Proceedings of the
7th Graz Brain-Computer Interface Conference 2017 DOI: 10.3217/978-3-85125-533-1-45

DECODING HAZARDOUS EVENTS IN DRIVING VIDEOS

H. Kolkhorst[1], W. Burgard[1], M. Tangermann[1]

[1]Cluster of Excellence BrainLinks-BrainTools
Department of Computer Science
University of Freiburg, Freiburg, Germany

E-mail: kolkhorst@informatik.uni-freiburg.de

ABSTRACT: Decoding the human brain state with BCI methods can be seen as a building block for human-machine interaction, providing a noisy but objective, low-latency information channel including human reactions to the environment. Specifically in the context of autonomous driving, human judgement is relevant in high-level scene understanding. Despite advances in computer vision and scene understanding, it is still challenging to go from the detection of traffic events to the detection of hazards.
We present a preliminary study on hazard perception, implemented in the context of natural driving videos. These have been augmented with artificial events to create potentially hazardous driving situations. We decode brain signals from electroencephalography (EEG) in order to classify single events into hazardous and non-hazardous ones. We find that event-related responses can be discriminated and the classification of events yields an AUC of 0.79. We see these results as a step towards incorporating EEG feedback into more complex, real-world tasks.

INTRODUCTION

Humans can hardly compete with machines in purely computational tasks. Though the progress in machine learning and artificial intelligence in general has led to computers outperforming humans in difficult tasks such as playing the game of Go, it is still challenging to provide adequate interaction policies between humans and machines. This challenge is faced in application areas in which machines and humans both are actors, such as in collaborative manipulation tasks with robot arms or autonomous cars sharing the road with humans. Robots often require a high amount of adaptation to the human user, specifically by learning from her or him. In the following, we will focus on the driving domain, where many challenges in the interaction between intelligent vehicles and humans (be it as passengers, drivers or pedestrians) arise [1].
When considering complex (e.g., residential) driving environments, it is not sufficient to consider humans in the scene merely as (dynamic) *obstacles*. Rather, it is desirable to have a task-specific label for these obstacles, such as the respective hazardousness.

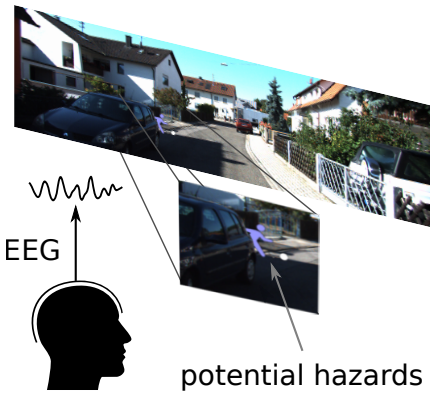

Figure 1: Experimental Paradigm: Recordings of driving scenes have been augmented with potentially hazardous events (artificial pictograms). The resulting videos are shown to the subjects while brain signals are recorded using electroencephalography.

We propose to utilize electroencephalography-based signals (EEG) to gather human feedback about the environment in a passive way. Alternative approaches to incorporating human teaching input, such as learning from demonstrations, can also be valuable tools [2]. However the situation- and context-dependence of preferred behavior (e.g., whether a situation should be treated as hazardous or not) suggests to instead gather feedback from the human, while she or he is acting within the target environment or within a reasonable approximation thereof. EEG signals, as opposed to behavioral feedback like button presses, offer the advantages of being non-intrusive and having a low latency. Additionally, brain-computer interfaces (BCIs) provide an unbiased feedback channel that corresponds to the subject's *individual* scene perception.
As the low signal-to-noise ratio of EEG poses practical challenges, it may be reasonable to seek for a balanced trade-off between a constrained lab environment and the final application environment to run experiments. Therefore, in this study we investigate passive viewing behavior of humans in continuous driving videos as a step towards monitoring humans as passengers in cars, with a possible application in the area of autonomous driving.

