Assisting Telemanipulation Operators via Real-Time Brain Reading

Elsa Andrea Kirchner*,†, Jan Hendrik Metzen†, Timo Duchrow†, Su Kyoung Kim*,†, Frank Kirchner*,†
*Robotics Lab
University of Bremen
Robert-Hooke-Straße 5, 28359 Bremen, Germany
† Robotics Innovation Center (RIC)
German Research Center for Artificial Intelligence (DFKI GmbH)
Robert-Hooke-Straße 5, 28359 Bremen, Germany
Email: {elsa.kirchner, jan_hendrik.metzen, timo.duchrow, su_kyoung.kim, frank.kirchner}@dfki.de

***Abstract*—In this paper, we present a concept for a new kind of man-machine interface that is based on the monitoring of brain activity and aimed at supporting operators in telemanipulation scenarios. Monitoring takes place unnoticed by the subject and is called brain reading. A brain reading interface (BRI) is a highly integrated control environment that observes the brain signals in real time. Consciously recognized and classified stimuli evoke a certain response in the operator's brain activity that will be detected by the BRI. Based on the detection of these changes in brain responses in the electroencephalogram (EEG), a brain reading system is able to discern whether a piece of information that has been presented to the operator was acknowledged or not. Hence, the BRI ensures that environmental alerts are processed and classified by the operator. Thus, BRIs can be a crucial component of control systems ensuring that operators perceive and cognitively process alerts presented to them during highly demanding tasks, like complex manipulations. We show that brain activity changes that correlate with the classification of important, task-relevant stimuli in multi-task telemanipulation-like scenarios are stable. Furthermore, we will outline a concept for a BR system that allows the detection of these brain activity changes in single trial EEG epochs based on machine-learning methods.**

I. Introduction

In many situations it is highly desirable for a machine to have information about the current (mental) state of its user in order to choose proper actions. In highly demanding situations, like space exploration, it is common to monitor subjects by recording and analysing their body signal data, such as ECG (electrocardiogram), pulse or GSR (galvanic skin response) that can be evaluated to measure stress level and exhaustion [5]. Furthermore, analysis of brain data can be applied to obtain insight into mental states. Electroencephalography (EEG) is a favored method to observe brain activity since it combines good time resolution and a sufficient spatial resolution without the need to implant artifacts, like electrode arrays, invasively under the skull. Several event-related potentials (ERPs) as well as changes in brain wave frequency bands and activity patterns are known to be coupled with mental or cognitive states or state changes. A well-investigated ERP is the so-called P300 [31]. The P300 (details in section II-A) is a positive fluctuation in the EEG, evoked by infrequent, important (task-relevant) stimuli that are attended, recognized, and cognitively evaluated by the subject. Thus, P300 can be used as a marker for successful information processing.

In many telemanipulation scenarios[1] it is of interest to know whether the operator perceived and understood important information (e.g. warnings or certain task-relevant messages). At the same time, the operator has to work highly concentrated without becoming distracted by repeated presentations of the same warning that he deliberately ignores. Our approach is to monitor the operator's EEG in order to detect changes or evoked activity like the P300 potential that indicate the processing and classification of an important, task-relevant stimulus (like a warning). This passive monitoring is called *brain reading (BR)*. BR denotes the external observation of brain activity (e.g. by means of EEG) without the active participation of the subject. Thus, BR can take place fully unnoticed by the user.

Unlike BR, classical Brain-Computer Interfaces (BCI) are not suited for the purpose of monitoring since those interfaces are used to control a machine, computer or prostheses [35] via the brain and need the user's attention. Even though this control can be learned by the subject and by thus turned into a highly automated behavior, it will still use cognitive resources of the operator and because of this does not improve a situation where an operator's brain is already under a high level of workload. In contrast, BR can be the method of choice to monitor the operator's brain signals in realtime to ensure that environmental alerts have been consciously processed by the operator. Since the operator will not be aware of this monitoring, he will be able to concentrate on the task, e.g. telemanipulating a complex robot. A further difference to typical BCI systems is that due to the real-time constraints, processing of brain patterns has to be done based on the individual EEG epochs (so-called "single-trial" analysis) instead of an average of several EEG epochs that have been obtained under similar conditions like in most BCIs (e.g. P300-based spellers [30]). Average analysis is easier

[1] e.g. the remote control of a robotic arm of an underwater vehicle by a human operator situated in a control center of a submarine or marine ship.

because averaging increases the signal-to-noise[2] ratio since the noise in the individual EEG epochs is not correlated and largely cancelled out by averaging. In contrast, single-trial analysis must deal with low signal-to-noise ratios since the relevant information is typically significantly weaker than background activity and noise. One increasingly popular approach for single-trial analysis of EEG data is the adoption of (supervised) machine-learning techniques [22]. This is due to the fact that after a short calibration session in which "typical" EEG epochs are recorded from a subject under the respective conditions, the machine is able to adapt to individual brain patterns of the user (in contrast, many classical BCIs require the user to adapt their brain waves so that they are understandable by the machine).

