Technical Design Space Analysis for Unobtrusive Driver Emotion Assessment Using Multi-Domain Context

DAVID BETHGE*, Porsche AG, LMU Munich, Germany
LUIS FALCONERI COELHO*, Porsche AG, CODE University, Germany
THOMAS KOSCH, HU Berlin, Germany
SATIYABOOSHAN MURUGABOOPATHY, Porsche AG, Germany
ULRICH VON ZADOW, CODE University, Germany
ALBRECHT SCHMIDT, LMU Munich, Germany
TOBIAS GROSSE-PUPPENDAHL, Porsche AG, Germany

Driver emotions play a vital role in driving safety and performance. Consequently, regulating driver emotions through empathic interfaces have been investigated thoroughly. However, the prerequisite - driver emotion sensing - is a challenging endeavor: Body-worn physiological sensors are intrusive, while facial and speech recognition only capture overt emotions. In a user study (N=27), we investigate how emotions can be unobtrusively predicted by analyzing a rich set of contextual features captured by a smartphone, including road and traffic conditions, visual scene analysis, audio, weather information, and car speed. We derive a technical design space to inform practitioners and researchers about the most indicative sensing modalities, the corresponding impact on users' privacy, and the computational cost associated with processing this data. Our analysis shows that contextual emotion recognition is significantly more robust than facial recognition, leading to an overall improvement of 7% using a leave-one-participant-out cross-validation.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing**; • **Computing methodologies** → *Machine learning*.

Additional Key Words and Phrases: Emotion Sensing, Affective Computing, Remote Sensors, Automotive, Empathic Interfaces, In-the-Wild Analysis

ACM Reference Format:
David Bethge, Luis Falconeri Coelho, Thomas Kosch, Satiyabooshan Murugaboopathy, Ulrich von Zadow, Albrecht Schmidt, and Tobias Grosse-Puppendahl. 2022. Technical Design Space Analysis for Unobtrusive Driver Emotion Assessment Using Multi-Domain Context. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 159 (December 2022), 30 pages. https://doi.org/10.1145/3569466

*Both authors contributed equally to the paper

Authors' addresses: David Bethge, Porsche AG, LMU Munich, Stuttgart, Germany, david.bethge@ifi.lmu.de; Luis Falconeri Coelho, Porsche AG, CODE University, Berlin, Germany, luis.coelho@code.berlin; Thomas Kosch, HU Berlin, Berlin, Germany; Satiyabooshan Murugaboopathy, Porsche AG, Stuttgart, Germany; Ulrich von Zadow, CODE University, Berlin, Germany; Albrecht Schmidt, LMU Munich, Munich, Germany; Tobias Grosse-Puppendahl, Porsche AG, Stuttgart, Germany.