 DOI: 10.3217/978-3-85125-533-1-45

As illustrated in Figure 1, we use videos of natural driving scenes as stimulus material and augment them with realistic salient pictograms of hazardous and non-hazardous events. With this, we focus on the domain-specific meaning (hazardousness) of the event rather than on the sole detection of an event (as in an oddball setting).
In the context of this paper, events are considered as hazardous when they would be dangerous (to the pedestrian or the driver), are hard to predict and, most importantly, require special attention or reaction by the (robot) driver (e.g., more defensive behavior or slowing down). As an example, a child or a pedestrian walking on the sidewalk would be considered as non-hazardous whereas a child running from occlusion onto the street (c.f, Figure 1) would be hazardous. While we focus on hazardousness here, we view it merely as an example for a high-level semantic scene information.

RELATED WORK

Substantial prior work addresses scene understanding for intelligent vehicles in the presence of humans (c.f., the survey by Ohn-Bar and Trivedi [1]). As a relevant example, Møgelmose et al. [3] present an integrated approach on pedestrian detection, tracking and hazard inference. For the latter, they leverage map data (proximity to street) to assign hazardousness to pedestrians. However, as also discussed in the following section, the mere presence of humans in the vicinity of the car does not necessarily imply a hazard.
Utilizing BCIs in the context of human-machine interaction, substantial previous work has been performed on decoding user state from brain signals for improved user experience or performance. For example, workload or drowsiness can be detected from EEG in different task settings [4], [5] and can be used to adapt tasks based on the decoded user state [6].
At the intersection of BCI research and, both simulated and real, driving, several works have addressed the utility of brain responses for human-machine interaction. Haufe et al. [7] investigated using brain signals in early detection of emergency braking. They report that using event-related potentials for detection of braking signals is feasible both in simulation and real-world driving, whereas oscillatory signals do not provide complementary information. Khaliliardali et al. [8] focused on the anticipation and prediction of the type of driver's actions in an automotive go/no-go paradigm. Zhang et al. [9] investigated the response to directional cues presented by driving assistant systems and classified whether these correspond to the user's intention based on error-related brain activity.
Whereas the subject's desired reaction to a stimulus is immediately clear in the first two studies or only requires a comparison with a street sign in the third, in this paper we consider a more unconstrained stimulus setting in which both the context and partly the movement of stimuli is relevant for the class assignment of an event as a step towards more ecological validity [10].

Figure 2: Exemplary events from the stimulus material. The top row consists of events that have been labeled as hazardous, whereas events in the center and bottom row are labeled as non-hazardous. Only half of the actual width of the video frame is displayed for layout purposes.

MATERIALS AND METHODS

Five healthy subjects participated in the study by watching natural video sequences of traffic scenes. All subjects gave their written informed consent and the study has been approved by the Ethics Committee of the University Medical Center Freiburg.

Experimental Design: The stimulus material consists of video sequences based on actual car recordings from the KITTI dataset [11] with a resolution of 1242x375 px. Parts of the sequences were edited by inserting events with pictograms in order to introduce potential hazards. A selection of exemplary events is depicted in Figure 2. The pictograms introduced in the natural scenes are generally salient and easily discoverable. However, a substantial portion of events consists of pictograms appearing from occlusion (both with or without prior appearances in the scene). The appearance of a pictogram from occlusion does not automatically imply that it is a hazard, which needs to be inferred from the context instead.
Different types of pictograms (such as children, pedestrians, cyclists) and different colors are used. However, the type of pictogram or color does also *not* imply the class label, i.e., hazardousness of the event (c.f., the color distribution by event class in Figure 3). Similarly, events in close proximity to the car can be both hazardous or non-hazardous (e.g., a child running close to the curb compared to a pedestrian with a dog in Figure 2).
In total, each subject watched 240 scenes (videos) of 20 s each. The total of 240 scenes is grouped into blocks of 12 scenes. Each block is balanced between scenes in simple (e.g., highways) and complex (e.g., residential) environments.

