Influence of User Tasks on EEG-based Classification Performance in a Hazard Detection Paradigm

Henrich Kolkhorst[1], Saku Kärkkäinen, Amund Faller Raheim, Wolfram Burgard and Michael Tangermann[2]

***Abstract*— Attention-based brain-computer interface (BCI) paradigms offer a way to exert control, but also to provide insight into a user's perception and judgment of the environment. For a sufficient classification performance, user engagement and motivation are critical aspects. Consequently, many paradigms require the user to perform an auxiliary task, such as mentally counting subsets of stimuli or pressing a button when encountering them. In this work, we compare two user tasks, mental counting and button-presses, in a hazard detection paradigm in driving videos. We find that binary classification performance of events based on the electroencephalogram as well as user preference are higher for button presses. Amplitudes of evoked responses are higher for the counting task—an observation which holds even after projecting out motor-related potentials during the data preprocessing. Our results indicate that the choice of button-presses can be a preferable choice in such BCIs based on prediction performance as well as user preference.**

I. Introduction

Brain-computer interfaces (BCIs) are characterized by the experimental paradigm that is utilized to elicit discriminative brain signals. Yet for some paradigms—especially complex stimuli—it is not well established which exact task a user should perform in order to optimize the usability of the BCI or to allow for studying a targeted neuroscientific research question.

In a visually simple detection task, Wentzel and colleagues reported that mental counting outperformed mental arithmetic and memorization in terms of classification accuracy obtained for BCI [1]. In the context of motor tasks in BCIs, the task spectrum reaches from overt motor execution on the one end over attempted motor execution (in paralyzed patients) to motor imagery at the other end. In this context, Nikulin and colleagues [2] have proposed an interesting and efficient quasi-movement task for experimenting with healthy users. In attention-based BCI paradigms, the user tasks reported in literature range from the sole observation of stimuli over mentally counting of stimulus subclasses to explicit motor responses upon events, such as button presses, pressing a foot pedal [3], or a combination thereof [4]. For a movement-related visually evoked potential paradigm, Guo and colleagues have reported that mental counting evokes stronger responses than gazing [5] .

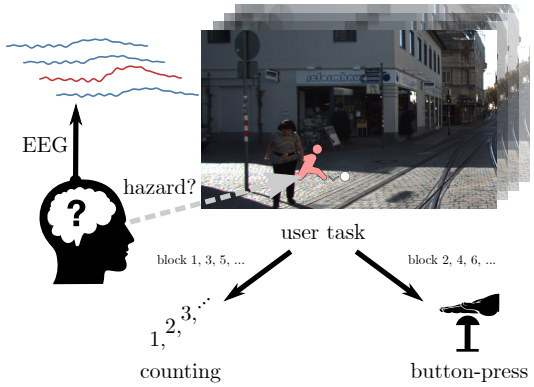

Fig. 1. Overview of experiment: Subjects watched driving videos augmented with potentially hazardous pictograms while performing a task: mentally counting hazards or pressing a button after each hazard. Tasks alternated between experiment blocks.

Given the choice of a user task when designing an attention-based experimental paradigm, multiple criteria influence a task's suitability: First, it should be feasible and well-accepted by the user. Second, it should not distract the user from the actual primary task (attention to the stimuli) in order to elicit the desired brain signals. Third, the task should allow for a sufficient classification performance to be useful with respect to the BCI's objective. Additionally, the task should not introduce an unnecessary amount of artifacts or confounding brain activity detrimental to the use case (e.g., in rehabilitation paradigms). Naturally, these criteria are not independent of each other. For example, a more engaging user task will likely also contribute to a good classification performance, but may at some point entail a higher rate of muscular or ocular artifacts.

In this paper, we present a pilot study in which we compare mental counting and button presses as user tasks in a paradigm for hazard detection in driving videos [6]. The paradigm is characterized by its naturalistic stimulus material, which we use to learn about the user's perception of the environment. The two user tasks investigated here, mental counting and pressing a button upon hazard perception, are representative for a common trade-off encountered during experiment design: While mental counting is known to perform well, it is usually only loosely related to the application and therefore hard to motivate from a user's perspective. Button presses can be seen as a simpler alternative that can be easier to incorporate and motivate (e.g., performing an emergency stop). However, the resulting motor activity can potentially create confounders for the later analysis.