2474-9567/2022/12-ART159 $15.00
https://doi.org/10.1145/3569466

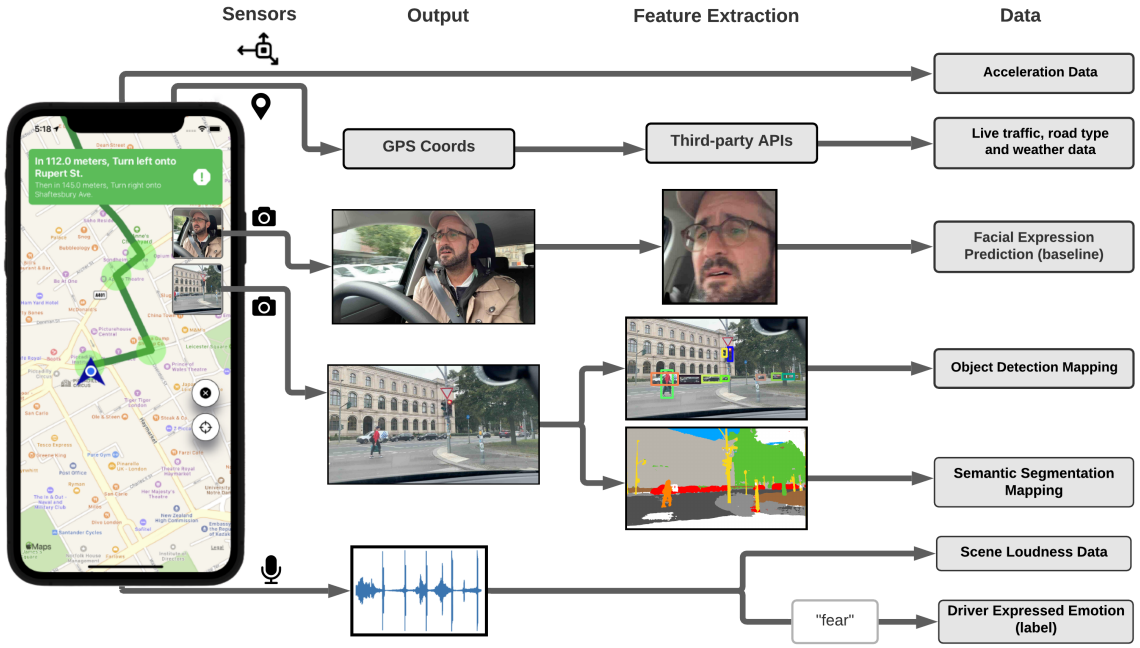


Fig. 1. Multi-Domain Context Sensor Information used for the technical design space analysis. We predict emotions using a smartphone app, which employs five different sensors: accelerometer, GPS, front-facing camera, back-facing camera, and microphone. We use the latitude and longitude output by the GPS sensor to fetch data on live traffic, road type, and weather from third-party APIs. The front-facing camera captures the driver's face to perform facial expression recognition (used as a baseline). Visual scene segmentation and object detection are performed on the back-facing camera input.

1 INTRODUCTION

Emotions considerably impact drivers' performance, safety, and health [51]. Aggressive driving, for instance, plays a significant role in most fatal highway collisions each year [19, 45], leading to more severe injuries and fatalities [14]. Statistics show that more than 90% of traffic accidents can be attributed to human errors [57]. Negative emotional states while driving are associated with "poorer physical and mental health and quality of life", leading to an overall deteriorating driving performance [24].

Even sadness can seriously increase driving errors and decrease driving efficiency [25]. Consequently, the design, implementation, and evaluation of so-called *empathic car interfaces* has been the subject of previous research [8, 22, 62]. Empathic car interfaces aim to regulate driver emotions, thus improving the driving experience and reducing the risk of accidents. However, the unobtrusive assessment of driver emotions remains an open challenge.

Facial expression recognition, a frequently used method to detect emotions in driving contexts, often performs poorly [23, 34]. It requires subjects to overtly express their emotion through their facial muscles, failing to detect covert affective states. [6, 30]. As an alternative, past research used physiological sensing as a real-time measure to estimate driver emotions [22, 62]. Although this provides accurate assessments, it requires body-worn physiological sensors that reduce user acceptance. To address these shortcomings, contextual and behavioral driving data analysis emerged as an unobtrusive alternative to detect driver emotions [6, 26, 36, 44, 46]. Previous work in this area focused on a limited set of features frequently requiring internal car data access. Furthermore,

it did not approach the topic of how different sensing modalities contribute to the classification performance of emotions.

Contrary to previous approaches that require body-mounted sensors [42], access to internal car information [36], or rely on a limited set of contextual features [6]; this paper investigates how emotions can be predicted by analyzing a rich set of contextual features unobtrusively captured by a smartphone, including audio-visual data. Furthermore, we derive a technical design space analysis to inform practitioners and researchers about the most indicative sensing modalities, their advantages and drawbacks.

We collect contextual and audio-visual data in an in-the-wild study with 27participants. The collected data comprises different context domains: weather, traffic, road type, and motion, including speed and acceleration. We record in-cabin audio and video, with the front-facing camera recording the driver and the back-facing camera recording the road view. We annotate the data using the participants' self-assessed emotional states. We compare our classification approach against emotion classification through facial expression as a baseline. A Random Forest classification using all features yields a classification accuracy of 59% (F_1: 0.45), outperforming facial expression classification by 7% and contextual classification by 13%. Finally, we present a technical framework showing how contextual and audio-visual sensing modalities influence the accuracy of emotion classification. Our work discusses how designers can select sensing strategies to prototype empathic car interfaces considering trade-offs related to computational cost and privacy concerns.

CONTRIBUTION STATEMENT

The contribution of this paper is threefold:

C1: An analysis of the technical design space for empathic car interfaces using a rich set of sensor streams from smartphones.
C2: A smartphone system and extensive data collection from in-the-wild driving evaluating contextual and audio-visual driving data for ubiquitous driver emotion assessments.
C3: Guidelines and considerations for application developers taking specific features for computational costs and privacy into account.

2 RELATED WORK

This section outlines the current understanding of emotions, how they are affected in a driving context, current emotion assessment practices, and driver emotion regulation methods.

2.1 Understanding Emotions while Driving

Emotional states can be schematized in different ways, the two most common categories being discrete and continuous emotion representations. Discrete representations of emotions derive from the works of Paul Ekman [15] who identified six basic emotions (i.e., anger, disgust, happiness, sadness, surprise, and fear) which are universally recognizable and encodable in facial muscles. In contrast, continuous emotion representation models encode the emotional state into a continuous value spectrum. Russell's circumplex model of affect [50] is one of the most commonly adopted continuous approaches. In our work, we employ a discrete emotion categorization.