 DOI: 10.3217/978-3-85125-533-1-45

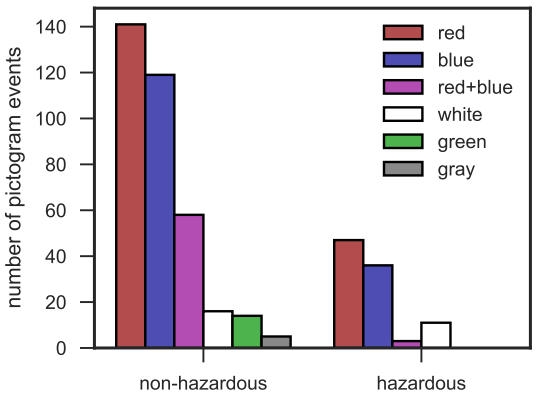

Figure 3: Event counts from a full experiment session, grouped by stimulus color and class label. The "red+blue" group consists of stimuli consisting of multiple participants (e.g., mother and child). It is apparent that the assignment of an event to a class cannot be performed solely based on stimulus type or stimulus color.

Embedded in these scenes are 262 unique events, out of which a portion was repeated over the course of the experiment, resulting in a total of 450 events in an experiment session. Out of all events in an experiment session, 97 are labeled as hazardous and 353 as non-hazardous.

During the experiment, subjects were seated approximately 80 cm in front of a 24 inch monitor, where videos were presented at 10 frames per second (corresponding to the recording rate of the source material). Subjects were instructed that they should assume being passengers in an autonomous vehicle and that they should press a button (in their right hand) in case of hazardous situations. Pressing the button could be seen as relaying the desire to drive more defensively to the vehicle. During the experiment, however, the videos continued normally, disregarding the button press.

EEG signals were recorded from 63 passive Ag/AgCl electrodes (EasyCap), which were positioned according to the extended 10-20 system and referenced against the nose. Impedances were kept below 20 kΩ. The signals were registered by multichannel EEG amplifiers (BrainAmp DC, Brain Products) at a sampling rate of 1 kHz.

Data Analysis: The recorded data was analyzed offline. It has been bandpass-filtered from 1.1 Hz to 15 Hz and downsampled to 100 Hz. Subsequently, the continuous recording has been divided into one segment per event. Each segment has a duration of 1200 ms, consisting of 200 ms preceding the first frame containing the pictogram and 1000 ms succeeding it. Note however that due to occlusions the pictogram is often not yet fully visible in the first frame of its appearance.

Before subsequent processing steps, channels whose variance was smaller than 0.5 for more than 10 % of the segments were rejected.

Additionally, segments that violated either a min-max threshold of 70 µV at frontal electrodes or whose variance was excessively large were rejected as artifactual. For base correction, the mean amplitude of the first 200 ms (corresponding to the duration of the two video frames preceding the pictogram) is subtracted from the signal.

Each segment was labeled as hazardous or non-hazardous (c.f., Figure 2). We use annotated class labels instead of using the behavioral button response of subjects to have constant class distributions and therefore better comparability across subjects.

As features for classification, mean voltages in 100 ms windows from 100 ms to 900 ms after the first visible pictogram frame were used. Both single time intervals and cumulative time intervals (i.e., concatenating the channel means of the interval with all preceding ones) were used as feature vectors (see Figure 5 for the used intervals).

Analyzing each subject individually, we train and evaluate classifiers in a chronological 5-fold cross-validation. Classification was performed by regularized linear discriminant analysis (with analytic determination of the shrinkage parameter). Classification results are reported as the area under the receiver operating characteristic (AUC). Assigning predictions at random would result in a chance-level AUC of 0.5.

RESULTS

Participants gave qualitative feedback that the events labeled as hazardous were perceived as such, and subjects pressed the button in 74 % of hazardous events. Based on the rejection policies described in the previous section, 12 % of all epochs were rejected. Rejection rates were similar for both classes such that the original class distribution was preserved.

We observed event-related responses to both hazardous and non-hazardous events with peak amplitudes around 600 ms after the first (partial) appearance of a potentially hazardous stimulus. Spatially, we observed a predominantly non-lateralized response throughout the subjects. Central and parietal electrodes offer highest discriminative information between hazardous and non-hazardous events for four of the five subjects.

Figure 4 visualizes data of a subject with average decoding results. Both hazardous and non-hazardous events show event-related potentials compared to baseline intervals taken from video segments at least 5 s from each annotated event. However, hazardous events elicit a stronger response. This is most prominent from 500 ms to 800 ms after the event's first video frame, as depicted by the color bars representing the channel-wise discriminatory information between hazardous and non-hazardous events.