In this paper, we outline a concept for applying single-trial analysis of EEG data in a brain reading scenario, namely for monitoring and supporting operators, and thus not for a direct BCI control (compare [22] for examples of both approaches). We believe that brain reading can be especially useful in scenarios where operators telemanipulate complex robotic systems. To investigate EEG potentials in a scenario that exhibits many of the characteristics of brain reading in a real-world telemanipulation scenario, we set up a test bed (called "Brio oddball scenario") that requires elevated levels of concentration, fine motor control as well as response to presented information. We show that certain brain potentials are elicited after the cognitive processing of important information (see Section II), showing the principle feasibility of single-trial analysis in such a scenario. Thereupon, we outline a concept and a software framework for real-time single-trial brain reading (see Section III).

II. Paradigm and EEG Offline Analysis

A. P300 Under Cognitive Load

In our experimental setup, the P300 potential and accompanying changes in brain activity (e.g. changes in frequency) will be investigated regarding their usability in a BRI. This paper will focus on the P3b potential [31], [34] (further called P300) which is evoked by task-relevant stimuli to answer the question whether task-relevant information was processed and classified in a manipulation scenario. P300 is a well-known and thoroughly studied potential. On the one hand this potential is stable and strong, allowing its use in classic BCI [6], [2] applications, on the other hand, peak latency of the P300 will shift regarding the complexity of the cognitive task to evaluate a stimulus task relevance [18] and the amplitude is sensitive regarding the subjective rarity, importance and unambiguousness of the stimulus [16]. Beside this, the magnitude of the P300 amplitude also depends on whether subjects devote high amounts of effort to the task [14].

Subsequently, we present an offline analysis of EEG data recorded under two different experimental conditions, where subjects have to solve one task in one scenario (simple oddball) or two different tasks in the other scenario (Brio oddball). We focus on P300 stability (changes in peak amplitude) and latency shifts due to the different complexity of both experimental setups.

[2]We refer to the relevant potential as the "signal" and to all other brain activities as "noise".

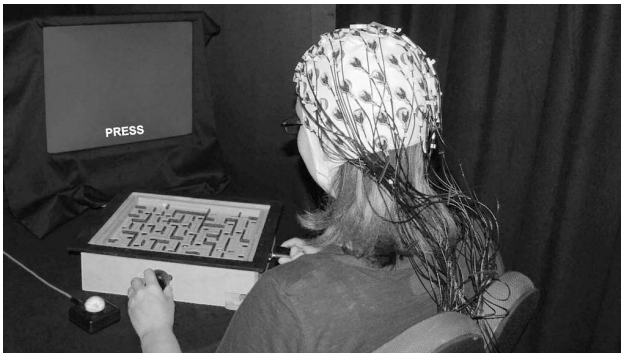


Fig. 1. **Experimental setup**: Subject is playing BRIO® and is reacting towards rare target stimuli (alerts) by pressing a buzzer.

B. Manipulation-Like Scenario - Brio Oddball

To analyse P300 in a manipulation-like scenario we set up a test bed, the "Brio oddball scenario", to be able to record data in a rather controlled environment. The test bed allows to investigate how an operator's EEG changes in regard to visually presented warnings while performing a manipulation task that requires elevated levels of concentration and fine motor control. The manipulation task of the subject is to play a BRIO® labyrinth game with the goal to manoeuvre a ceramic ball from a starting point along a partly bordered, marked path to the target position by tipping the board so that the ball rolls without falling into any of the holes. Detailed information about the test bed can be found in [21].

In Brio oddball, subjects were asked to play the labyrinth game as well as possible (contest situation) and at the same time they had to react to stimuli presented on a monitor (see figure 1). In simple oddball, subjects only reacted to the same stimuli without playing the game. Stimuli were presented with an interstimulus interval (ISI) between 600 and 800 ms. Stimuli consisted of a high number of irrelevant frequent information (**standards**, n = 720) and were mixed up with infrequent events that appeared one or several trials before a target to warn the subject of incoming target stimuli (**deviants**, n = 60), and rare target stimuli (**targets**, n = 60) that required a response (pressing a buzzer). The presentation of rare stimuli within a sequence of frequent stimuli is called oddball discrimination paradigm [28], [26], [25].

To investigate whether high cognitive load in Brio oddball influences the latency, amplitude, and stability of the P300 potential, we compared Brio oddball data with data recorded in a standard oddball paradigm (simple oddball), where subjects only reacted toward stimuli in the same experimental setup without playing the game and only focused on the monitor.

C. EEG Offline Average Analysis

1) Method of offline data analysis:

a) subjects: Eight undergraduate and graduate students (two female and six male; age from 19 to 29 with mean age of 24.38 ($\pm$4,033) participated. All subjects were right-handed and had normal or corrected-to-normal vision. Declaration of consent in writing was obtained from each participant.

b) task: All subjects (except for one subject) performed two experiments (simple and Brio oddball) at the same day. One subject attended on two different days. All subjects entered the simple oddball experiment first. Experimental setup is explained in II-B.

c) data acquisition: EEGs were recorded continuously from 64 electrodes (extended 10-20 system with reference at FCz), using an actiCap system (Brain Products GmbH, Munich, Germany). EEG signals were amplified by two 32 channel BrainAmp DC amplifiers (Brain Products GmbH, Munich, Germany) and filtered with a low cutoff of 0,1 Hz and high cutoff of 1000 Hz. EEGs were sampled at 2500 Hz. The impedance was kept below 5 kΩ. EEGs were analyzed off-line with BrainVisionAnalyser Software Version 2.0 (Brain Products GmbH, Munich, Germany). EEGs were off-line re-referenced to an average reference and filtered (0.2 Hz low cutoff, 45 Hz high cutoff). Artifacts (e.g. eye movement, blinks, muscle artifacts, etc.) were rejected semi-manually (amplitude 100/-100 μV, gradient 75 μV). EEGs were off-line segmented in epochs from 100 ms before stimulus onset to 1000 ms after stimulus onset. Epochs were averaged based on trial events. Baseline correction was performed before averaging (pre-stimulus interval: -100 to 0 ms).

