, 2 (2012), 228–235. https://doi.org/10.1109/TNSRE.2011.2174652

[10] Yucel Cimtay, Erhan Ekmekcioglu, and Seyma Caglar-Ozhan. 2020. Cross-Subject Multimodal Emotion Recognition Based on Hybrid Fusion. *IEEE Access* 8 (2020), 168865–168878. https://doi.org/10.1109/ACCESS.2020.3023871

[11] Charles S. DaSalla, Hiroyuki Kambara, Makoto Sato, and Yasuharu Koike. 2009. Single-trial classification of vowel speech imagery using common spatial patterns. *Neural Networks* 22, 9 (2009), 1334–1339. https://doi.org/10.1016/j.neunet.2009.05.008

[12] Arnaud Delorme and Scott Makeig. 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods* 134, 1 (2004), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

[13] Thomas C Ferree, Phan Luu, Gerald S Russell, and Don M Tucker. 2001. Scalp electrode impedance, infection risk, and EEG data quality. *Clinical Neurophysiology* 112, 3 (2001), 536–544. https://doi.org/10.1016/S1388-2457(00)00533-2

[14] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti S. Hämäläinen. 2013. MEG and EEG Data Analysis with MNE-Python. *Frontiers in Neuroscience* 7, 267 (2013), 1–13. https://doi.org/10.3389/fnins.2013.00267

[15] John K Haas. 2014. A History of the Unity Game Engine. https://api.semanticscholar.org/CorpusID:86824974

[16] Chien-Ming Huang, Sean Andrist, Allison Sauppé, and Bilge Mutlu. 2015. Using gaze patterns to predict task intent in collaboration. *Frontiers in Psychology* 6 (2015). https://doi.org/10.3389/fpsyg.2015.01049

[17] Young-Min Jang, Rammohan Mallipeddi, and Minho Lee. 2014. Identification of human implicit visual search intention based on eye movement and pupillary analysis. *User Modeling and User-Adapted Interaction* 24, 4 (2014), 315–344. https://doi.org/doi/10.1007/s11257-013-9142-7

[18] Young-Min Jang, Rammohan Mallipeddi, Sangil Lee, Ho-Wan Kwak, and Minho Lee. 2014. Human intention recognition based on eyeball movement pattern and pupil size variation. *Neurocomputing* 128 (2014), 421–432. https://doi.org/10.1016/j.neucom.2013.08.008

[19] Jun-Su Kang, Ukeob Park, V. Gonuguntla, K.C. Veluvolu, and Minho Lee. 2015. Human implicit intent recognition based on the phase synchrony of EEG signals. *Pattern Recognition Letters* 66 (2015), 144–152. https://doi.org/10.1016/j.patrec.2015.06.013 Pattern Recognition in Human Computer Interaction.

[20] Marius Klug and Klaus Gramann. 2021. Identifying key factors for improving ICA-based decomposition of EEG data in mobile and stationary experiments. *European Journal of Neuroscience* 54, 12 (2021), 8406–8420. https://doi.org/10.1111/ejn.14992

[21] Zhen-zhong Lan, Lei Bao, Shoou-I Yu, Wei Liu, and Alexander G Hauptmann. 2014. Multimedia classification and event detection using double fusion. *Multimedia tools and applications* 71, 1 (2014), 333–347. https://doi.org/10.1007/s11042-013-1391-2

[22] Warren J Levy. 1987. Effect of epoch length on power spectrum analysis of the EEG. *Anesthesiology* 66, 4 (1987), 489–495. https://doi.org/10.1097/00000542-198704000-00007

[23] Yongqiang Liang, Wei Wang, Jue Qu, and Jie Yang. 2019. Application of Eye Tracking in Intelligent User Interface. *Journal of Physics: Conference Series* 1169, 1 (feb 2019), 012040. https://doi.org/10.1088/1742-6596/1169/1/012040

[24] Kip Ludwig, Rachel Miriani, Nicholas Langhals, Michael Joseph, David Anderson, and Daryl Kipke. 2009. Using a Common Average Reference to Improve Cortical Neuron Recordings From Microelectrode Arrays. *Journal of neurophysiology* 101 (03 2009), 1679–89. https://doi.org/10.1152/jn.90989.2008

[25] Päivi Majaranta and Andreas Bulling. 2014. *Eye Tracking and Eye-Based Human–Computer Interaction*. 39–65. https://doi.org/10.1007/978-1-4471-6392-3_3