Using eight time intervals within the range of 100 ms to 900 ms after the first visible frame of the event, classification yielded a mean AUC of 0.79 over all subjects, with a minimum AUC of 0.75 and a maximum of 0.87 across the five subjects.

 DOI: 10.3217/978-3-85125-533-1-45

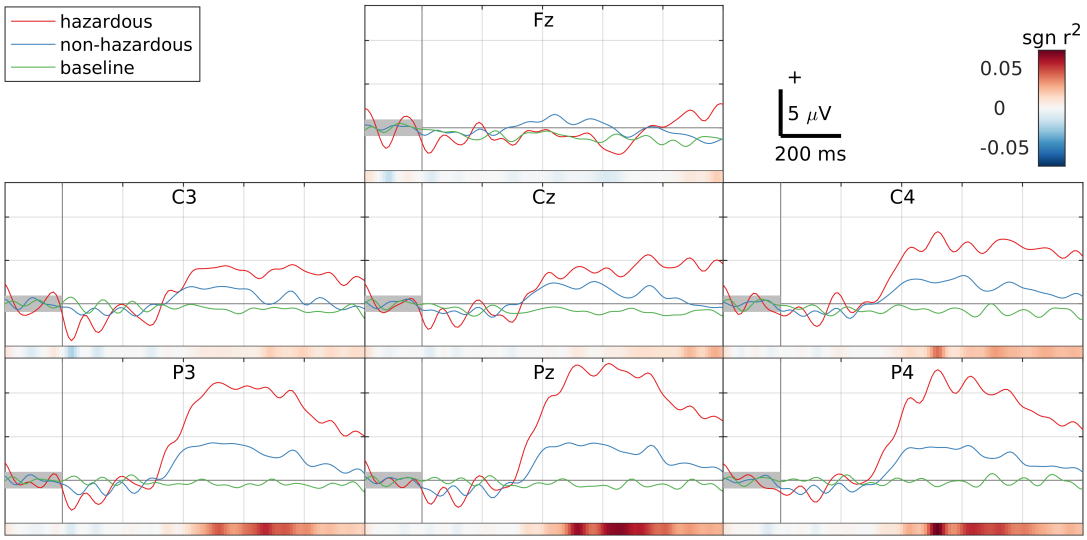

Figure 4: Event-related responses of an exemplary subject (with average classification quality) at seven electrodes. Lines show the mean voltage at the given electrode over all hazardous, non-hazardous events or baseline segments, respectively ($N = 79, 297, 315$). Baseline segments are extracted from parts of scenes that are separated by at least 5 s from other events. The mean of the first 200 ms of the interval (i.e., before the appearance of the pictogram) has been subtracted from each channel. The colorbars depict the signed r^2 value between hazardous and non-hazardous events.

Focusing on the features' influence on single-trial decoding quality (see Figure 5), one can see that classifiers trained on features from 300 ms to 400 ms after onset show the first reasonable performance (AUC of 0.66). The best single time windows have latencies of 500 ms to 600 ms and 600 ms to 700 ms after onset. Cumulatively using all time windows up to a given point, we see gains until 800 ms after the event.

DISCUSSION

We find that distinguishing between hazardous and non-hazardous events is possible with a reasonably good quality for all five subjects (minimum AUC of 0.75). Although results need to be supported by more subjects, the observation that discriminative information from the first 600 ms already result in a mean AUC of 0.76 suggests the possibility for close-to-realtime utilization as an information source in online systems.

For this study we did not perform exhaustive feature engineering or hyperparameter optimization, but rather focused on obtaining a realistic decoding result using "best practice" methods in order to evaluate the feasibility of distinguishing between high-level event classes. It appears reasonable to expect that better decoding results are possible by, e.g., adapting time intervals or spatial filters to individual subjects.

On a cursory glance, the question might arise whether the results just resemble a "classical" P300 effect in an oddball scenario. We argue that this is not the case since we aim to distinguish between two different types of events (hazardous and non-hazardous) which are both similarly (un)expected. Observed effects are not based on the sole occurrence of an event compared to a baseline stimulus.