All authors are with the Department of Computer Science, University of Freiburg, Germany. H.K., W.B. and M.T. are with the Autonomous Intelligent Systems group and H.K. and M.T. are also with the Brain State Decoding Lab. This work was (partly) supported by BrainLinks-BrainTools, Cluster of Excellence funded by the German Research Foundation (DFG, grant number EXC 1086). Additional support was received from the German Research Foundation through grant INST 39/963-1 FUGG and from the Ministry of Science, Research and the Arts of Baden-Württemberg for bwHPC.

[1] kolkhorst@informatik.uni-freiburg.de
[2] michael.tangermann@blbt.uni-freiburg.de

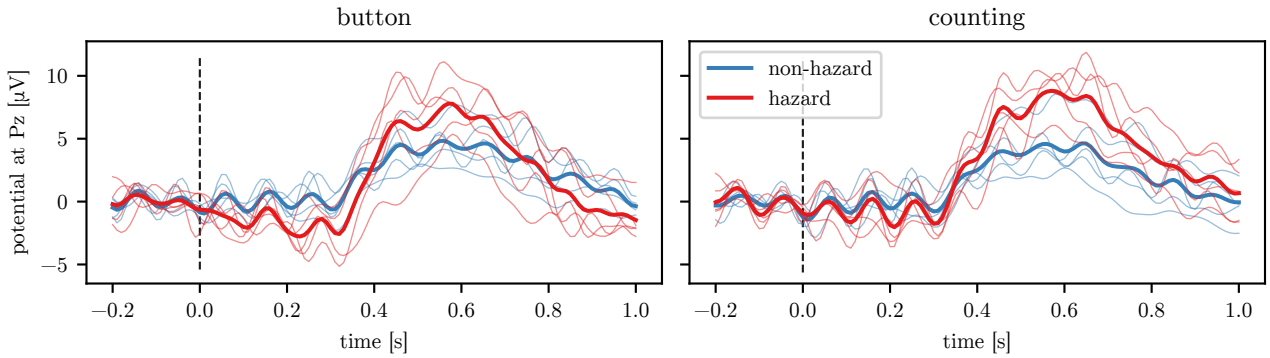

Fig. 2. Responses to hazardous and non-hazardous stimuli at electrode Pz for both the button-press and the counting task. The stimulus appears at time 0 s. The thin lines represent the class-wise average for each of the subject and the thick lines the grand average over all five subjects.

In the remainder of this paper, we evaluate and compare the two user tasks with regard to the user's preferences, the evoked event-related potentials (ERPs) and the classification performance both on the same task as well on the complementary one. We show how a reduction of the confounding motor activity by projecting out corresponding motor-related components of the recorded data can reduce the influence of confounders in the button press condition and improve classification performance. Additionally, we discuss these results in the context of choosing an adequate user task when designing a BCI paradigm.

II. Methods

A. Paradigm and Experimental Data

We recorded the electroencephalogram (EEG) from six healthy subjects, who each participated in a single experiment session. Following the declaration of Helsinki, we received approval by the local ethics committee for this study and obtained written informed consent from participants prior to the session. Subjects were seated approximately 80 cm in front of a screen, on which they were shown videos of natural driving environments based on the KITTI dataset [7]. In order to create (potentially) hazardous situations, videos were augmented with pictograms (e.g., pedestrians or playing children). More details on the used stimulus material can be found in [6].

Each session consisted of 20 blocks. Every block encompassed 12 videos with a duration of 20 s each that were separated by a short break. A single video contained between 0 and 8 occurrences of different pictograms, which we will subsequently call *events*. Events were labeled as either non-hazardous (e.g., a pedestrian walking on the sidewalk) or hazardous (e.g., a pedestrian crossing the street in front of the virtual car). Hence, the context rather than the identity of the stimulus determines the class label. A session consisted of 871 events out of which 19.40 % were hazardous.

Prior to experiment start, subjects were instructed that they should assume being a passenger in an autonomous vehicle. Depending on the experiment block, they had to execute one out of two tasks: button-press or counting. In the *button-press* task, subjects were told to press a button, which was placed in their right hand, once they (subjectively) considered a situation to be hazardous. In the *counting* task, they were told to silently count the number of hazards they observed while avoiding any movement. During a short pause after each video, subjects reported their counts to the experimenter. Blocks were assigned to tasks in an alternating fashion.

We acquired the brain signals using a cap holding 63 Ag/AgCl gel-based passive EEG electrodes positioned according to the extended 10-20 system with a nose reference. Channel impedances were kept below 20 kΩ. The amplifier sampled the EEG signals at 1 kHz.