[26] Philipp Müller, Ekta Sood, and Andreas Bulling. 2020. Anticipating Averted Gaze in Dyadic Interactions. In *Proceedings of ACM International Symposium on Eye Tracking Research and Applications (ETRA)*. 1–10. https://doi.org/10.1145/3379155.3391332

[27] Shahla Nemati, Reza Rohani, Mohammad Ehsan Basiri, Moloud Abdar, Neil Y. Yen, and Vladimir Makarenkov. 2019. A Hybrid Latent Space Data Fusion Method for Multimodal Emotion Recognition. *IEEE Access* 7 (2019), 172948–172964. https://doi.org/10.1109/ACCESS.2019.2955637

[28] Judith F Nottage and Jamie Horder. 2016. State-of-the-art analysis of high-frequency (gamma range) electroencephalography in humans. *Neuropsychobiology* 72, 3-4 (2016), 219–228. https://doi.org/10.1159/000382023

[29] Anneli Olsen. 2012. The Tobii IVT Fixation Filter Algorithm description. https://api.semanticscholar.org/CorpusID:52834703

[30] Ukeob Park, Rammohan Mallipeddi, and Minho Lee. 2014. Human implicit intent discrimination using EEG and eye movement. In *International Conference on Neural Information Processing*. Springer, 11–18. https://doi.org/10.1007/978-3-319-12637-1_2

[31] Vandana Roy, Shailja Shukla, Piyush Kumar Shukla, and Paresh Rawat. 2017. Gaussian elimination-based novel canonical correlation analysis method for EEG motion artifact removal. *Journal of Healthcare Engineering* 2017 (2017). https://doi.org/10.1155/2017/9674712

[32] Hosnieh Sattar, Sabine Müller, Mario Fritz, and Andreas Bulling. 2015. Prediction of search targets from fixations in open-world settings. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 981–990. https://doi.org/10.1109/CVPR.2015.7298700

[33] Mansi Sharma, Maurice Rekrut, Jan Alexandersson, and Antonio Krüger. 2023. Towards Improving EEG-Based Intent Recognition in Visual Search Tasks. In *Neural Information Processing: 29th International Conference, ICONIP, Proceedings, Part III*. Springer, 604–615. https://doi.org/10.1007/978-3-031-30111-7_51

[34] Gino Slanzi, Jorge A. Balazs, and Juan D. Velásquez. 2017. Combining eye tracking, pupil dilation and EEG analysis for predicting web users click intention. *Information Fusion* 35 (2017), 51–57. https://doi.org/10.1016/j.inffus.2016.09.003

[35] Sven Stauden, Michael Barz, and Daniel Sonntag. 2018. *Visual Search Target Inference Using Bag of Deep Visual Words: 41st German Conference on AI, 2018, Proceedings*. 297–304. https://doi.org/10.1007/978-3-030-00111-7_25

[36] Julian Steil, Philipp Müller, Yusuke Sugano, and Andreas Bulling. 2018. Forecasting User Attention During Everyday Mobile Interactions Using Device-Integrated and Wearable Sensors. In *Proceedings of ACM International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI)*. 1–13. https://doi.org/10.1145/3229434.3229439

[37] Florian Strohm, Ekta Sood, Sven Mayer, Philipp Müller, Mihai Bâce, and Andreas Bulling. 2021. Neural Photofit: Gaze-based Mental Image Reconstruction. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. 245–254. https://doi.org/10.1109/ICCV48922.2021.00031

[38] Julia Trabulsi, Kian Norouzi, Seidi Suurmets, Mike Storm, and Thomas Zoëga Ramsøy. 2021. Optimizing fixation filters for eye-tracking on small screens. *Frontiers in neuroscience* (2021), 1257. https://doi.org/10.3389/fnins.2021.578439

[39] Nigel G. Ward, Chelsey N. Jurado, Ricardo A. Garcia, and Florencia A. Ramos. 2016. On the Possibility of Predicting Gaze Aversion to Improve Video-Chat

Efficiency. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*. 267–270. https://doi.org/10.1145/2857491.2857497

[40] Thorsten Zander, Jonas Brönstrup, Romy Lorenz, and Laurens Krol. 2014. *Towards BCI-Based Implicit Control in Human–Computer Interaction*. 67–90. https://doi.org/10.1007/978-1-4471-6392-3_4

[41] Minrui Zhao, Hongni Gao, Wei Wang, and Jue Qu. 2020. Research on Human-Computer Interaction Intention Recognition Based on EEG and Eye Movement. *IEEE Access* 8 (2020), 145824–145832. https://doi.org/10.1109/ACCESS.2020.3011740