NOT READY FOR REVIEW Detection of target object recognition in simulated driving

R Aydarkhanov[1], M Uscumlic[2], L Gheorghe[2], R Chavarriaga[3], J d R Millan[4]

[1]EPFL, Switzerland
[2]EPFL, Switzerland
[3]EPFL, Switzerland
[4]TU Austin, USA

E-mail: `ruslan.aydarkhanov@epfl.ch`

Abstract. Decoding visual recognition during everyday life is challenging. Classical Brain Computer Interfaces based on Event Related Potentials show impressive results in controlled conditions by decoding P300 component. Loosening the experimental conditions by introducing dynamic visual input and natural free view allows to embed these systems in daily activity. We address limitations of the existing systems by investigating the neural and behavioral correlates of visual recognition during simulated driving. Our protocol resembles active driving at a comfortable speed in an empty city with the objects to recognize popping up at a close distance. We decode the recognition of target objects from Eye Fixation Related Potentials (EFRP) in a closed loop scenario. By integrating decoding probabilities of multiple EFRP the accuracy reached 0.37 on average in identifying one out of four object types. These results show that visual recognition can be decoded during active driving with natural dynamics except for sudden appearance of task-relevant objects.

Keywords: Brain-computer interfaces, Electroencephalography, Eye tracking, Driving

1. Introduction

Decoding of neural and behavioral correlates of visual recognition is complex in everyday life. One of the common everyday activities is driving which is a complex activity involving a range of neural processes from motor control to visual information processing, thus illustrating many BCI challenges in real world application.

EEG signature of visual recognition was reported under various conditions. In the controlled experiments where subjects had to recognize rare target stimuli in the sequence, the recognition process is reflected in EEG as a well-known P300 Event Related Potential (ERP). It was successfully used in various BCI applications such as P300-based speller because of high decoding performance. With the typing speed of 10 characters per minute [1], the home use of such spellers can improve the quality of life of patients with strong motor disabilities, such as ALS [2, 3]. For healthy users, however, this setup and typing performance does not bring any value.

Successful decoding of visual recognition in free viewing tasks would open space for BCI application for healthy users. In free viewing conditions people need to fixate or track moving objects to perceive it in all details. Fixations evoke a type of ERP called Eye Fixation Related Potentials (EFRP) where later components of EFRP resemble P300 component from oddball paradigm [].

Visual stimuli mostly used in these studies range from simple geometric shapes randomly positioned in static scenes and natural images to synthetic dynamic scenes []. Only a few attempts on EFRP decoding in videos or VR simulations have been reported []. However, the experimental conditions in those studies do not fully reflect the real world dynamics. In human action recognition from a cartoon animation the video playback was sped up to limit the time of the recognition process []. In a maze navigation experiment the subjects experienced fast autonomous driving with a simple motor task of button press when the car in front brakes suddenly [].

In our study subjects perform visual recognition task while primarily engaged in active driving in a car simulator. The subject are faced with natural radially expanding optic flow. We study EFRP while driving facing optic flow yet we ensured that recognition happens upon the fixation by introducing pop up effect. Otherwise, there would be 2 challenges: 1) case when driver gazes at the distant object until he approaches it close enough to recognize and 2) case when he returns gaze to the object multiple times until the object is recognizable. The latter challenges collecting clean data while the former challenges the EFRP decoding itself.

2. Materials and Methods

2.1. System

Our system is based on the driving simulator previously used for EEG-based BCI experiments [4, 5]. We extend the protocol published in [6]. It allows for immersive driving experience through the utilization of real Nissan driving chair with steering wheel and two pedals (gas and brake). The visual input is provided with three 3D monitors which create multiple renders for different angles. The virtual environment is implemented on a basis of the open source driving simulator project VDrift [7]. The environment resembles a regular grid city with static objects, i.e. building, traffic lights, fields. The task-related objects include direction indications on the road, target cue, boards with symbols and finish lines.

2.2. Tasks

The experimental session is completed in 2 phases offline and online with similar instructions for both phases. First of all, it is required to drive through the city while following the direction indications. Every road begins and ends with a left or right turn at the crossing. Within each road subjects perform the cognitive task while driving. In the beginning of the road the target symbol (the cue) is depicted on the ground. Subjects must look at the cue and remember it. While driving trough the road multiple boards appear one by one on both sides of the road. Only a fraction of the boards have the target symbol on them. Subjects must visually attend all the boards and count the number of boards with the target symbol. The terminal part of the road is marked with a finish line and reserved for count reporting in offline or feedback in online protocols.

