Towards Imbalanced Multiclass Driver Distraction Identification

Kapotaksha Das
University of Michigan
Dearborn, Michigan
takposha@umich.edu

Mohamed Abouelenien
University of Michigan
Dearborn, Michigan
zmohamed@umich.edu

Mihai Burzo
University of Michigan
Flint, Michigan
mburzo@umich.edu

Rada Mihalcea
University of Michigan
Ann Arbor, Michigan
mihalcea@umich.edu

Abstract

Driver's distraction is one of the leading causes of driving-related accidents worldwide. The ability to detect driver's distraction preemptively is crucial to reducing the number of such accidents. This paper utilizes a novel multimodal dataset of thermal, visual, near-infrared, and physiological signals recorded from 45 subjects in order to identify distraction. We explore imbalanced distraction identification as a four-class problem across different types of distractions in order to resemble real-life scenarios, where the occurrence of different distractors varies. Our study analyzes the effectiveness of using early fusion across different modalities, a variety of window sizes, and data balancing schemas using synthetic instances. Moreover, we explore the effects of introducing subject-specific knowledge when training such identification models.

Introduction

The Center for Disease Control and Prevention (CDC) reported over 3,000 deaths in 2019 alone that were caused from accidents involving distracted driving [NHT, 2020]. A study conducted by Zendrive reported that despite a reduction in traffic due to the COVID-19 pandemic in 2020, there was a 63% increase in the number of collisions per million miles [Zen, 2021]. Furthermore, 57% of these collisions were caused due to phone-based distractions, which was also believed to be an under-reported statistic. The National Safety Council notes that over half of the US states have no provisions for local authorities to report phone-based triggers in crash reports, further highlighting the underestimated dangers associated with distracted driving [NSC, 2016]. Zendrive also reported that over the pandemic there was a 36% increase in rapid acceleration events, indicating stressors such aggression also play a role in distracted driving and associated accidents. All of these findings indicate the need for timely detection of distracted driving, as well as highlight the significance of identifying the nature of the distraction. This will allow us to better understand how different distractors relate to accidents in order to propose proper and specific solutions.

Previous work in this field has focused on unimodal distraction detection, with works of Wang et al. and Mao et al. [Wang et al., 2021; Mao, Zhang, and Liang, 2020] using the visual modality to detect driver's distraction using deep learning networks. However, the models proposed were limited to only the visual modality and could only detect a distraction versus an undistracted state, being unable to detect the type of distractor affecting the driver. Using other modalities, such as wearable devices [Jiang et al., 2018] or driver posture [Abouelnaga, Eraqi, and Moustafa, 2017] have also been explored by using the driver's movement to identify physical distraction. Multimodal approaches using vehicular driving data alongside physiological signals [Taamneh et al., 2017] or facial expressions and landmarks [Du et al., 2018] have also been researched to develop more robust distraction detectors. However, this past research did not focus on identifying the type of distractors used, as well as how they differ in affecting the drivers. Research has mainly focused on a binary classification problem regarding whether a driver is in a state of distraction.

Our past work explored the task of driver distraction detection, where we specified whether a driver was distracted [Das et al., 2021] using a novel cascaded late fusion multimodal system. Our ongoing work in this paper aims to further this novel work and introduce increased granularity in detection and make it possible to identify the type of distraction the driver is being subjected to using a multimodal approach, taking into consideration that different distractors might occur with different frequencies, which allows for monitoring and guidance systems that can develop better targeted responses to warn drivers about their behavior preemptively.

Dataset Description

In our experiments, we use a novel multimodal dataset collected from 45 subjects consisting of thermal, visual, near-infrared (NIR) and physiological signals. Each subject participated in two recording sessions, one in the morning (before 11 AM) and another in the evening (post 4 PM), with each session lasting around 45 minutes, while subjected to a simulated driving experience. The session consisted of a baseline phase with no driving, a 'freedriving' phase where the subject drove without any distractions, and a 'distractors' phase where the subject had to drive while responding to distractions. We used four different distractors for our experiments to emulate common distractions the drivers usually face, which are:

Physical - The subjects typed and sent out a predefined 8 word text message using a mobile device mounted next to the driver,

Cognition - The subjects took an N-back test that would challenge short term memory by determining whether a letter matched earlier letters using a prerecorded letter sequence. In our experiments N was 1,

Emotional - The subjects listened to a preselected proactive audio clipping of some news and then shared their opinions regarding that clip,

Frustration - The subjects verbally interacted with an intentionally misleading GPS system in an attempt to locate a specific destination, without being made aware that the misguidance was part of the experiment.