B. Data Analysis

Analysis was performed separately for each subject in an offline manner. Initially, the EEG data was filtered to a frequency band of 0.50 Hz to 16 Hz using a FIR filter. In order to investigate the influence of motor activity, we perform the subsequent analysis steps in two preprocessing settings. In the *raw* setting, filtered data is directly segmented into event epochs as described below. In the *redu-motor* setting, we first reduced motor-related activity by projecting out corresponding spatial components from the data. For this, we selected data segments from −0.70 s to 0.50 s relative to each button press in the session and baselined these segments based on the first 0.20 s. We extracted a full set of spatial filters using xDAWN [8], a supervised algorithm that is trained to increase contrast between the extracted motor-related segments and "background" segments extracted from all video portions of the recording.

Components revealing strong motor activity were projected out: We removed data corresponding to the motor components by setting the corresponding surrogate channels (as extracted by xDAWN) to be constant. For this study, we manually selected a single component per subject out of the six components with highest eigenvalues. Note that the data now has a reduced rank, which we counter in the following data processing by regularization as described below.

Subsequently, the continuous recording was segmented into a single epoch per event. Each epoch consists of the data from −0.20 s to 1 s relative to the onset of the stimulus (i.e., the first video frame in which a pictogram appears). Note that not all of the stimuli may have been fully visible in the first frame (e.g., when a person (re)appears from an

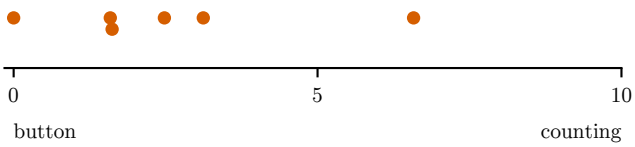

Fig. 3. User rating regarding the preferred task in a post-session questionnaire. Markers correspond to the rating of individual users.

occlusion). Epochs were labeled hazardous or non-hazardous based on the context of the event. While we did not use button presses to label events, for the behavioral analysis we considered them to be in response to an event if they are within $0.20\,\mathrm{s}$ to $1.20\,\mathrm{s}$ of the event onset. Epochs were corrected for signal drifts with respect to the first $0.20\,\mathrm{s}$ after event onset. We rejected epochs in which the peak-to-peak amplitude exceeded $70\,\mu\mathrm{V}$ in any channel. Data of one subject was excluded due to extremely high noise (more than $50\,\%$ of epochs had to be rejected).

For the classification of epochs into *hazard* and *non-hazard*, we used a covariance-based feature representation [9], [10]. On the training data, we extracted three xDAWN filters [8] for each class. Epoch data was projected using these filters. We augmented epochs with prototype responses (based on the xDAWN-filtered class-means in the training data) before calculating the covariance matrix of each epoch. For this, we used a Ledoit-Wolf regularization. These regularized covariances were projected into the Riemannian tangent space at the mean of the training data and subsequently classified using Logistic Regression. For details, see [11].

For the complete data as well as the within-task evaluation, we performed a 5-fold chronological cross-validation to estimate classification performances. For between-task transfer, classifiers were trained similarly on $80\,\%$ of the data of one task and were evaluated on $20\,\%$ of the data of the other task (repeated five times) in order to achieve comparable training data sizes to the within-class setting. In the following, we report results using the area under the receiver-operating characteristic curve (AUC).

III. Results

A. Behavioral Response

As subjects had been instructed to attend to events that they perceived as hazardous, the number of button presses as well as the reported hazard counts varied substantially between subjects (222 ± 68 button presses and 230 ± 58 counts). However, within subjects the numbers were consistent between tasks (Pearson correlation of 0.98 between number of button presses and hazard counts). The button presses had a mean latency of $0.65\,\mathrm{s}$ relative to the events.

A post-session questionnaire revealed that users on average preferred the button-press task over the counting task (see Fig. 3). As reasons they primarily mentioned that button presses are "easier" (three subjects) and more "interactive" (two subjects). The single subject who preferred the counting task similarly claimed that the button-press was easier, but reported to be more concentrated due to the higher workload setting of counting.