Offline task. The steering wheel is equipped with a button. After crossing the finish line the subject presses the button as many times as the number of target boards he/she counted along the current road.

Online task. In the online scenario the system decodes the target based on the neural responses of the subject. After crossing the finish line the predicted target is projected at the bottom of the screen. Subjects are instructed to pay attention to this feedback.

One quarter of the roads are designed empty to allow subjects to rest.

2.3. Stimuli presentation

The board presentation is carefully adjusted to guide the behavior of subjects. First of all, boards are invisible unless the driver approaches them close enough making them to pop up suddenly. Their positions are generated using the following rules. The boards appear on a regular grid along the road however randomly on either side of the road with maximum of 2 boards on the same side in a row. The number of boards on left and right sides are balanced. Since the pop up distance was greater than the distance between the boards along the road, multiple boards from the same side were visible in the same time so their horizontal and vertical position were adjusted to avoid the overlap for the driver view.

The maximum speed of the car was limited to ensure that all the targets can be attended. The subjects were allowed to slow down if it is necessary to attend all the boards and count the targets. Nonetheless, all the subjects practiced until they felt comfortable with completing the recognition task at the maximum speed during the EEG setup. Due to constant speed and regular placement of the boards they popped up at a regular pace with 900 ms period.

In order to link the perception of the symbols on the board with the eye fixations, the recognition by peripheral vision must be avoided. Therefore, the target and distracting symbols were similar and surrounded by # character. Additionally, we added a bright red border around the board similar to the traffic signs to create a contrast with the environment and facilitate their identification.

Offline stimuli. In the offline phase one of the two symbols were depicted on each board: E and horizontally flipped E, i.e. ∃. One of them was randomly chosen as a target and were presented as the cue at the beginning of the road. There were 2-5 targets out of 12 boards on each road with the average fraction of targets of 0.25 in total.

Online stimuli. In the online phase 4 different symbols were available. There were 3 boards of each type resulting into 12 boards on the road. Only one of them was a target on each road.

2.4. Data collection

We had 13 volunteers (N male and N female) with the average age of N. They participated in one 3 hour session which included 1 hour of the set up, 45 minutes per phase and 30 minute break in between. The offline phase consisted of 3 runs through the city whereas the online phase could have from 3 to 5 runs depending on the available time. One run included 20 non-empty roads with 240 boards in total. Before each run the subjects were asked to move their eyes up-down and left-right for one minute in order to collect the data for eye movement artifact removal.

The EEG was acquired with BioSemi ActiveTwo system with 64 electrodes at 2 kHz sampling rate. Additionally, we recorded 3 EOG channels to collect the eye movement data: two electrodes next to the outer canthi of the eyes and one above the nasion. The EEG data were captured and saved on the laptop. The real time processing of EEG in online phase was done on the same laptop using CNBI loop.

The eye gaze was recorded with SMI RED Eye tracking system with the sampling rate of 120 Hz. The chair and eye tracker positions were adjusted for each subject. The eye tracker was calibrate with 13 points only once after the EEG setup and before beginning of the experiment.

The driving simulator logged various information of the driver location, the controllers state and the 2D position of boards on the screen at the sampling rate of 256 Hz. In order to synchronize the data acquisition on three separate machines (EEG, eye tracking and driving simulator) at different sampling rates, a square pulse of 4 Hz was generated by the driving simulator and sent to the eye tracker through TCP connection and to BioSemi through the parallel port.

2.5. Fixation extraction and analysis

There exist numerous methods to extract eye movement events from the eye gaze direction. Some of them proved to provide a better quality according to the human experts however are more challenging to implement in real time. We used different methods for offline and online. Simulated online analysis is implemented identical to online phase.

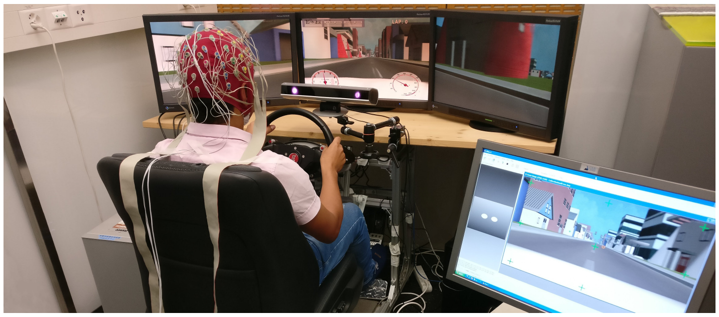

Figure 1. Experimental setup.