Furthermore, the order of the distractors and the content used for each of them were randomized for each subject to prevent any possible bias.

To capture multimodal data for these recordings, a camera and sensor suite were used as follows: Thermal - A FLIR SC6700 thermal camera capable of capturing thermal video at 100 FPS,

Visual - An IDS RGB camera capable of capturing RGB video at 20 FPS,

NIR - An IDS NIR camera capable of capturing NIR video at 20 FPS,

Physiological - Four sensors provided by Thought Technology to measure Blood Volume Pulse (BVP), Respiration Rate, Skin Temperature, and Skin Conductance at a 2048 Hz sampling rate.

After recording the data, each modality was processed for feature extraction. Due to various human factors as well as the nature of the four distractors, they did not have balanced recording times. Table 1 shows the distribution of the total recording lengths across the distractors.

Distractor	Total Length (in minutes)
Physical	$\sim$143
Cognition	$\sim$76
Emotion	$\sim$134
Frustation	$\sim$108

Table 1: Total Recording Lengths for each Distractor

Feature Processing and Extraction

For the thermal modality, each frame was split into five regions of interest (ROI); the face as a whole, the forehead, the eyes, the cheeks, and the nose. These ROIs were tracked across all the frames of a recording allowing us to monitor changes in temperature in the specific ROIs, even as the subjects moved their head through the course of the recording. The pixels captured in the ROIs represented temperature in Fahrenheit, which were used to extract a 20-bin histogram representing the temperature distribution for a given ROI for a frame. Four additional features, namely, the mean pixel value (mean temperature in the ROI), max pixel value (highest temperature seen), min pixel value (lowest temperature seen), and pixel range (temperature range) were also computed to give us a total of 24 features for each ROI. Across five ROIs, this gave us a total of 120 features extracted from the thermal modality.

The visual and NIR modalities followed the same processing steps for their recordings. We used the OpenFace image processing library to assist us in feature extraction. By using information, such as facial landmarks to approximate 3D positioning of the head and its pose, as well as the location of the eyeball sphere and the pupil center to track the gaze, and by using deviations in facial features and landmarks against those of a neutral expression to define and group Action Units (AU) by intensity, we created a feature vector for each frame consisting of spatial and behavioral information for the subject over the course of the recording. In total, 709 features were extracted for each of the visual and NIR modalities.

We used the Biograph Infiti software for feature extraction of the physiological modality. This involved the computation of statistical features across the four sensors signals that were collected. A set of 49 features was extracted from the BVP sensor, consisting of basic time domain statistical features, Normal to Normal heartbeat interval (NN) based features, and frequency domain statistical features across three spectral power frequency bands (very low, low and high) to give us a detailed insight across all aspects of the BVP sensor channel. For respiration rate, skin temperature, and skin conductance, six standard temporal statistical features for each sensor channel were extracted. Finally, four features that described the BVP and respiration rate with respect to each other statistically were also computed. In total, 73 features were extracted for the physiological modality.

With the feature extraction process completed, the feature vectors for each subject were normalized using the baseline from their corresponding evening baseline recording across all modalities. The data was then segmented using windows of fixed length across all our modalities. We used four window sizes: 2-seconds. 4-seconds, 8-seconds, and 16-seconds for our experiments in order to identify the best performing window size in identifying distractions. For each segment, the mean of each feature was computed to provide a single feature vector per segment per modality. This allowed for modalities with different framerates to be integrated with each other in each segment, enabling us to implement early fusion. By using early fusion, we were able to use data across multiple modalities to form truly multimodal feature vectors that were used to train our classifiers.

Additional details of the dataset collection and feature extraction methodologies can be seen in Das et al. [Das et al., 2021].

Experimental Setup

In this paper, we looked at a 4-class distractor identification problem for all of our experiments, where a model would have to correctly identify the type of distractor amongst the four that were used during the recording sessions, as discussed previously in the Dataset Description section. We used our multimodal dataset with the Extreme Gradient Boosted (XGB) classifier [Chen and Guestrin, 2016] for

SMOTE Balancing	Accuracy	Mean Recall	Mean F1
Balancing Applied	0.770	0.782	0.761
No Data Augmentation	0.771	0.783	0.762

Table 2: Average Performance Metrics with respect to usage of SMOTE

each experiment. This classifier was selected experimentally.