2) Results of offline analysis:

a) Behavioral data: For statistical analysis, one-way ANOVA (SPSS, version 16, Inc., Chicago, IL, USA) was applied to response time (RT), with one factor: scenario type (simple oddball, Brio oddball). We found an effect of scenario type [$F(1,14) = 39.49, p < 0.001$], reflecting a different RT between simple oddball and Brio oddball. Subjects responded to targets faster in the simple oddball scenario [median of RT: 496 ms] compared to the Brio oddball scenario [median of RT: 720 ms]. Concerning response accuracy, we also performed one-way ANOVA with one factor: scenario type (simple oddball, Brio oddball). We found an effect of scenario type [$F(1,14) = 6.552, p < 0.024$], reflecting a different response accuracy between simple oddball and Brio oddball. Subjects responded more accurately in the simple oddball scenario [median of response accuracy: 100%] compared to the Brio oddball scenario [median of response accuracy: 91.66%].

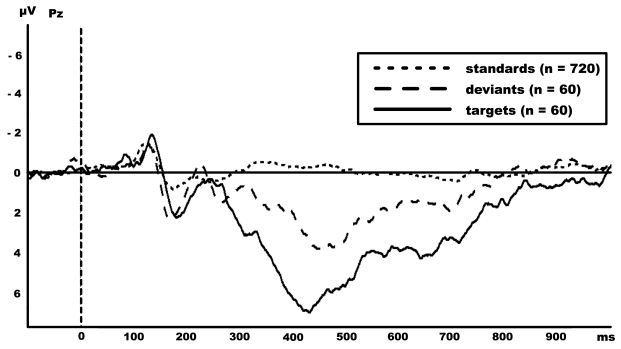
b) EEG/ERP data: For statistical analysis, two separate time windows were applied to the ERP data for amplitude and latency: 350-600 ms and 600-850 ms. We performed repeated measures ANOVA with two within-subjects factors: a) stimulus type (three levels: standard, deviant, target) and b) time windows (two levels: early time window, late time window). Besides, the scenario type (two levels: simple oddball, Brio oddball) was computed as a between-subject factor. Greenhouse-Geisser correction was applied and the corrected p-value was reported. For pairwise comparisons, Bonferroni correction was applied.

For P300 amplitude, we found a main effect of stimulus type [$F(2,28) = 33.02, p < 0.001$], reflecting a significant amplitude difference between standards and targets (i.e. P300 effect in the target condition) as well as standards and deviants (i.e. P300 effect in the deviant condition). There was a main effect of time window [$F(1,14) = 20.04, p < 0.002$], reflecting a significant amplitude difference between the early and the late time window. The time window interacted with the scenario type [$F(1,14) = 9.94, p < 0.008$]. Pairwise comparisons revealed a significant amplitude difference between the early and the late time window in simple oddball [$p < 0.001$], but not in Brio oddball [$p = n.s.$]. We found that a stronger P300 effect in the early time window compared to the late time window was observed only in simple oddball. In contrast, we found a broader P300 peak in Brio oddball. We found no interaction between stimulus type, time window and scenario type. Pairwise comparisons revealed the following three findings: First, we found a P300 effect in the target condition for each time window as well as for each scenario type [simple oddball: $p < 0.001$ for early time window, $p < 0.035$ for late time window; Brio oddball: $p < 0.001$ for early time window, $p < 0.002$ for late time window]. However, we found a P300 effect in the deviant condition for each time window only in simple oddball [early time window: $p < 0.001$, late time window: $p < 0.021$]. In the Brio oddball condition, the P300 effect in the early time window was absent, even though we found a P300 effect in the late time window [early time window: $p = n.s.$, late time window: $p < 0.019$]. Secondly, we found a stronger P300 effect in the early time window compared to the late time window. The stronger P300 effect in the early time window could be shown for both stimulus conditions in simple oddball [targets: $p < 0.003$; deviants: $p < 0.002$]. In Brio oddball, P300 in the early time window was just as strong as in the late time window for the target condition. Thirdly, we found a scenario-specific difference that could only be shown in the early time window for the deviant condition.[$p < 0.026$]

Concerning P300 peak latency, there was no main effect of stimulus type [$F(2,28) = 2.433, p = n.s.$] as well as scenario type [$F(2,28) = 3.139, p = n.s.$]. The stimulus type did not interact with the scenario type [$F(2,28) = 3.139, p = n.s.$]. Not surprisingly, there was a main effect of time window [$F(1,14) = 195.156, p < 0.001$]. The time window did not interact with the scenario type [$F(1,14) = 0.399, p = n.s.$] nor with the stimulus type [$F(2,28) = 1.0, p = n.s.$]. There was no interaction between stimulus type, scenario type, and time window. Pairwise comparisons revealed that there was a latency difference between simple oddball and Brio oddball in the early time window of target condition [$p < 0.007$]. There was also a latency difference between target and deviant in Brio oddball [$p < 0.044$].