Offline fixation analysis. The detection of fixation is done with the Identification by 2-Means Clustering (I2MC) method. We relied on the implementation provided by the authors of the method using the default parameters. The main idea behind is to find the transition between two consecutive fixations by applying 2-mean clustering in a sliding window manner. During fixation the eyes do not move so if we can clearly detect 2 clusters it means that they correspond to two fixations. This method is more precise and robust to noisy outliers which allows to obtain a training dataset of higher quality.

Online and simulated online fixation analysis. We could not use the provided implementation of I2MC in real time to extract fixations so we used the Identification by Dispersion-Threshold (IDT) supplied with our eye tracking system. Fixation in IDT is extracted when the signals lies within the dispersion thresholds for at least a minimum fixation duration. It requires two parameters: we used 100 ms for the minimum fixation duration and 200 pixels for the maximum dispersion.

The cognitive response is stronger when the stimulus is perceived and recognized for the first time. We assume that subjects categorized the symbol at the first attendance so we use only the first fixations on the boards for our analysis.

The visual input during the task is dynamic. Due to driving through the virtual environment the objects including the boards are also moving on the screen. So we assume that most of the board attendances are done with smooth pursuit rather than fixations. To the best of our knowledge there is no available algorithm for efficient extraction of smooth pursuit for eye movement data sampled at 120 Hz. The only consequence of extracting fixation from smooth pursuit is that a single smooth pursuit may be oversegmented into multiple fixations. For the sake of our analysis we do not need to differentiate between fixations and smooth pursuit movement. The onset of first fixation on a board will coincide with the onset of smooth pursuit.

For the behavioral analysis we estimate the total attendance time of boards for the first uninterrupted visit or dwell time. The dwells were created by merging all the fixations on the same board with saccade durations between them below 50 ms.

Each fixation and dwell were assigned to a target board, a non-target board or non-board. Due to a reading visual span of several degrees, the board movement and noisy eye tracking data we applied the following approach to assign the boards to fixations. For each eye gaze sample we estimate the probability of fixating eyes on the center of the board according to a normal distribution. After averaging log-probabilities across the dwell time we apply a hard threshold to assign the fixation to a board or a non-board class.

2.6. EEG data processing

All the EEG and EOG were filtered with Butterworth band-pass filter of order 4 within the band [1, 10] Hz forward and backward and downsampled from 2 kHz to 256 Hz. Due to low conductivity of the skull and the skin, EEG signal is spatially smoothed so a high contrast between nearby channels is a result of noise and movement artifacts. We remove this noise by keeping only low spatial frequency components after decomposition EEG with SPHARA. Horizontal and vertical components of eye movement were estimated which allowed to remove the eye movement artifacts from EEG using multivariate regression. The coefficients of multiple regression were estimated from the one-minute session of eye movements before the corresponding run. Then the signal is spatially filtered with common-average-reference (CAR). The epochs are extracted from time window of [200, 1000] ms after the fixation onset. We investigate and compare different sets of features, which include EFRP waveform and covariance-based features.

Offline EEG processing. For the offline analysis we chose the following combination of features and classifiers:

- **Linear**. Penalized logistic regression (PLR) trained on waveform features after reducing the dimensionality with PCA. Only the components which explain 90% of variance are kept.
- **Dwell**. PLR trained on dwell time on the boards.
- **Linear with Dwell**. PLR trained on the combination of waveform features with dwell time. We concatenate the two feature sets before applying PCA to keep 95% of variance. Since the dynamic range of dwell time in ms is greater than the one of EEG in uV, most of the information remains is projecte
- **Random Forest**. Random forest trained on waveform features. We use 100 decision trees and with maximum depth of 5.
- **Riemann**. PLR trained on Riemannian features from simple epochs. To build Riemannian features we estimate spatial covariance matrix with shrinkage and project it to the tangent space according to the classical Riemannian geometry on SPD matrices. We subselected 8 channels based on mean Fisher score across the epoch.
- **Riemann+**. PLR trained on Riemannian features from augmented epochs. Before computing the covariance matrix we augment the epoch with the averaged ERP

for each class (target and non-target). Otherwise, it is identical to the previous approach.