In order to explore the effects of subject-specific intuition when identifying distractions, we utilized two kinds of train-test splits in our experiments, as follows:

Leave One Subject Out (LOSO) Split - Here, each test fold consisted of all recordings from one subject only. This would ensure that the classifier had no prior information about the test subject in any training fold.

Global Subject Split - Here, each recording was split into training and testing batches by a given fraction. However, no identifying information about the subject was provided in testing, only the features pertaining to the subject were provided to prevent any identification bias.

For the Global Subject Split, we tested across three split fractions, a 25-75 train-test split, a 50-50 train-test split and an 80-20 train-test split. These splits would let us hypothesize on whether the introduction of subject-specific features would allow for better classification, and how the fraction of subject-specific data available during training might affect the performance.

We also analyzed whether the imbalanced nature of the dataset classes were detrimental to classification and whether this could be mitigated. Hence, for half of the experiments, we applied the borderline Synthetic Minority Oversampling Technique (SMOTE) to the dataset prior to training. SMOTE oversamples an imbalanced dataset with synthetically generated data that balances the classes in the training data. This in certain circumstances helps models to provide better classifications that are not biased towards the majority classes. We did not apply any form of data augmentation to the other half of our experiments to act as a form of control, as well as to verify the need of any augmentation.

In summary, we analyzed the performance of distraction identification across four window sizes, four data splits, and with as well as without data augmentation for an early fusion-based multimodal dataset using an XGB classifier for a total of 32 experiments.

Results & Analysis

To observe the effects of applying SMOTE data augmentation for balancing the dataset, we measured the average F1 score for all experiments with and without SMOTE, as seen in Table 2. We can see that there is no significant difference in the performance metrics on applying SMOTE. This is likely due to the utilization of segmentation to generate a large number of samples for each class, which greatly helps in offsetting any negative impacts that would occur from an imbalanced dataset during classification. It also might indicate that strong inter-class differences exist between the distractors, something we also observe in the upcoming results.

Next, we can see how window sizes affect the performance, with the results tabulated in Table 3 averaged across the other parameters, where the bold text highlights the best seen metrics for a given column. Here, we see that the 8-second window is the best performing window size with a mean F1 score of 78.9%. Moreover, we observe that there is no linear correlation between the windows sizes and the mean F1 score, or any of the class-specific recalls as well. In fact, the class for the emotion distractor has the highest recall of 63% seen when using a 2-second window. This can imply that different distractors might be easier to identify using different window sizes. For instance, smaller windows sizes seem to be sufficient to detect emotional and frustration distractors.

Finally, we look at how splitting the data affects the performance in order to provide subject-specific intuition. The results are shown in Table 4, with the results averaged across the other parameters, where the bold text highlights the best seen metrics for a given column. These results indicate that while having subject-specific data in training is beneficial to performance, a significant amount of training data, in the ratio of 80 training - 20 testing, is required to outperform the LOSO split type. This is however an incomplete picture, as when we look at the class recalls we see that the physical and emotion distractors perform significantly better in the LOSO split type, while the cognition and frustration distractors not only perform the worst in LOSO, but also improve in performance in the Global split as the available subject-specific segments used in training increase. This potentially indicates that global behavioral drivers' patterns can be observed for the physical and emotional distractors, which can be easily identified even with no prior information about a new test driver. On the other hand, cognition and frustration-based responses are more driver-specific and, therefore, benefit from prior information. Identifying distractors in general benefits from adding more samples, except for the frustration distractor, which can easily be detected with a smaller amount of training data. This is in line with our observation that a 2-second window is satisfactory to detect this specific distractor.

Fig. 1 plots a confusion matrix obtained for the best performing set of parameters, which is using an 8-second window, with an 80-20 Global data split without any data augmentation. Using these parameters, we achieve a mean F1 score of 91%. It can be seen that even with the cognition distractor having the least number of samples amongst all the four distractors, it still performs well with a recall of 92%. However, the emotion distractor consistently is the weakest performing class across all of our experiments, despite having the second highest number of samples available in that class. This could indicate that patterns in the responses of certain distractors are more apparent than others regardless of the number of samples.