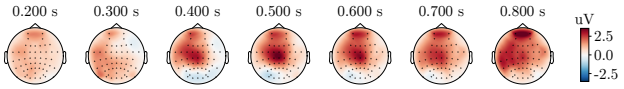

Fig. 4. Spatial distribution of difference in response to hazard stimuli (average activity of button task is subtracted from the one in the counting task). Each subplot shows the mean potential within a $100\,\mathrm{ms}$ window centered at the given time after event onset.

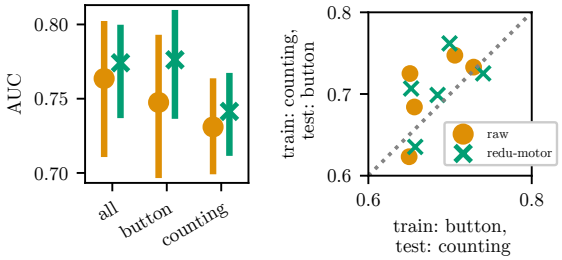

Fig. 5. Decoding quality (AUC) of single epochs (hazard/non-hazard). Colors denote the reduction of motor influence after removal of corresponding components (*redu-motor*) and the original data (*raw*). Left: Results from a cross-validation within the same task ("all" corresponds to pooling the data of both conditions) . Right: Generalization across tasks (trained on one task and evaluated on the other). Each marker corresponds to the classification results of one subject in one data setting.

B. Electrophysiology of Responses

As depicted in Fig. 2, we observed ERPs for both hazardous and non-hazardous events and under both user tasks. Hazardous events generally showed a stronger positive deflection between approximately $350\,\mathrm{ms}$ and $750\,\mathrm{ms}$ relative to the appearance of the stimulus. Comparing the two tasks, the latencies of ERPs are similar, whereas the amplitudes differ. As shown in Fig. 4, the potential in the counting task is higher at central electrodes from $0.40\,\mathrm{s}$ to $0.80\,\mathrm{s}$ after event onset. Neither the waveform characteristics (Fig. 2) nor the spatial distribution of the grand average ERP response (Fig. 4) for each task are affected by removing motor components in the *redu-motor* setting.

C. Hazard Decoding Performance

Pooling all data of both tasks in the *raw* condition, we achieved an AUC of 0.76 via a cross-validation as indicated by the leftmost yellow marker in the left subplot of Fig. 5. A comparison of decoding accuracies of the two single tasks reveals a loss of performance compared to the pooled data, but a higher classification performance in the button task (0.75 compared to 0.73 in the counting task). The scatter plot in the right subplot provides the individual performance values which can be achieved by transferring classifiers between tasks (i.e., training on the data corresponding to one task and evaluating on the other). The transfer for classifiers trained on counting and evaluated on button (mean AUC of 0.70) resulted in higher performances compared to the transfer in the opposite direction (mean AUC of 0.68).

Projecting out motor-related xDAWN components from data in the *redu-motor* setting (green markers) improved classification performance in each of the subjects and tasks. As expected, the AUC for the button task improved strongest by this removal from 0.75 to 0.78. Interestingly, we also

observed a small improvement in the counting task (0.73 to 0.74) after removing motor components. The improvements could also be observed for the cross-task transfer scenarios (green markers in right subplot of Fig. 5). The transfer from counting to button still outperformed the transfer from button to counting with regard to AUC (0.71 to 0.69), yet the transfer performance did not match the performance of a classifier trained on the same task.

IV. Discussion

In order to reduce the risk of confounders during the analysis of brain signals, motor tasks that are not mandated by the primary use case are often avoided in experimental paradigms of BCIs. In this paper, we find—in an attention-based visual hazard detection paradigm—that they are not necessarily detrimental to the analysis of evoked responses nor classification performance. Since the motor activity caused by button presses is not precisely time-locked to the onset of the visual event, we observe similar ERPs in the grand average in Fig. 2.

Similarly, the classification of hazards in the button-press task works well, with performance superior to performance during mental counting when each is evaluated within the respective task. However, the transfer to the other task works better for mental counting. At first glance, this might lead to the interpretation that the motor activity is a confounder that solely due to correlation with event classes can improve the classification performance (and, conversely, the lack thereof diminishes performance on the counting task). However, in the *redu-motor* setting we observe that projecting out components strongly tied to motor activity actually improves performance of the button-trained classifier on both tasks. Instead, the signals in the button-press portion of the data appear to be easier to discriminate into hazards and non-hazards. Due to the similar stimulus material as well as interleaved design, the higher user engagement would be a plausible explanation for this effect (i.e., training on the counting data does not necessarily generalize "better", but test data in this transfer direction is easier).