Despite the smaller amount of hazardous events compared to non-hazardous ones (since traffic scenes should maintain some degree of realism), both classes are "odd" events that differ strongly from regular parts of the scenes (also in their brain response, as depicted by the baseline class in Figure 4). Hence, rather than discriminating rare *unexpected* events from regular ones, our classes distinguish between the *contextual meaning* of an event. Furthermore, due to the priming of participants (e.g., by pedestrians appearing before occlusions) and some repetitions in the later course of the experiments, we argue that the sole occurrence of (both classes of) events is not always unexpected.

The comparatively high latency of the event-related response could be attributed to the fact that stimuli are still partly occluded at time 0 s. Additionally, it has to be noted that the decision whether a pictogram is hazardous could not always be made immediately at its appearance since movement with respect to the scenes is critical for judging the event. Alongside the different latencies after which participants noticed the pictograms, these differences between time-alignment of events could potentially be mitigated by relying on fixation-related potentials [12]. Nevertheless, these latencies are common to both classes so we expect only minor influence upon the quality of classification.

Since subjects performed button presses during the experiment, the question arises whether the decoding results might be solely based on the motor activity of the subject. However, this appears to not be the case since there is not a clear lateralization of the response as would be expected from a single-handed motor activity. Additionally, artifact rejection should diminish the effects of muscular artifacts in the analyzed signal.

 DOI: 10.3217/978-3-85125-533-1-45

As the preceding paragraphs suggest, the complex set of stimuli might lead to several effects that might be considered confounders in the context of the classification. As discussed above for a selection of important candidates, we aimed to either control for these or check that they did not heavily affect the results. More importantly, we would like to stress that real-world use cases of BCIs most likely also implicate substantial confounders to the main task and in the light of ecological validity of BCI studies, these have to be dealt with in the analysis rather than solely by restricting the experiment.

While we focus on hazardous and non-hazardous events in the context of the paper, these labels should be considered as representatives of semantic classes that can easily and "intuitively" be distinguished by humans whereas it is challenging to infer them from alternative sensor data.

Regarding the applicability of the performed experiments to real-world driving, we want to discuss two major impediments of the current setup. First, the experiments are limited to video-based stimuli in a laboratory setting and events have been artificially introduced into the scenes. While this has been motivated by having repeatable experiments with material that is similar in quality, style and salience of stimuli, there is still an apparent mismatch with car recordings in real traffic. Regarding the signal quality, an automotive environment including abrupt movement is certainly a more difficult recording environment. However, such an environment can also be expected to create a much higher immersion than the laboratory setting, which might also transfer to more distinct subject reactions. Additionally, it is reasonable to expect some generalization to a car setting since the stimulus material is comparatively realistic and a transfer across similar ERP-based tasks has been shown to be feasible [13].

A second constraint of the presented analysis is the assumption of having temporal alignments for the potentially hazardous candidate events. However, in the context of autonomous driving it is reasonable to assume that additional sensor equipment (such as cameras or laser scanners) are able to detect candidate events (e.g., the appearance of dynamic obstacles or identification of pedestrians [1]).

Generally, we find that the combination of BCI-based monitoring in the context of autonomous machines promises to be especially helpful in the generation of labels from humans with high temporal resolution. During development of systems, this can be utilized to directly associate training data from other modalities with continuous human feedback, e.g., to evaluate the compliance of the machine to the human's requirements. Additionally, due to the potentially subject-specific but unbiased nature of responses, BCIs can become building blocks for adapting complex systems to individual users, e.g., by optimizing parameters based on the *perceived* hazardousness.

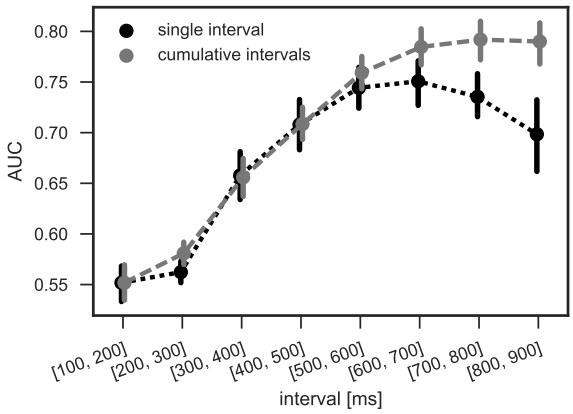

Figure 5: Classification results on different time intervals relative to the event's first visible video frame. Each point represents the mean AUC over all subjects along with a bootstrapped $68\,\%$ confidence interval. The dotted line represents classification results on the respective $100\,\mathrm{ms}$ interval whereas the dashed one shows results based on including all time preceding intervals as features.