Research on driving contexts has found interesting relations between specific scenarios and their potential for eliciting emotional states. Dittrich considered the "spatial-temporal distribution of drivers' emotions and their determinants" [12]. The study found that road intersections cause considerable amounts of emotional activity in drivers. Positive emotions are more likely at the beginning and end of a ride, adding strength to the claim made in this work that drivers' emotions can be inferred from contextual information.

Hancock et al. [20] concluded that as drivers' affective states change, so do the "measures of both longitudinal and lateral control of the vehicle", indicating that different emotions correlate with different mean vehicle speeds

and the number of lane excursions. Another study [43] further examined the link between driver affect and driving styles, verifying that maladaptive driving styles, including reckless, careless, angry, hostile, and anxious, were associated with a lower capacity for emotional self-regulation. This finding is confirmed by Mesken et al. [41], which investigates the impact of driving context in eliciting certain emotions, listing anxiety, anger, and happiness as the most likely emotions to fluctuate in traffic situations.

A dynamically changing external environment can manipulate drivers' perceived workload and emotions. Faure et al. [18] showed that visually changing driving environments influence the driver's perceived cognitive workload. Frequent and unexpected changes in visual processing can change driver stress levels [52], resulting in differently perceived emotions by the driver [51].

2.2 Emotion Assessment

Detecting the driver's emotional state is essential for developing in-vehicle empathic systems to improve the driving experience. Related psychological research has shown that negative affective states can negatively impact driving performance [25], potentially causing traffic violations, driver distraction, and accidents. Therefore, previous research on in-car interventions has been designed to alleviate extreme emotional states. Braun et al. [8] show that these extreme states correspond to danger, while states with medium arousal levels and positive valence are recognized as optimal for safe driving.

Different approaches have been used to assess drivers' affective states, including physiology, facial expression, self-reports, or biosignals [4, 62]. To an extent, driving behavior and context have also been researched as an alternative assessment of emotions. Liu et al. [36] presented an emotion sensor based on CAN-BUS data and external environmental factors. In a long-term user study, they collected facial expressions, heart rate variability, CAN-BUS data, and environmental data to predict driver emotions using three classification models: CAN-BUS data only, video-only, and a fusion of both models. Their results show a participant-dependent classification accuracy of 71% and a leave-one-participant-out accuracy of 59.2% using a fusion-based model considering both video and CAN-BUS data. Our system is inspired by this approach, extending the collected video-only data by contextual driving semantics such as live traffic, weather, and audio data.

Furthermore, we evaluate the impact of different variables, including facial expression analysis as a common emotion assessment [15, 16], on the classification accuracy. Universal emotion assessments through facial expressions are disputed in previous work [23, 34], potentially requiring individual training for each person [30]. Here, we aim for a universally applicable approach using the driver's smartphone only to collect contextual and environmental data. We label the collected data using the participant's verbally self-assessed emotions.

In addition to driver context, previous research pointed out that environmental events influence driver emotions and stress levels—however, exclusively using driving context for predicting driver emotions is relatively new. Recently, Bustos et al. [10] proposed a system that recognizes driver stress levels by analyzing outside-view camera input during real-world driving conditions. The authors propose three models to predict a three-class stress level (i.e., low, medium, and high) from the image stream: (1) image classification with object presence features, (2) end-to-end image classification via a CNN, and (3) end-to-end video classification by temporal segment networks. Their results showed that the best average test accuracy of 72% was obtained using a video CNN. While their work focused on a second-person annotated stress label which should reflect the driving scene complexity [21], our work uses self-reported subjective emotion ratings. Bethge et. al [6] proposed a smartphone application that detects subjective discrete driver emotions. Their app uses GPS and third-party APIs to obtain road and traffic data representing environmental characteristics. While their approach offers a rich set of features to classify emotional states, their sensor set is constrained, containing no visual or auditory features. We added their work as a baseline for our study.

2.3 Driver Emotion Regulation

Empathic car interfaces can counteract emotion-related hazards by sensing the driver's state and intervening when potentially dangerous behavior is detected. Different mechanisms, including interventions, adaptive music [28], and lighting [9] were proposed in the literature. Such empathic car interfaces require continuous monitoring of the driver's emotional state, preferably via remote sensing, making our contribution relevant for application designers.

Summary

Previous research informs how emotions are interpreted, how they change while driving, and how they can be assessed in real-time to implement empathic car interfaces. However, they present drawbacks that may hinder the adoption of empathic car interfaces in the real world. Currently, most reliable assessments rely on body-worn sensors or are not universally applicable, e.g., by relying on internal car interfaces. Though there is a large availability of research evaluating different data streams, it remains unclear which features indicate emotions. We address this gap by collecting a rich contextual and audio-visual data set in an in-the-wild study using consumer smartphones. We analyze the indicativeness of the data streams to present a technical framework, depicting the contribution of contextual and audio-visual feature sets for the accuracy of driver emotion classifications.