Since PLR is a linear regularized classifier we standardize all the features to z-score when using PLR.

Online EEG processing. The online phase required the real time processing. SMI system provides a real time eye fixation detection. The fixations were buffered by a parallel process within the driving simulator, matched with the boards, and a trigger was sent to the BioSemi system 3 s after the onset of each the fixation on a board. We choose 3 s delay because we apply non-causal filter on EEG data. EEG processing was identical to the offline procedure except for 2 steps:

- the spectral filtering was done on a 5 s buffer of data, approximately around [-2, 3] s around the fixation onset;
- the eye movement artifacts were removed based on the multiple regression coefficients trained with offline phase eye movement data.

On the data obtained in the offline phase we trained Random Forest classifier and applied it in real time. The probability for the target class was sent back to the driving simulator. After crossing the finish line the probabilities were averaged per each symbol (1 out of 4). The symbol which had the highest probability of being a target was shown to the subject on the screen.

Simulated online EEG processing. The EEG processing was identical to the offline analysis except for the eye movement artifacts removal. The multiple regression model was obtained from offline data.

2.7. Performance estimation

Performance is estimated differently for the data from offline and online phases.

Offline performance evaluation. We employ nested cross validation to adjust various hyperparameters in the inner loop: regularization term for PLR and the tree depth in Random Forest. The purpose of the outer loop is to obtain an unbiased performance estimation so it is critical to avoid training and testing on correlated data. We achieve it by performing leave-one-run-out for the outer loop, although we had only 3 offline runs. The inner loop is implemented with 4-fold cross validation while keeping the temporal order of the trials before the split. Since the classes of target and non-target eye fixations are unbalanced, we utilized AUC to measure the classification performance.

Simulated online performance evaluation. After training the classifiers on the offline data we applied them to the online data and assessed AUC.

Table 1. Board attendance rate

	Offline	Online
Targets	0.87	0.45
Non-Targets	0.87	0.43

Online performance evaluation. During online phase we predicted the target symbol from the EFRP classification. We assess the overall performance with accuracy and confusion matrices for 4 symbols.

3. Results

3.1. Behavioral analysis

Board attendance Subject attended most of the boards in the offline phase and approximately half of them in the online phase. The average attendance rate is shown in the Table 1. Repeated measures ANOVA shows significant difference between all 4 groups: targets in offline, targets in online, non-targets in offline and non-targets in online, with $p - value < 0.0001$. Post-hoc analysis shows that it is driven by the difference between offline and online phases with p-value < 0.0001 (paired t-test). The difference between targets and non-targets is not significant with p-value $= 0.03$ so we can assume that subjects could not differentiate the symbols with peripheral vision.

Counting The total number of targets in offline phase is 173. We analyzed the button presses which should be equal to the number of targets on each road. The average number of incorrect counts (both missed and extra counts) was 5, the worst performance was at 15 errors (Figure 5).

Dwell time We analyzed the distribution of dwell times on targets vs non-targets in offline and online phases (Figure 2). Most of the dwells are limited to the time between the boards pop up equal to 900 ms. The dwell times are identical for non-targets in both phases and significantly shorter than for targets (p-value < 0.0001). The median dwell time for targets is significantly longer in online phase (p-value < 0.0001).

3.2. EFRP waveform

We present the analysis of the EFRP waveform for the 4 subjects who demonstrated the highest classification performance in the offline phase with the Linear classifier. The univariate discriminant power is shown on the Figure 3. The results are similar for the offline and online phases with online data having twice as higher discriminant power with up to 0.01 of signed R^2. The greatest values are mainly confined within the region between 100 and 700 ms. The higher discriminant power is spread across the whole scalp which can be the consequence of using CAR in the processing. Nonetheless, the