In summary, in the simple oddball scenario, we found a P300 effect elicited by targets as well as a reduced P300 effect elicited by deviants. In the Brio oddball scenario, we found

a)



b)

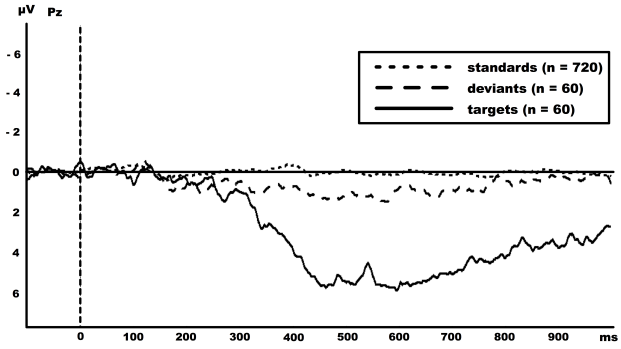


Fig. 2. Grand averages over 8 data sets each scenario (2a: **simple oddball** scenario; 2b: **Brio oddball** scenario); [artifact-free segments [standards/deviants/targets (hits)]: simple oddball [48%/54%/52%], Brio oddball [45%/47%/32%]

a P300 effect with an extended peak latency in the target condition, whereas a P300 effect was absent in the deviant condition within the early time window. Also we found a delayed peak latency of P300 in Brio oddball compared to simple oddball, when we concern the early time window.

3) Discussion of offline data analysis: In our discussion we will focus on four main findings in the early time window: (1) the stable P300 in the manipulation like scenario, (2) the delayed peak latency of P300, (3) the broader P300 after targets in the early time window in the Brio oddball scenario in comparison to P300 elicited by targets in the simple oddball scenario, and (4) the missing P300 after deviants in the Brio oddball scenario.

One of our main findings is a stable P300 in the Brio oddball scenario. There was no significant reduction in P300 amplitude due to the dual task. Thus, P300 is a stable measure for information processing in a manipulation-like scenario. Similar results could be shown by Fowler at al. [8].

However, we found a delayed peak latency of P300 in the Brio oddball scenario compared to the simple oddball scenario (see Fig 2a vs. Fig 2b). The Brio oddball scenario makes higher demands on the subject's cognitive processing, since the subject has to pay attention to two tasks, motor response after successful perception and classification of targets and playing the BRIO® labyrinth game, which require different cognitive procedures. Actually, both tasks involve stimulus processing of the same modality, namely visual processing. Thus, cognitive resources have to be shared. Selective attentional processes are involved in generating a P300. Since motor response after correctly classified target stimuli as well as playing the game involve selective attentional processes, both tasks compete for the same cognitive resources involved to direct attention. This in turn makes stimulus evaluation more difficult and complex and results in a delay of stimulus classification in the Brio oddball scenario reflected in a delayed P300 [8], [18], [11].

The finding of a broad P300 elicited by targets in the Brio oddball scenario (see Fig 2b) can be interpreted in three ways. Firstly, we could see stronger inter-subject differences in P300 latency in the Brio oddball scenario than in the simple oddball scenario, which causes a broader morphology of the grand averaged P300 in the Brio oddball scenario. Those inter-subject differences might result from the demand that subjects had to manage a conflict situation of solving two possible tasks at the same time (playing the BRIO® labyrinth game and pressing the buzzer). An individual subject might solve this conflict situation differently. We noticed that subjects that played the labyrinth very well and concentrated had more problems to focus on the oddball task. Secondly, playing the BRIO® labyrinth game might be different at any individual trial[3], namely more or less complicated and the subject might at any time play more or less concentrated. The more difficult a situation in the game is, the more a subject is distracted from paying attention to the oddball task. This might in turn influence P300 peak latency in every single trial P300. Inter-trial differences in P300 single trial peak latency can result in a broader averaged P300 potential for each subject [20]. Thirdly, though a flat grand average curve can be caused by latency differences between averaged P300 from individual subjects as well as differences between single trials within subjects, it can also result from overlapping potentials [19], e.g. P300 overlapping with a SW [32].

The higher cognitive load in the Brio oddball scenario resulting from the dual task might also be a reason for the absence of a P300 effect in the deviant condition in the early time window of the Brio oddball scenario (see Fig 2b). Our results indicate that the deviants were ignored, i.e. no longer perceived as warnings. This finding is consistent with the P300 pattern of the deviant condition in the simple oddball scenario which is characterized by a stronger P300 effect of deviants compared to the Brio oddball scenario (see Fig 2a). This can be explained with the quality of P300 to be associated with the information processing of the task-relevant stimuli. In our case, target stimuli are task-relevant and thus targets elicit the stable P300 effect irrespective of the experimental scenario type. In contrast, the occurrence of P300 elicited by deviants depends on whether deviants are perceived as warning, i.e.

[3]Trial stands for every single EEG epoch time-locked to a certain event, containing evoked brain activity.

as task-relevant stimuli or not. Since in the Brio oddball scenario attention was divided and cognitive load was much higher than in the simple oddball scenario, subjects ignored the deviant stimuli which were not necessarily task-relevant. This is because they predicted the occurrence of a target stimulus not reliably but weakly only[4]. In contrast to the simple oddball senario, subjects were not able to keep enhanced attention on the oddball task after the occurrence of a deviant stimulus and in anticipation of a target stimulus since they were forced to perform the game and thus were distracted from keeping the attention on the oddball task. Our findings confirm results from Israel et al. [15].