When projecting out motor-related components (*redu-motor* setting), performance of the classifiers trained on button-press data improved, providing a practical approach for reducing undesired effects on an individual epoch level. Observing both tasks within the same session, we could also apply the redu-motor setting to the counting task. Interestingly, we also see (smaller) gains in performance in this task (in which no button presses occurred). Possible explanations for this include (i) that components may capture not only motor activity, but also other confounders and (ii) that due to the interleaved design, subjects might be tempted to press the button even in counting tasks, leading to brain activity that can be captured by the button-trained components.

Overall, both mental counting and button-press tasks are feasible, evoked similar average responses, kept the user focused and allowed a classification into hazardous and non-hazardous events (albeit with performance differences between tasks). This leads to the question which task to choose when designing an experiment. The button-press task performs better than counting and is preferred by users, hence might be the clear choice when optimizing for classification performance. However, classification of the response to visual stimuli might also include motor performance. Projecting out components in the *redu-motor* tries to address this, with the higher performance potentially allowing a better BCI. Notwithstanding, depending on the application the counting task might also be preferable: In scenarios where feedback shall be provided on specific brain activity only (e.g., non-motor-related rehabilitation training paradigms) button presses may be prohibitive, as a contribution of motor-related brain activity in the classification cannot be ruled out. Consequently, the choice of the user task of an experimental paradigm should be made based on the trade-offs of the application. Based on the results of our pilot study, the use of button presses can be desirable, mirroring the preference of users.

V. Conclusion

Investigating the influence of button presses and mental counting as user tasks in a visual attention-based BCI, the results of our pilot study indicate that using a button-press task is preferred by most users and improves classification performance compared to counting, especially when motor-related components can be removed from the data during preprocessing. Counting constitutes a feasible alternative if influence of motor activity should be minimized.

References

[1] M. A. Wenzel, I. Almeida, and B. Blankertz, "Is Neural Activity Detected by ERP-Based Brain-Computer Interfaces Task Specific?" *PLOS ONE*, vol. 11, no. 10, p. e0165556, Oct. 2016.

[2] V. V. Nikulin, F. U. Hohlefeld, A. M. Jacobs, and G. Curio, "Quasi-movements: A novel motor–cognitive phenomenon," *Neuropsychologia*, vol. 46, no. 2, pp. 727–742, 2008.

[3] Z. Khaliliardali, R. Chavarriaga, L. A. Gheorghe, and J. d. R. Millán, "Action prediction based on anticipatory brain potentials during simulated driving," *J. Neural Eng.*, vol. 12, no. 6, p. 066006, 2015.

[4] S. Haufe, J.-W. Kim, I.-H. Kim, A. Sonnleitner, M. Schrauf, G. Curio, and B. Blankertz, "Electrophysiology-based detection of emergency braking intention in real-world driving," *J. Neural Eng.*, vol. 11, no. 5, p. 056011, Oct. 2014.

[5] F. Guo, B. Hong, X. Gao, and S. Gao, "A brain–computer interface using motion-onset visual evoked potential," *J. Neural Eng.*, vol. 5, no. 4, pp. 477–485, Nov. 2008.

[6] H. Kolkhorst, W. Burgard, and M. Tangermann, "Decoding Hazardous Events in Driving Videos," in *Proc. of the 7th Graz Brain-Computer Interface Conference*, Graz, Austria, Sept. 2017, pp. 242–247.

[7] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets Robotics: The KITTI Dataset," *The International Journal of Robotics Research*, p. 0278364913491297, Aug. 2013.

[8] B. Rivet, H. Cecotti, A. Souloumiac, E. Maby, and J. Mattout, "Theoretical analysis of xDAWN algorithm: Application to an efficient sensor selection in a p300 BCI," in *2011 19th European Signal Processing Conference*, Aug. 2011, pp. 1382–1386.

[9] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Multiclass Brain-Computer Interface Classification by Riemannian Geometry," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 4, pp. 920–928, Apr. 2012.

[10] A. Barachant and M. Congedo, "A Plug&Play P300 BCI Using Information Geometry," *arXiv:1409.0107 [cs, stat]*, Aug. 2014.

[11] H. Kolkhorst, M. Tangermann, and W. Burgard, "Guess What I Attend: Interface-Free Object Selection Using Brain Signals," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2018, pp. 7111–7116.