3 DATA COLLECTION SYSTEM

In this section, we present a system that captures the contextual driving data from a smartphone using a combination of virtual and on-device sensor streams. We built an end-to-end data pipeline composed of a data-gathering mobile application with remote-sensing features and a data-processing pipeline. Informed by related work, we considered the following requirements: (1) in-the-wild data acquisition with a smartphone, (2) seamless integration of usable in-the-wild emotion sensing, (3) acquisition of features related to driving tasks or emotional state, (4) unobtrusive remote sensing and (5) the effect of the environment's physical characteristics (e.g., weather, road type and motion metrics) and visual complexity-related features [6, 36]. The end-user application seamlessly integrates in-the-wild contextual gathering to everyday driving tasks; hence, it features standard navigation functionality.

3.1 Mobile Application

The mobile application, written in Swift[1], allows users to enter text-based descriptions of locations to obtain turn-by-turn spoken directions. We show the application user interface in Figure 2. The app requires an internet connection and runs on iPhones [2] that have iOS 13 or higher installed.

Figure 3 presents the mobile application's application architecture. The input modalities utilized by the application as data sources are the smartphone's front-facing camera, back-facing camera, microphone, GPS sensor, and accelerometer.

3.1.1 Application Main Loop. The audio/video controller monitors the application's main loop. Its output frequency is configured to 10 frames per second (FPS) for optimal performance and constitutes the central processing trigger. When the user activates the navigation mode, session recording begins. The application starts writing a sequence of RGB images from the front-facing camera, facing the driver, and back-facing camera, directed at the road, to local storage. A journey snapshot summary JSON object is generated for each frame pair. Each summary includes the frame number reference for posterior retrieval of images and a summary of the most up-to-date sensor-merged data.

[1] https://developer.apple.com/swift

[2] 11 Pro Max, 11 Pro, 11, Xs Max, Xs, Xr, SE 2. These iPhones enable acquiring front-facing and back-facing camera input at the same time.

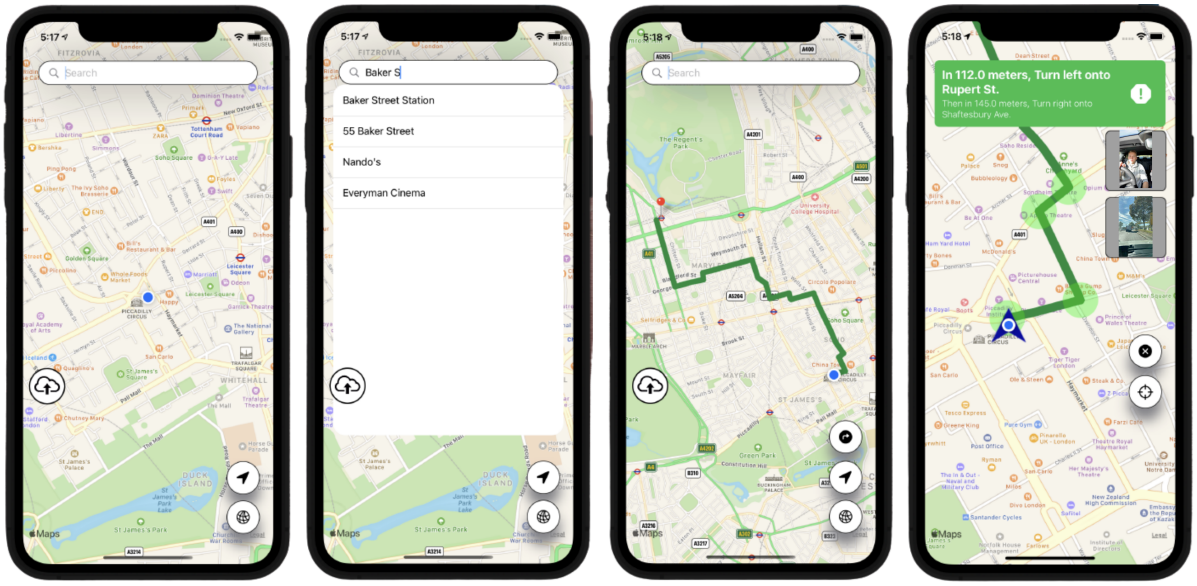


Fig. 2. Navigation User Flow. Screenshots of the mobile application demonstrate user flow from the home screen (a), followed by route search (b), route overview (c) and after navigation is activated and session recording beings (d). In (d), we also show a small preview of the front-facing and back-facing camera stream to the user, e.g., to allow for position adjustments and to be transparent about all recorded information.