Window Size	Accuracy	Mean Recall	Mean F1	Physical - Recall	Cognition - Recall	Emotion - Recall	Frustration - Recall
2 seconds	0.759	0.767	0.750	0.750	0.747	**0.630**	0.941
4 seconds	0.753	0.768	0.746	0.715	0.805	0.607	**0.947**
8 seconds	**0.798**	**0.809**	**0.789**	**0.836**	**0.843**	0.613	0.944
16 seconds	0.771	0.787	0.759	0.816	0.835	0.550	0.946

Table 3: Average Performance Metrics with respect to Window Sizes

Split Type	Accuracy	Mean Recall	Mean F1	Physical - Recall	Cognition - Recall	Emotion - Recall	Frustration - Recall
Global 25-75	0.684	0.709	0.675	0.575	0.766	0.515	0.981
Global 50-50	0.743	0.768	0.734	0.751	0.862	0.477	0.982
Global 80-20	**0.850**	**0.863**	**0.844**	0.888	**0.913**	0.662	**0.987**
LOSO	0.804	0.791	0.792	**0.902**	0.689	**0.746**	0.826

Table 4: Average Performance Metrics with respect to Data Splits

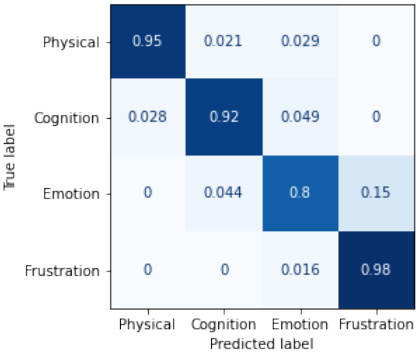

Figure 1: Confusion Matrix using the optimal parameters

Conclusion

In this paper, we used a novel mutlimodal dataset consisting of thermal, visual, NIR and physiological signals to identify distracted driving across four distractor types: physical, cognition, emotion and frustration in order to allow for a better understanding of how distractions affect drivers. In our experiments, we analyzed the effects of window size (for data sampling), as well as imbalance in our data. Our experiments showed that the 8-second window overall was the best window size for distraction identification. However, depending on the type of distractor, smaller window sizes might be useful as well, as can be seen with the emotional and frustration distractors. Having subject-specific data in training also proved to be beneficial, however, it required a high fraction of the subject's data to be available for optimal performance. We also observed that the imbalanced classes had no significant effect on performance, if a large number of data points for all distractors were available for training.

Acknowledgements

This content is based in part on research funded by the Toyota Research Institute ("TRI"). Any opinions, observations, assumptions, or recommendations shared in this content are the authors' only and do not necessarily represent the views of TRI or any other Toyota entity.

References

Abouelnaga, Y.; Eraqi, H. M.; and Moustafa, M. N. 2017. Real-time distracted driver posture classification. *arXiv preprint arXiv:1706.09498*.

Chen, T., and Guestrin, C. 2016. Xgboost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Das, K.; Sharak, S.; Riani, K.; Abouelenien, M.; Burzo, M.; and Papakostas, M. 2021. *Multimodal Detection of Drivers Drowsiness and Distraction*. New York, NY, USA: Association for Computing Machinery. 416–424.

Du, Y.; Raman, C.; Black, A. W.; Morency, L.; and Eskénazi, M. 2018. Multimodal polynomial fusion for detecting driver distraction. *CoRR* abs/1810.10565.

Jiang, L.; Lin, X.; Liu, X.; Bi, C.; and Xing, G. 2018. Safedrive: Detecting distracted driving behaviors using wrist-worn devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1(4).

Mao, P.; Zhang, K.; and Liang, D. 2020. Driver distraction behavior detection method based on deep learning. *IOP Conference Series: Materials Science and Engineering* 782:022012.

2020. Overview of motor vehicle crashes in 2019. Accessed: 2022-01-19.

2016. Undercounted is underinvested: How incomplete crash reports impact efforts to save lives. Accessed: 2022-01-19.

Taamneh, S.; Tsiamyrtzis, P.; Dcosta, M.; Buddharaju, P.; Khatri, A.; Manser, M.; Ferris, T.; Wunderlich, R.; and Pavlidis, I. 2017. A multimodal dataset for various forms of distracted driving. *Scientific Data* 4(1):170110.

Wang, J.; Wu, Z.; Li, F.; and Zhang, J. 2021. A data augmentation approach to distracted driving detection. *Future Internet* 13(1).

2021. Zendrive collsion report. Accessed: 2022-01-19.