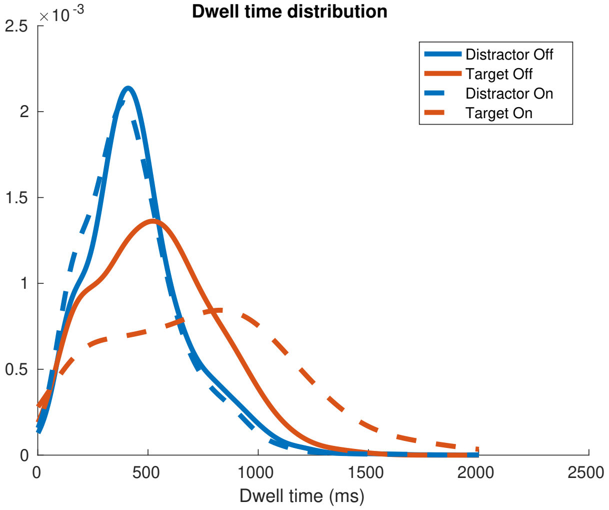

Figure 2. Dwell time distribution for targets vs distractors in offline and online phases.

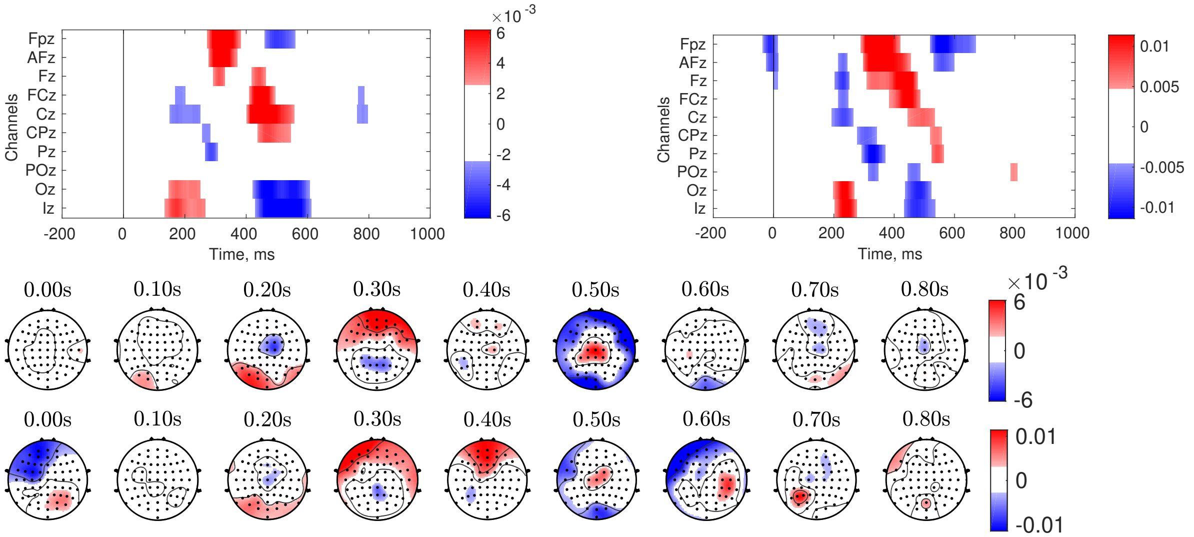

Figure 3. The discriminant power for the 4 best subjects with the offline AUC above 60. Signed R^2 is demonstrated for midline channels across around the eye fixation onset (top left: offline, top right: online) and on topographic maps (top map: offline, bottom map: online).

P300-like component can be seen at 500 ms after the fixation onset. The representative channel for this component is Cz.

We visualized Cz signal for each eye fixation while ordering them by the dwell time (Figure 4). The amplitude of the presented EFRP is limited to the range [-1.5, 1.5] uV. The complex of components right after the fixation onset ranging from 100 to 300 ms reflects the evoked potentials from the fixation itself. It contains negative and positive deflections. We can observe the same complex of components after the dwell offset. The shift of gaze happens right where we expect the P300-like component so it can be masked by this evoked activity. The positive deflection occurs at the end of the dwell