3.1.2 Data Synchronization. At the end of each session, the application concatenates the 'journey snapshot' objects accumulated and adds non-sensitive user-specific information (self-identified gender, age, car model - provided on the first app launch) and session metadata, e.g., start time, end, day of the week, to produce a JSON file summarizing the whole ride (see section 5 for a more detailed description of the dataset). At this point, the app outputs the audio recording file and uploads only the JSON summary to cloud storage (we used AWS S3[3] bucket) to register the session's occurrence. This upload does not include the audio and the images. Due to the size of the file bundles (i.e., approximately 2.5 gigabytes per session), we designed the upload process to be initiated by users at their leisure, making usage more convenient and avoiding unnecessary mobile network charges. This process does not conflict with future real-time capabilities and local predictions, but rather is necessary to explore potential design decisions on a fully functional dataset of source data. For real-time predictions, we imagine that remote context data (e.g., current weather) is collected from remote services, and ML models are run locally using frameworks such as CoreML[4]. Developers of future applications will additionally have to consider the trade-off between prediction frequency and energy consumption.

3.2 Data Post-Processing Pipeline

First, the data processing pipeline ingests data from the cloud storage and checks that the session data has been uploaded and there are no corrupted files. After, we perform multiple extraction mechanism steps: (1) visual features extraction, (2) audio analysis, (3) facial expression classification, (4) road data acquisition, (5) weather

[3] https://aws.amazon.com/s3
[4] https://developer.apple.com/documentation/coreml

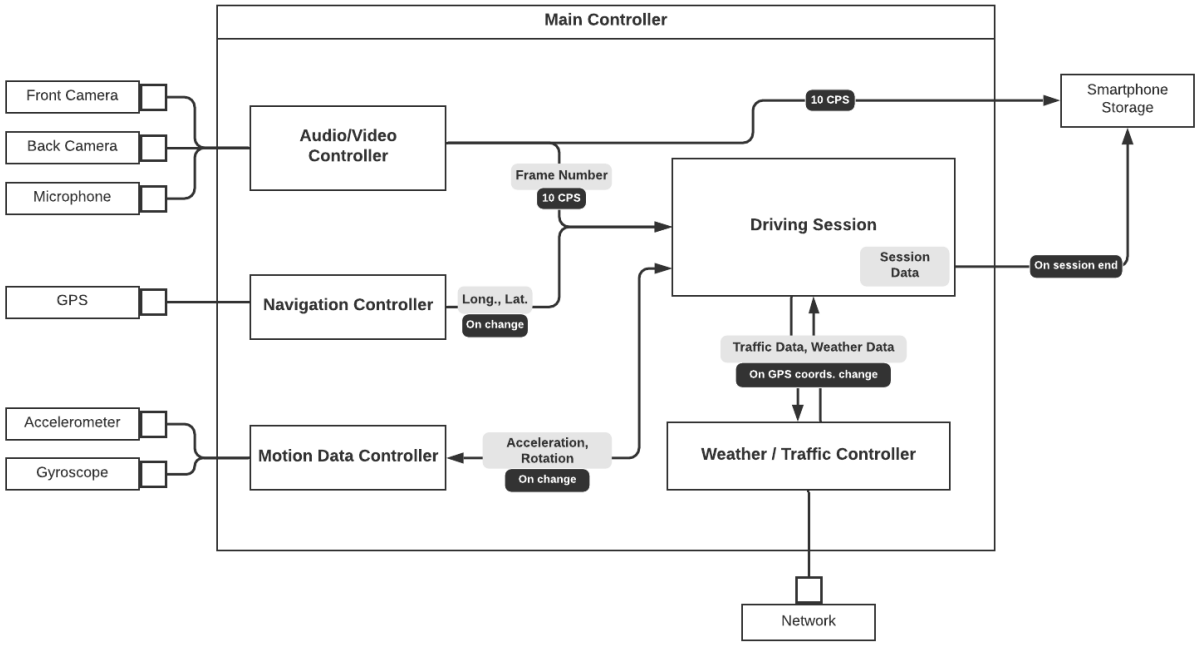


Fig. 3. Application's Reference Model. To the left of the *Main Controller*, sensors employed are displayed. *Sensing Controllers* intermediate the *sensor fusion* process by providing the *Driving Session* with up-to-date data from sensors. The Main Controller houses all the other controllers and manages data exchange. The Navigation Controller employs Apple's Mapkit framework to implement navigation and gather location data. The Weather/Traffic Controller calls both a weather API and a traffic API to fetch live data. The Session Data model holds at all times the most up-to-date value for each of the variables.

and traffic flow estimation, and (6) modeling. We describe the data pipeline in detail in the following paragraphs. The complete list of available features is presented in Table 1.

3.2.1 Visual Feature Extraction. We employ two parallel approaches to extract visual-related features from the road-facing frames: (a) object detection and (b) semantic segmentation.