In summary, we could show that the P300 potential is elicited after task-relevant stimuli in a stable manner and is reduced or absent for rare but non task-relevant stimuli. Also, we found that in a scenario that requires complex manipulation and multiple tasking, only stimuli that are very important, task-relevant and cognitively processed elicit a P300 potential. Based on both findings, we presume that P300 is a good indicator for cognitive processing in multitasking scenarios. Regarding the broader P300 morphology in Brio oddball, we assume that there are two possible reasons: different P300 latencies at the level of single trials and inter-subjects differences in P300 latency at the level of averages. An approach for single trial ERP detection has to be robust to inter-trial and inter-subject variances, which can be achieved by using appropriate feature generation techniques (see section III-A3).

III. Single-Trial Brain Reading

A. Concept

In this section, we present a concept for single-trial classification of an operator's information processing in a telemanipulation scenario. More precisely, in this scenario the single-trial brain reading device has to make a decision whether an operator did perceive an important message or whether he did not. For such a system, we have identified the following steps for the processing: (1) Subdividing the continuous EEG into fixed-length time windows (windowing), (2) increasing the signal-to-noise ratio (preprocessing), (3) extracting stable features from the EEG time windows (feature generation), and (4) deciding whether the one or the other condition was present based on the extracted features (classification). Due to the large inter-subject (and inter-session) variance, feature generation and classification (and partly preprocessing) should be adapted to the specific subject for each session. This can be achieved by a separate calibration session prior to each actual session, in which some representative examples under the different conditions are recorded (see Section III-B). Based on this training data, machine learning techniques can be applied in order to detect promising features and to learn good classification strategies.

[4]A deviant stimulus occurred one or several trials before each target stimulus.

1) Windowing: For each message presented, exactly one decision has to be made. All information that can be used for this decision is usually contained in a certain, fixed time-range around the message presentation. Thus, the decision could be done based solely on the EEG recorded in the second after message presentation. The process of extracting this time window is called "Windowing". Windowing allows to simplify computation since it allows to work always on instances of the same shape (length of the signal frame).

2) Preprocessing: Preprocessing refers to operations aimed at increasing the signal-to-noise ratio. It requires some assumptions on which components of the time window are considered useful and which are considered as noise. For example, high frequency noise can be removed through a low-pass filter. Another preprocessing method is spatial filtering, which refers to methods that combine information of several channels and create a new (usually smaller) set of pseudo channels. The objective is to create channels that contain a high signal content while the noise is more concentrated in the remaining channels.

3) Feature generation: Finding features that are not strongly influenced by inter-trial and inter-session variances is important since it increases the probability that a classifier trained on these features achieves a good performance also under conditions that have not been tested during the calibration phase. Possible kinds of features are the power of a certain frequency band in a certain channel, the correlation of two channels within a certain time bin, or the value of a channel at a certain point in time. Our approach is to generate a large set of features (in the order of 10^3 to 10^4 features) and to use supervised feature selection methods (see for example Guyon and Elisseeff [10] for an overview) to identify subject-specific features that have a high discriminative power regarding the two classes. Supervised feature selection methods require labeled examples that can be obtained during the calibration session. In order to find not only predictive but also stable features (i.e. features that have a high predictive power over a broad range of problems), the feature selection can be performed on different subsets of the whole calibration data set (or from different calibration sessions). Features that are selected in a high percentage of the subsets are likely to be stable. Studying feature stability is of special importance in the light of identifying new feature generation methods, which we are investigating, that are well-suited to handle inter-trial variances due to varying latencies of P300 components and correlated brain activity changes.

4) Classification: Based on the extracted features, a subject-specific classification strategy needs to be derived. Given the windowed data along with the respective labels from the calibration session, that imposes an instance-based machine learning task, any kind of classification algorithm suited for binary decision tasks can be used to learn a user-specific classification strategy. We plan to systematically evaluate which combination of features and classification algorithms maximizes the predictive performance over a broad range of subjects.

B. Calibration

Calibration refers to the process of collecting representative example recordings of an operator's brain activity along with label information indicating the conditions a classifier should predict later. These training data should cover different situations that are likely to occur during usage. For example, in a telemanipulation scenario, the operator should actually manipulate something during the calibration, he should be situated in the same environment, and he could be put under time pressure. During the calibration session, we ask subjects to press a buzzer to obtain the information if a presented stimulus has been perceived.

C. Software

In this section, we describe the software framework that has been developed in order to implement the concept outlined above. This framework consists of two main parts: an EEG acquisition infrastructure and a data processing part called *Brain Reading Interface - Data Processing (BRI-DP)*. See Figure 3 for a dataflow diagram of software framework.