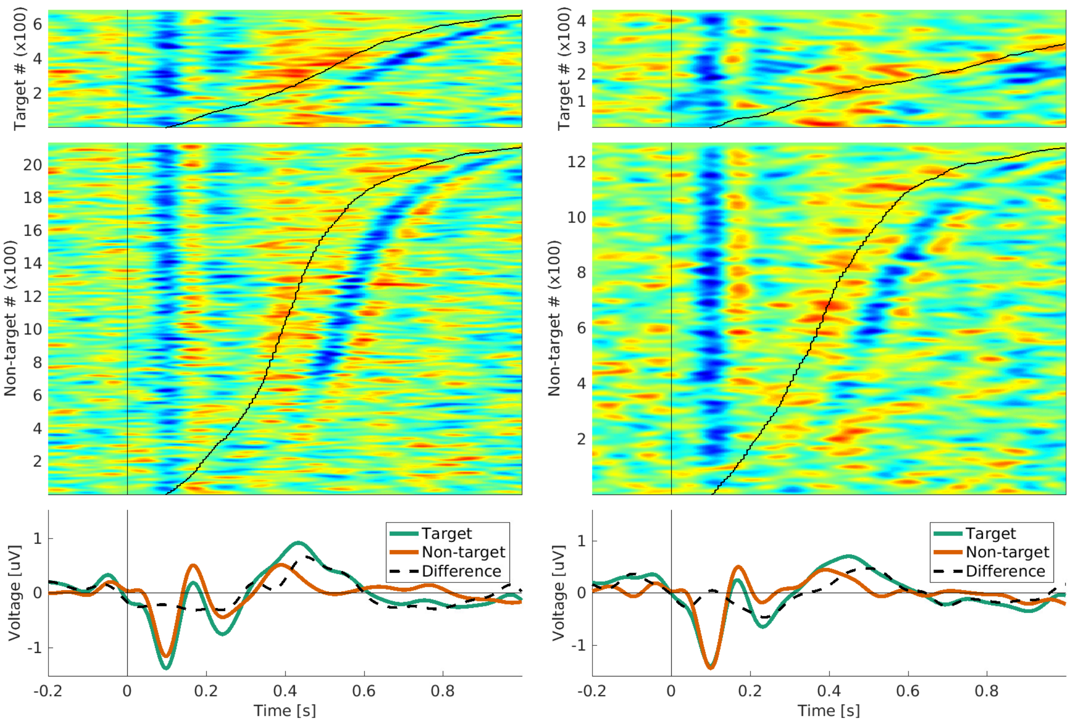

Figure 4. The signal of Cz channel for all epochs for the 4 best subjects. Left: offline, Right: online. The black curve shows the offset of the fixation.

for both targets and non-targets, however it has a greater amplitude for targets as seen on the averaged EFRP.

3.3. Comparison of decoding approaches

Offline All the classification approaches lead to the performance between 53 and 60 AUC points on average (Figure 5) which is statistically significant against random level of 50 (p-values < 0.001 with Student's t-test after Bonferroni correction). 8 out of 13 subjects achieve performance above 60 for at least one of the approaches based on neural correlates. Although each subject has a different preferred approach, the differences between the approaches are not statistically significant (p-value $= 0.16$ with repeated measures ANOVA). It is worth noticing that the combination of both dwell time and EEG waveform is not always better than just one of these feature sets.

Simulated online First of all, we note that the performance of neural-based approaches on online data is consistent with the training performance on offline data. The average AUC values lie between 56 and 59 for each approach which is statistically significant against 50 (p-values < 0.001 with Student's t-test after Bonferroni correction).

For approaches relying on the dwell time, however, the performance drastically improved for multiple subjects. The average AUC for *dwell* classifier increased from 56