Object Detection. For the object detection module, we used a PyTorch implementation[5] of a Yolo5 object detection model pre-trained on the COCO dataset[6]. The machine learning model outputs a list of objects detected, and their respective 2D bounding boxes (BB) expressed as normalized pixel coordinates (x, y, width, and height). The object classes are filtered to include only those of interest: cars, people, bicycles, motorcycles, buses, trucks, traffic lights, and traffic signs. We use the BBs to calculate the relative area of the object to the complete frame, representing the object's perceived size. We classify the relative area values using predefined thresholds into five different distance/perceived size classes (very far, far, medium, close, very close). Thresholds are devised based on observations about the frequency of occurrence of relative area values. The final output is a dictionary providing a scene summary with the number of objects for each detected class and the number of objects in each distance/perceived size class.

[5] https://pytorch.org/hub/ultralytics_yolov51
[6] https://cocodataset.org

Table 1. List of all available features in the dataset. We group the features by context and provide exemplary values or a description in the details column. The columns 'frame_number', 'timestamp', 'audio_file_path', 'front_file_path', 'latitude' and 'longitude' are not used as input for the machine learning models.

Context	Feature	Details
Reference Data	frame_number	The number reference for the session snapshot frame pair.
	timestamp	e.g. 21/10/15, 18: 55:39:0025
	audio_file_path	p_01/session_id/audio.mp4
	front_frame_path	p_01/session_id/imgs/front_frame_501.jpg
	back_frame_path	p_01/session_id/imgs/back_frame_501.jpg
Personal	sex	male, female, other
	car_model	e.g. VW Polo, Porsche Taycan
	age	Participant's age.
	participant_id	e.g. p_01, p_02
	emotion_before	Emotion before ride.
Session	session_id	e.g. 0751B8E9-3357-47E3-A862-CBFC60B88555
	session_start	e.g. 21/10/15, 18: 54:49:0015
	session_end	eg. 21/10/15, 19: 14:69:0485
Session Time	weekday	Mon. Tue. Wed., Thurs. Fri., Sat., Sun.
	daytime	Morning, Afternoon, Evening, Night.
Motion	acceleration_x	Acceleration on the x axis.
	acceleration_y	Acceleration on the y axis.
	acceleration_z	Acceleration on the z axis.
	vemotion_acceleration	(or acceleration_v1) Acceleration as in VEmotion [6].
GPS	speed	Vehicle speed in km/h.
	latitude	Latitude value of current location.
	longitude	Longitude value of current location.
Traffic Data	current_travel_time	Current travel time in seconds.
	free_flow_speed	The free flow speed expected under ideal conditions.
	current_speed	The current average speed at the selected point.
	free_flow_travel_time	The travel time in seconds under ideal free flow conditions.
	reduced_speed	Calculated with *free_flow_speed* minus *current_speed*.
Weather Data	wind_speed	Outside wind speed in km/h.
	precipitation_24h_mm	Rain fall measurement in millimetres.
	feel_temp_outside	"Feels like" temperature in Celsius.
	cloud_cover	Percent representing cloud cover.
	weather_term	e.g. cloudy, mostly cloudy, mostly sunny, sunny
Road Data	road_type	e.g. cycleway, footway, living_street, motorway, residential.
	max_speed	Maximum allowed speed for the current road.
	num_lanes	Count of available lanes on the road.
Facial Expression Pred.	facial_expression_label	Front-facing camera's classified emotion|.
Perceived Emotion	label	Emotion expressed by the participant during the experiment.
Audio	audio_amplitude	Audio amplitude averaged for duration of correspondent chunk.
	audio_loudness	Audio recording average loudness for duration of correspondent chunk.
	audio_zero_crossings	Audio zero crossing rate of correspondent chunk.
Visual Complexity (Object Detection)	num_cars, num_people, bycicles, pedestrians, motorcycles, buses, trucks, traffic_lights, traffic_signs	Num. of objects detected in the back-facing camera frame per class.
	num_med_close_objs, num_very_close_objs, num_close_objs, num_very_far_objs, num_far_objs	Num. of objects at an estimated distance from camera.
Visual Complexity (Segmentation)	road, sidewalk, building, wall, fence, pole, traffic light, traffic sign, vegetation, terrain, sky, person, rider, car, truck, bus, train, motorcycle, bicycle	Percentage pixels in back-facing frame representing class.