The *EEG acquisition infrastructure* is designed so that (1) it is suited for both online processing of EEG data and offline benchmarking of signal processing and machine learning methods, (2) EEG acquisition and processing can take place on two different machines, and (3) it is real-time capable. The first requirement is fulfilled by a component that provides a common interface to access EEG data. Internally, this component can read this data from a file or acquire it online from a subject. The second requirement is fulfilled through a TCP-based communication layer that allows to stream EEG data from the machine where it is acquired/stored to the machine where it is actually processed. Sending of EEG is done by the BRI EEG data protocol server and receiving by the BRI EEG data protocol client. The third requirement will be achieved by parallel processing of whole windows or of the individual channels on multiprocessor and/or graphics processing units (GPUs) based architectures. Furthermore, the EEG acquisition infrastructure is also responsible for subdividing the data into windows of fixed length. This windower component can be configured by means of a configuration file in which rules are defined that specify when to extract a window. For instance, whenever an important message is issued to the operator, a marker is inserted into the EEG. A typical rule for the windower would be to extract the 1 second of EEG that follows such a marker.

The BRI-DP is based on the Modular toolkit for Data Processing (MDP) [36]. The MDP allows to specify a data processing procedure by means of a data flow, in which every processing step is modelled as a node and a sequence of nodes constitutes a (data-)flow. This allows to easily "plug together" different algorithms and to exchange one component of a flow by another in order to compare their relative performance. This is particularly useful for the empirical comparison of different preprocessing, feature selection, and classification methods. The MDP already offers a magnitude of signal processing and machine learning algorithms, for instance the spatial filtering methods Independent Component Analysis (ICA) [13], Principal Component Analysis (PCA) [17], and the classification method Linear Discriminant Analysis [3]. These data processing units can be combined into data processing flows and also more complex feed-forward network architectures. BRI-DP extends MDP in two ways: on the one hand, further algorithms have been added like low-pass and band-pass filters, the Common Spatial Patterns (CSP) algorithms (see for example Blankertz et al. [4]), and several feature extraction methods based on properties like frequency band power, amplitudes, and the pairwise correlation and coherence of channels. On the other hand, the semantics of flows has been changed slightly so that cross validation is supported, intermediate results can be stored and loaded, and that BRI-DP flows can be specified by means of a configuration file. Furthermore, BRI-DP can be easily integrated into benchmarking frameworks.

D. Evaluation

The signal frames that are extracted from the continuous EEG data stream during windowing are preprocessed and presented to the classification algorithm after the treated signal has been transformed into a feature representation. Frames are extracted from the stream following the occurrence of a marker that signals the presentation of a message to the subject. During the calibration phase, the subject presses a buzzer to acknowledge perception of a message (see section II-B). The buzzer press event is recorded as a second marker type in the EEG stream and extracted windows are labeled according to the existence of such a buzzer press event following presentation of informational messages. As such, each extracted signal frame constitutes a labeled instance (example) used to train the machine learning algorithm during the training phase. During the usage phase, the classification algorithm is asked to assign a class label ("message was perceived" vs. "message was not perceived") or corresponding probability estimate to each unlabeled signal frame, after it has been subjected to the same preprocessing and feature extraction treatment as frames during the training phase.

The windowing-based approach adopted here lends itself to the use of classical instance-based performance metrics that are long established in machine learning. When evaluating preprocessing and feature extraction methods in combination with a classification algorithm on test data, a confusion matrix can be calculated showing the frequencies of true and false positive (TP, FP) and true and false negative (TN, FN) predictions on the test data. A very simple performance estimate is the *accuracy* which denotes the total fraction of correct predictions ($\text{acc} = \frac{\text{TP}+\text{TN}}{\text{P}+\text{N}}$). Accuracy, however, is not an adequate performance measure in applications with imbalanced classes. For example, in the scenario presented in section II-B, 720 standard messages versus only 60 target messages (relevant information) are shown to the user. A trivial classifier that needs no training and always predicts the majority class (standards) would yield an accuracy of 92.3% on these data, which clearly does not give a useful estimate of the actual performance in the application. Measures from information retrieval, such

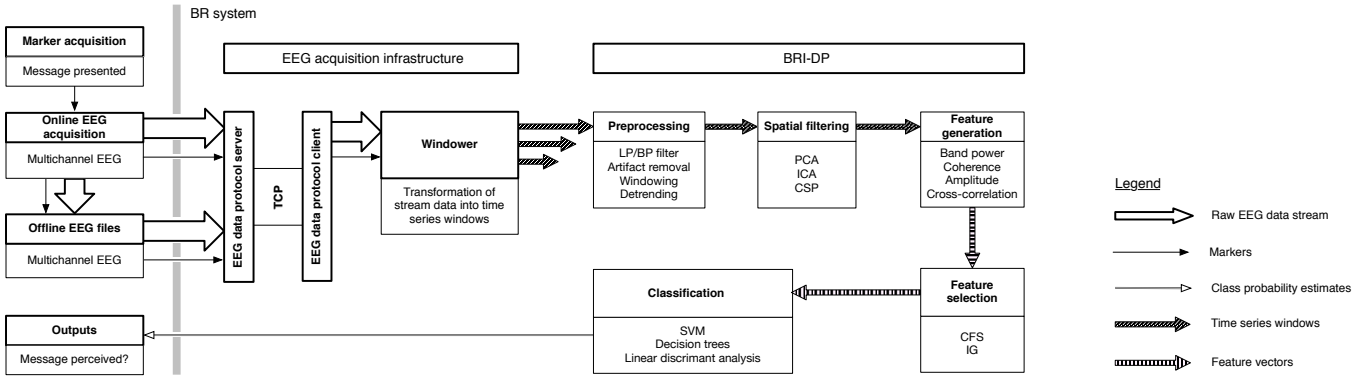


Fig. 3. Data processing architecture of the adaptive Brain Reading Interface. EEG data can be streamed via a TCP-based protocol from one machine to another so that acquisition and processing can take place on two different machines. The continuous EEG data stream is subdivided into consecutive parts by the windower. The different processing stages of the BRI-DP are applied to these parts in order to obtain a binary classification about whether an important message has been perceived or not. The software implementation of the BRI-DP is modular so that the different algorithms of the processing stages can be combined arbitrarily.