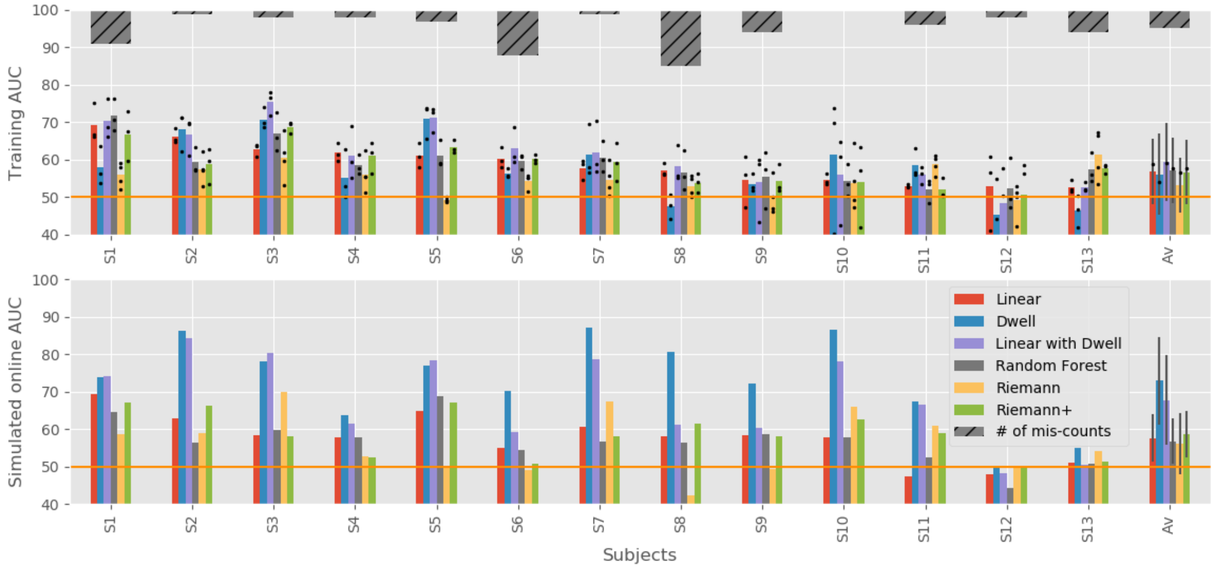

Figure 5. Performance of EFRP classification with various approaches in offline analysis (top) and simulated online analysis (bottom). Each dot shows single fold performance in a leave-one-run-out cross validation for the corresponding classification approach.

Table 2. Online performance

Subjects	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
Accuracy	0.44	0.38	0.45	0.29	0.44	0.35	0.39	0.25	0.31	0.38	0.38	0.4
Accuracy test	**0.003**	0.29	**0.002**	5.27	**0.003**	0.62	0.08	12.0	2.40	0.29	0.29	0.16
Independence test	**0.004**	2.58	0.06	8.91	0.24	0.84	0.39	4.13	5.26	1.14	1.53	2.07

to 73 and for the *Linear with Dwell* classifier from 59 to 67. This improvement is a direct consequence of the changes in target dwell time together with the constant non-target dwell time shown in 2.

3.4. Online performance

We assessed the task performance of each subject as the accuracy of target decoding (Table 2). The averaged accuracy equals 0.37 and it is significantly different from random level of 0.25 for a classification of 4 between balanced classes (p-value < 0.0001 with Student's t-test). Additionally, we applied statistical test to assess the accuracy per subject. After Bonferroni correction only 3 subjects performed statistically better than random performance.

To verify the independence of the 4 classes in online phase we computed the aggregated confusion matrices of across all symbol (Figure 6). We applied an independence test for confusion matrices per subject. After Bonferroni correction only

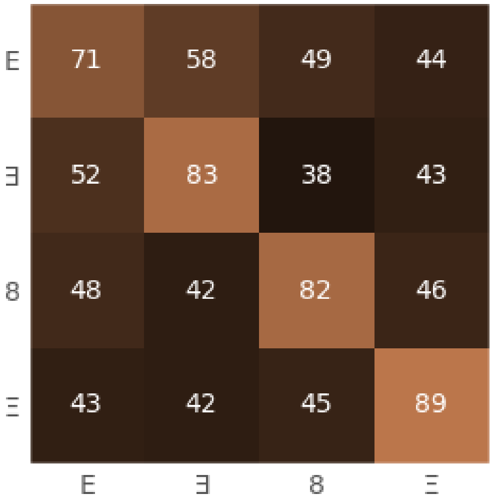

Figure 6. The aggregated confusion matrix of online decoding for each type of symbol being a target.

one subject shows significant imbalance in the confusion matrix. It is linked to a high accuracy of the symbol Ξ and low accuracy for the symbol E.

4. Discussion

Decoding of cognitive process from EFRP has a potential to augment real life interaction with machines for healthy users. However, there are still challenges to solve to achieve a satisfactory performance.

In this study we investigated the scenario of simulated driving. We limit the driving task to following the simple route at a comfortable and natural speed without other participants of the traffic or other moving objects. However, due to the movement of the car, the drivers were subjected to a dynamic visual input. Although subjects followed the objects moving on the screen, we could analyze their eye behavior by approximating smooth pursuit with consecutive fixations combination.

In the online phase subjects looked longer on targets compared to the offline phase. Although the subjects dwell time was not decoded directly, they were aware that they could potentially influence the decoding quality. This might lead to deliberate or unconscious changes in their behavior. Some subjects could achieve a high decoding performance based only on dwell time in offline phase. But with the changes in the behavior during online phase most of the subjects drastically improve in their decoding performance.