Semantic Segmentation. We trained a Deeplabv3-ResNet model[7] on the Cityscapes Dataset[8] to perform semantic segmentation on the back-facing frames. We limited the training data to the classes relevant to our study: road,

[7] https://pytorch.org/hub/pytorch_vision_deeplabv3_resnet101

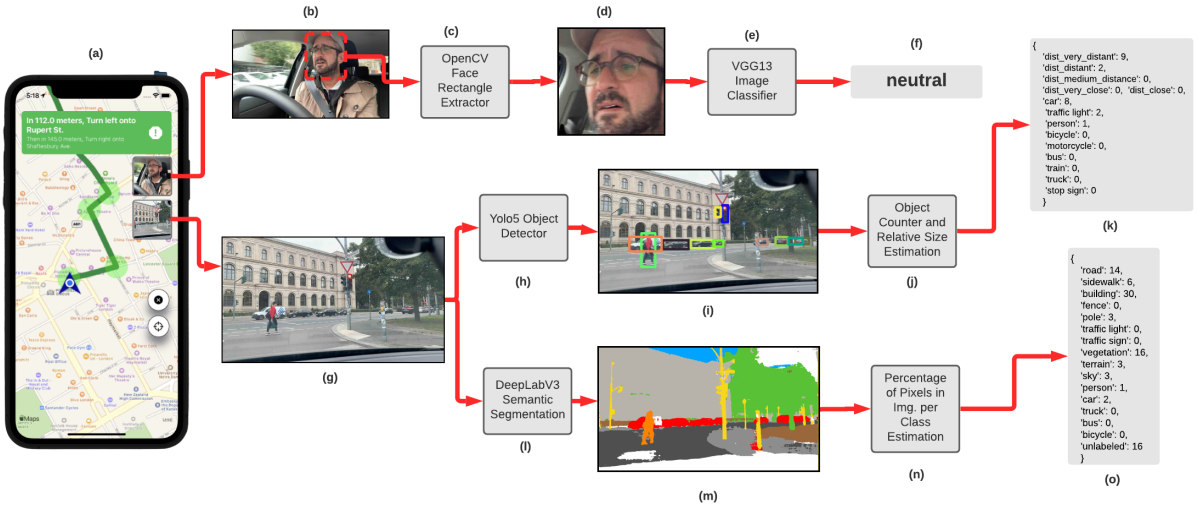


Fig. 4. Contextual Unobtrusive Sensor Feed Pipeline showing image processing from acquisition (a) to output features used for emotion prediction (f, k, o). The frame containing the driver's face (b) is cropped into a face rectangle (d) with Python OpenCV (c) and run through an emotion classifier (e). The back-facing camera frames (g) are processed with: 1) a Yolo5 object detection model (h), whose output (i) is further processed by (j), 2) a DeepLabV3 semantic segmentation model (m) which outputs a dictionary (o) of pixel per class.

sidewalk, building, wall, fence, pole, traffic light, traffic sign, vegetation, terrain, sky, person, rider, car, truck, bus, train, motorcycle, bicycle. This model's processed output consists of an array of shapes in which each pixel is assigned a class (m). Following, we calculate the percentage of pixels occupied by each class into a dictionary (o). The percentage of pixels associated with a specific road class attribute helps the system to understand how complex the visual field may be to the driver. For instance, a high number of pixels associated with cars and pedestrians may be due to a traffic jam and challenging driving scenarios. The visual segmentation engine's output is shown in Figure 5.

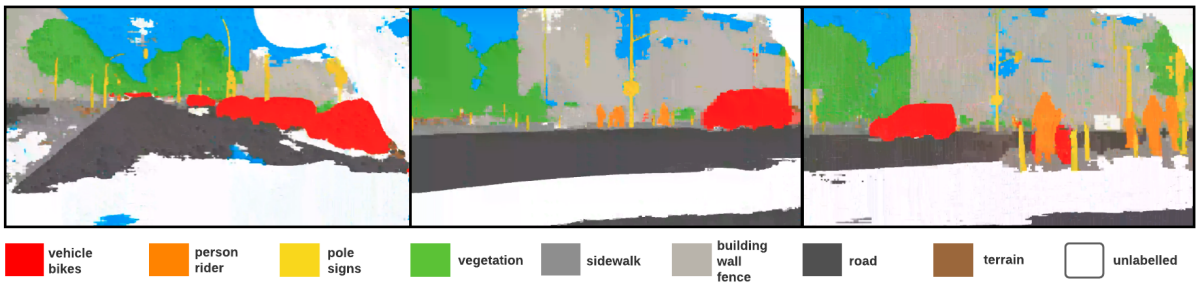


Fig. 5. Semantic Image Segmentation. Images showing segmentation results with different colors representing the predicted semantic classes. We color vehicles (cars, trucks, buses, trains) and bikes (motorcycles and bicycles) red. Yellow shows poles, traffic lights, and traffic signs.

[8] https://www.cityscapes-dataset.com

3.2.2 Audio Analysis. The smartphone's microphone is used in two ways: (1) to extract in-car loudness, (2) to compute the zero-crossing rate of the audio signal, and (3) to extract the annotated emotional labels expressed by the participants.