as precision (pre $= \frac{\mathrm{TP}}{\mathrm{TP+FP}}$) and recall (rec $= \frac{\mathrm{TP}}{\mathrm{P}}$) are much more suited to the application at hand. We define the positive class to consist of those signal frames recorded from an individual that has just consciously perceived a presented message. Then precision gives an intuitive measure of how many of the messages that are predicted by the BR system have been perceived by the user have indeed been consciously processed. Naturally, one would aim to optimize this measure for certain applications where a (possibly critical) piece of information that was wrongly classified is not brought to the attention of the user for a second time and thus lost. Recall, on the other hand, gives an estimate of the number of consciously perceived messages that have not been classified as such. For uncritical warnings, it might be sensible to optimize this metric to avoid disturbing the operator by a repeated display of warning messages until the classifier has finally recognized the operator's conscious processing of the information.

There is a trade-off between precision and recall that can be made by choosing a threshold when mapping class probability estimates (if the chosen classifier outputs these) to nominal classes or selecting classification algorithms that perform particularly well in either dimension. This trade-off can be visualized by ROC plots [7]. Precision and recall can also be combined into a single value called F-measure [33].

The goal of benchmarking is not only to evaluate the performance of different methods in isolation, but also to determine which method performs best [12]. We adopt an exploratory benchmarking strategy to identify the best-performing preprocessing and machine learning workflows for our application using an automated benchmarking process in conjunction with grid computing. To this end we perform a large number of benchmarking experiments to compare different methods on a set of EEG data from usually one or several individuals. Due to the nature of the problem, individual benchmarking experiments are not statistically independent, hence, the correct choice of appropriate test statistics gains special importance to ensure that winner methods identified during automatic benchmarking perform equally well in the real-world application [24], [29].

IV. Outlook

In the future, we will focus on the acquisition of EEG data from subjects that are situated in virtual environments since telemanipulation can be more effective by placing an operator in a virtual environment to allow telepresence [9], [1]. Such a scenario imposes additional challenges; for instance, devices that are needed to situate the subject in the virtual reality like headsets might cause artifacts in the EEG (e.g. 50 Hz or muscle artifacts due to the extra load of wearing a heavy headset). These artifacts are an additional kind of "noise" and thus reduce the signal-to-noise ratio. We will investigate how a single-trial brain reading system can deal with this kind of noise. Beside this, subjects might behave differently in a virtual environment than in reality and might be under even greater cognitive load since they might be confused by artifacts of the simulation environment, i.e., situations where the simulation behaves slightly different than the reality.

Furthermore, we will examine not only ERP signals but also correlated changes in different EEG frequency bands and the sources of ERP signals in the brain. Besides ERPs that are evoked by the processing of information, ERPs that precede motor behavior as well as ERPs that are correlated with attentional processes will be investigated. We will analyze whether the combination of the monitoring of different cognitive processes lead to more precise prognoses of mental states of the operator and could maybe be used to forecast future behavior. Both, motor-related ERPs and attention-related ERPs can give further insights into the planning and execution of behaviour. This is valuable for both monitoring of operators in telemanipulation scenarios and in other complex scenarios

that involve the direct control of machines or devices via parts of the human body, like the control of a robotic manipulation arm via an exoskeleton.

For single-trial analysis, the choice of appropriate features is crucial for a good predictive performance of a classifier. We will investigate which kind of features are stable, i.e. contain a high information content across sessions and subjects. Besides the inter-trial variance, during a session there might be also systematic changes in the response of an operator's information processing. For instance, in a telemanipulation scenario, an operator might fatigue over time. This in turn might increase the latency and decrease the amplitude of the P300 potential after presentation of an important message [23]. Because of this, a brain reading system has to adapt over time to these changes. Online adaptation will be a focus of further work.

Furthermore, we will conduct studies that compare the influence of different spatial filtering and classification algorithms on the performance. Based on the results of these offline studies, an online system will be developed that is able to classify in real time whether an operator has consciously perceived a message. This brain reading system will be tested in a real-world telemanipulation scenario. In the long run, the processing must be implemented on portable hardware (like FPGA) to increase the usability of the system and reduce energy consumption. This, together with the recent progress in the development of drycaps [27], promises that in the future EEG recordings can be used in application scenarios without the need of time-consuming preparations of EEG electrodes and restrictions due to high cost and bulky analysis hardware.

ACKNOWLEDGMENT

The authors would like to thank Manfred Fahle for reading the manuscript and helpful comments, Johanna Marquardt for help in EEG data acquisition and Alexander Boettcher for writing programs for behavioral data analysis.

REFERENCES

[1] M Anvari. Robot-assisted remote telepresence surgery. *Surgical Innovation*, Jan 2004.

[2] JD Bayliss. Use of the evoked potential P3 component for control in a virtual apartment. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):113–116, 2003.

[3] CM Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, August 2006.