We observed a statistically significant difference in the board attendance between offline and online phases. On one hand, we expected the subjects to be more engaged in the task due to interactive feedback part. Observed longer dwells on targets make it more challenging to attend all the boards. On the other hand, behavioral reports were not required which released the pressure to complete the task properly.

Comparing targets vs non-targets we found no difference in the attendance rate which confirms that subjects could recognize the symbol only by directly looking at it. The behavioral reports show that subjects properly counted the number of targets on the road with a limited number of errors. So most of the board attendances led to a proper recognition in the offline phase. Moreover, the balanced confusion matrices in the online phase confirm that subjects perceived all the symbols equally with the regards to the task. The presence of new types of stimuli did not alter the cognitive process.

The discriminant analysis of EFRP shows similar results for both offline and online phases. Most of the relevant features lie within [200, 700] ms window which coincides with the dwell durations. The spatial localization of relevant features is consistent with the typical spatial distribution of P300 component in oddball paradigm. The EFRP waveforms are known to contain a strong P1 component at the occipital area that reflects the beginning of the visual processing of a stable visual input after the saccade. In the analysis of Cz channel it corresponds to the negative deflection at 100 ms. It is clearly present in most of the fixations on boards. Moreover, we could also see it for the following fixations. It leads to the overlap between P300-like component and the evoked fixation-related components. This overlap contaminates the data and complicates the decoding of cognitive process. Removing the evoked activity from the EEG can be done by modeling it from various characteristics of the previous fixations and saccades []. However, there is a risk to remove the cognitive signal even with EFRP collected from more controlled conditions with static visual input.

We limited the removal of artifacts to high frequency spatial noise with SPHARA and direct eye movement potential propagation by regressing it out from the EOG signal.

We compared multiple classification approaches based on EFRP on offline data and simulated their application to online data. All approaches including waveform-based linear and non-linear and covariance-based methods result in similar performance on average across subjects which is significantly above the chance level. However there is no single best approach for all subjects.

One can argue that the high performance of EFRP-based classifiers is due to the strong and well-aligned evoked potentials after the fixations, which reflects the difference between the target and non-targets dwell times. In this case we would see an improvement in performance for online data similar to the classification approaches based on dwell time. The combination of two sources of information (EEG and dwell time), nonetheless, does not lead to significant improvement and in the simulated online decoding it results in intermediate performance.

The actual online performance measured by the target symbol identification is significantly above the chance level only for 3 subjects. Nonetheless, on average across subjects the accuracy of 0.37 is significantly higher than 0.25.

References

[1] Aya Rezeika, Mihaly Benda, Piotr Stawicki, Felix Gembler, Abdul Saboor, and Ivan Volosyak. BrainComputer Interface Spellers: A Review. *Brain Sciences*, 8(4), March 2018.

[2] Eric W. Sellers, Theresa M. Vaughan, and Jonathan R. Wolpaw. A brain-computer interface for long-term independent home use. *Amyotrophic Lateral Sclerosis*, 11(5):449–455, October 2010.

[3] Elisa Mira Holz, Loic Botrel, Tobias Kaufmann, and Andrea Kbler. Long-Term Independent Brain-Computer Interface Home Use Improves Quality of Life of a Patient in the Locked-In State: A Case Study. *Archives of Physical Medicine and Rehabilitation*, 96(3, Supplement):S16–S26, March 2015.

[4] Zahra Khaliliardali, Ricardo Chavarriaga, Lucian Andrei Gheorghe, and Jos del R. Milln. Action prediction based on anticipatory brain potentials during simulated driving. *Journal of Neural Engineering*, 12(6):066006, 2015.

[5] H. Zhang, R. Chavarriaga, Z. Khaliliardali, L. Gheorghe, I. Iturrate, and J. d R. Milln. EEG-based decoding of error-related brain activity in a real-world driving task. *Journal of Neural Engineering*, 12(6):066028, 2015.

[6] Hadrian Renold, Ricardo Chavarriaga, Lucian Andrei Gheorghe, and Jos del R. Milln. EEG correlates of active visual search during simulated driving: An exploratory study. In *2014 IEEE International Conference on Systems, Man, and Cybernetics*, 2014.

[7] About - VDrift.