CONCLUSION

In order to investigate the discriminability between different high-level semantic events in complex environments with passive BCIs, we describe preliminary experiments with five subjects on distinguishing hazardous and non-hazardous appearances of pictograms in natural driving videos. We find that the event-related responses differ not only compared to baseline stimuli but also between classes. Single-event classification yields a mean AUC of 0.79, suggesting that reasonable discrimination is possible in the context of complex realistic baseline stimuli.
We view these results as a step towards utilization of BCIs as a monitoring and feedback channel of human scene understanding and assessment for improved human-machine interaction.

ACKNOWLEDGMENT

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 the project ZAFH-AAL (Az: 32-7545.24-9/18/1) and for bwHPC.

REFERENCES

[1] Ohn-Bar E and Trivedi M. M. Looking at Humans in the Age of Self-Driving and Highly Automated Vehicles. *IEEE Transactions on Intelligent Vehicles*, 1(1):90–104, March 2016.

[2] Bestick A, Bajcsy R, and Dragan A. D. Implicitly Assisting Humans to Choose Good Grasps in Robot to Human Handovers. In *International Symposium on Experimental Robotics (ISER)*, Tokyo, Japan, 2016.

 DOI: 10.3217/978-3-85125-533-1-45

[3] Møgelmose A, Trivedi M. M, and Moeslund T. B. Trajectory analysis and prediction for improved pedestrian safety: Integrated framework and evaluations. In *2015 IEEE Intelligent Vehicles Symposium (IV)*, pages 330–335, June 2015. 00000.

[4] Brouwer A.-M, Hogervorst M. A, van Erp J. B. F, Heffelaar T, Zimmerman P. H, and Oostenveld R. Estimating workload using EEG spectral power and ERPs in the n-back task. *J Neural Eng*, 9(4):045008, August 2012.

[5] Borghini G, Astolfi L, Vecchiato G, Mattia D, and Babiloni F. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews*, 44:58–75, July 2014.

[6] Schultze-Kraft M, Dähne S, Gugler M, Curio G, and Blankertz B. Unsupervised classification of operator workload from brain signals. *J Neural Eng*, 13(3):036008, 2016.

[7] Haufe S, Kim J.-W, Kim I.-H, Sonnleitner A, Schrauf M, Curio G, and Blankertz B. Electrophysiology-based detection of emergency braking intention in real-world driving. *J Neural Eng*, 11(5):056011, October 2014.

[8] Khaliliardali Z, Chavarriaga R, Gheorghe L. A, and Millán J. d. R. Action prediction based on anticipatory brain potentials during simulated driving. *J Neural Eng*, 12(6):066006, 2015.

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

[10] Brouwer A.-M, Zander T. O, van Erp J. B. F, Korteling J. E, and Bronkhorst A. W. Using neurophysiological signals that reflect cognitive or affective state: Six recommendations to avoid common pitfalls. *Frontiers in Neuroscience*, 9, April 2015. 00005.

[11] Geiger A, Lenz P, Stiller C, and Urtasun R. Vision meets Robotics: The KITTI Dataset. *The International Journal of Robotics Research*, page 0278364913491297, August 2013.

[12] Finke A, Essig K, Marchioro G, and Ritter H. Toward FRP-Based Brain-Machine Interfaces—Single-Trial Classification of Fixation-Related Potentials. *PLOS ONE*, 11(1):e0146848, January 2016.

[13] Wenzel M. A, Almeida I, and Blankertz B. Is Neural Activity Detected by ERP-Based Brain-Computer Interfaces Task Specific? *PLOS ONE*, 11(10):e0165556, October 2016.