[4] B Blankertz, R Tomioka, S Lemm, M Kawanabe, and KR Müller. Optimizing spatial filters for robust EEG Single-Trial analysis. *Signal Processing Magazine, IEEE*, 25(1):4156, 2008.

[5] J Day. Review of nasa-msc electroencephalogram and electrocardiogram electrode systems including application techniques. *Technical note. United States. National Aeronautics and Space Administration*, Jan 1968.

[6] LA Farwell and E Donchin. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalogr Clin Neurophysiol*, 70(6):510–23, Dec 1988.

[7] T Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, June 2006.

[8] B Fowler. P 300 as a measure of workload during a simulated aircraft landing task. *Human factors*, 36(4):670–683, 1994.

[9] SM Goza, RO Ambrose, MA Diftler, and IM Spain. Telepresence control of the nasa/darpa robonaut on a mobility platform. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 623–629, New York, NY, USA, 2004. ACM.

[10] I Guyon and A Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:11571182, 2003.

[11] J Hohnsbein, M Falkenstein, and J Hoormann. Effects of attention and time-pressure on P300 subcomponents and implications for mental workload research. *Biological Psychology*, 40(1-2):73–81, 1995.

[12] T Hothorn, F Leisch, A Zeileis, and K Hornik. The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics*, Jan 2005.

[13] A Hyvärinen, J Karhunen, and E Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

[14] J Isreal, G Chesney, C Wickens, and E Donchin. P300 and tracking difficulty: Evidence for multiple resources in dual-task performance. *Psychophysiology*, 17(3):259–73, Jan 1980.

[15] J Isreal, G Chesney, C Wickens, and E Donchin. P300 and tracking difficulty: Evidence for multiple resources in dual-task performance. *Psychophysiology*, 17(3):259–73, Jan 1980.

[16] R Johnson. A triarchic model of P300 amplitude. *Psychophysiology*, 23(4):367–84, Jan 1986.

[17] IT Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, 2002.

[18] M Kutas, G McCarthy, and E Donchin. Augmenting mental chronometry: the P300 as a measure of stimulus evaluation time. *Science*, 197(4305):792–5, Aug 1977.

[19] SJ Luck. *An introduction to the event-related potential technique*. MIT Press, Cambridge, Jan 2005.

[20] A Magliero, TR Bashore, MGH Coles, and E Donchin. On the dependence of p300 latency on stimulus evaluation processes. *Psychophysiology*, 21(2):171–186, 1984.

[21] JH Metzen, EA Kirchner, L Abdenebaoui, and F Kirchner. Learning to play the BRIO labyrinth game. *Special Issue "Reinforcement Learning" of the KI Magazine*, 2009. Accepted.

[22] KR Müller, M Tangermann, G Dornhege, M Krauledat, G Curio, and B Blankertz. Machine learning for real-time single-trial eeg-analysis: from brain-computer interfacing to mental state monitoring. *J Neurosci Meth*, 167(1):82–90, Jan 2008.

[23] A Murata and A Uetake. Evaluation of mental fatigue in human-computer interaction-analysisusing feature parameters extracted from ebent-related potentials. *10th IEEE International Workshop on Robot and Human Interactive Communication*, Jan 2001.

[24] C Nadeau and Y Bengio. Inference for the generalization error. *Mach Learn*, 52(3):239—281, Jan 2003.

[25] T Picton. The P300 wave of the human event-related potential. *Journal of clinical neurophysiology*, 9(4):456–479, Jan 1992.

[26] J Polich and A Kok. Cognitive and biological determinants of P300: an integrative review. *Biological Psychology*, 41(2):103–146, 1995.

[27] F Popescu, S Fazli, Y Badower, B Blankertz, and KR Müller. Single trial classification of motor imagination using 6 dry eeg electrodes. *PLoS ONE*, 2(7):e637, Jul 2007.

[28] W Ritter and H Vaughan. Averaged evoked responses in vigilance and discrimination: A reassessment. *Science*, Jan 1969.

[29] SL Salzberg. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1:317—327, 1997.

[30] G Schalk, DJ McFarland, T Hinterberger, N Birbaumer, and JR Wolpaw. Bci2000: a general-purpose brain-computer interface (bci) system. *IEEE transactions on bio-medical engineering*, 51(6):1034–43, Jun 2004.

[31] NK Squires, KC Squires, and SA Hillyard. Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli. *Electroencephalogr Clin Neurophysiol.*, 38(4):387–401, April 1975.

[32] S Sutton and DS Ruchkin. The late positive complex. advances and new problems. *Ann N Y Acad Sci*, 425:1–23, Jan 1984.

[33] CJ Van Rijsbergen. *Information retrieval (2nd edition)*. Butterworths, London, Jan 1979.

[34] R Verleger, W Heide, C Butt, and D Kömpf. Reduction of P3b in patients with temporo-parietal lesions. *Brain research Cognitive brain research*, 2(2):103–16, Sep 1994.

[35] JR Wolpaw, N Birbaumer, DJ McFarland, G Pfurtscheller, and TM Vaughan. Brain-computer interfaces for communication and control. *Clin Neurophysiol*, 113(6):767–91, Jan 2002.

[36] T Zito, N Wilbert, L Wiskott, and P Berkes. Modular toolkit for data processing (MDP): a python data processing framework. *Frontiers in Neuroinformatics*, 2:8, 2008. PMID: 19169361.