Loudness Extraction. We use audio amplitude and loudness in decibels (dB) to represent in-cabin driver auditory stimuli. We segment the audio stream into chunks of 0.5 seconds to calculate its mean amplitude. The formula for deriving loudness in dB is as follows:

$$loudness = 20 * log_{10}\left(\sqrt{\overline{chunk^2}}\right) \quad (1)$$

$$amplitude = \overline{chunk^2} \quad (2)$$

Zero-Crossing Rate. The Zero-Crossing Rate (ZCR) of an audio frame is a measurement of an audio signal's human-perceived *noisiness*. We calculate ZCR by counting the number of times a given audio signal crosses the zero axis and dividing it by the length of the frame. Unlike *loudness*, it incorporates spectral aspects of the signal and is widely used by applications in speech analysis [38, 48, 56] and musical genre classification [59]. Therefore, it is a good representation of driver auditory stimuli. Again, we derive a ZCR value for each chunk of 0.5 seconds.

Emotion Labels. Annotated labels are extracted manually onto a text file from the session's audio recordings. Label definitions and procedure of labeling are discussed later in the experiments section 4.

3.2.3 Facial Expression Classification. We use facial expression recognition to obtain baseline metrics for model performance comparison. Thus facial expressions are not included as features in the modeling phase. To extract facial recognition predictions, we use a face rectangle extractor and run its resulting image through a VGG13-based image classifier trained on the FERPlus Dataset[9]. Furthermore, as an additional baseline, we apply the Microsoft Face Recognition API classifier[10]. This step outputs a facial expression prediction for each of the session's front-facing frames. We stress that the emotion label provided by the facial expression classifier is not used as a feature for our model but represents a baseline metric.

3.2.4 Road Type Data. In order to detect the road infrastructure components of in-the-wild driving thoroughly, we acquire road-type-related features via reverse geocoding from OpenStreetMap[11] with the Python package OSMnx[12]. We download a high-definition map for each unique combination of GPS coordinates and search for the closest road node object to extract the relevant data. From the closest road node object, we extract the following attributes: 'road_type' (e.g., residential), the number of available lanes on the current road ('n_lanes'), and the maximum allowed speed on the current road ('max_speed').

3.2.5 Weather and Traffic Flow. We request weather information for each distinct GPS coordinate pair from the Microsoft Azure Maps API[13]. We include the following weather conditions: weather description, approximate outside temperature, cloud coverage, and wind speed. We also infer the traffic flow by requesting the speeds and travel times of the road fragment closest to the given coordinates using the Microsoft Maps Traffic Flow API[14].

[9] https://github.com/microsoft/FERPlus
[10] https://azure.microsoft.com/en-gb/services/cognitive-services/face
[11] https://www.openstreetmap.org
[12] https://github.com/gboeing/osmnx
[13] https://atlas.microsoft.com
[14] https://atlas.microsoft.com/traffic

4 EXPERIMENT

In the following section, we describe the details of the in-the-wild driving experiment.

4.1 Participants

In total, 27 participants (five female, ages 21 to 63, $\mu_{age} = 30.9$, $\sigma_{age} = 9.8$) took part in the experiment. The participants drove in total 48 sessions, with a total duration of 663.93minutes and a mean duration of 13.83 minutes (min = 9.0, max = 28.62, $\sigma_{duration} = 3.54$). The average number of unique emotions reported per session was 2.79.

4.2 Procedure

We asked participants to download the mobile app from Apple's beta-testing platform "TestFlight" and instructed them to use it as their navigation tool during two to three journeys with a duration of between 10 and 15 minutes. We also recommended that journeys be at different times of the day and, preferably, on different days to diversify the data collected as much as possible.

We needed the participant's front-facing camera to be pointed towards their face and the back-facing video stream to capture the road from the driver's perspective. Participants had to attach their smartphones to the windscreen.

The first time the subjects launched the app, they were asked for driver-specific context information, precisely their age, self-perceived gender, and car model. After that, we instructed the participants to choose their destination freely. Upon entering navigation mode after accepting the route proposed by the app, subjects were asked to select their pre-ride felt emotion and provide their emotion after the ride. Ethical approval for the experimental procedure was granted by the institutional review board of the university department.

4.2.1 Annotation of Emotions. To link the acquired contextual data with emotions on the road, we present the experimental design of emotion annotation in the following section. First, we explain how subjectively felt emotions can be acquired in the vehicle context. Second, we explain the mapping procedure and trade-offs between the expressed emotion and contextual ongoings. In our case, the driver expresses their emotions during the ride via voice without the need to let go of the steering wheel. In preparation for the experiments, subjects were asked whether they felt confident expressing their emotional states while driving. The verbally expressed emotion was recorded while driving and analyzed offline using a speech-to-text algorithm. We triggered a beep tone every 60 seconds for the drivers to express their discrete emotional state (a list of valid responses was given to the participants beforehand). Based on Ekman's basic emotion theory [16], we selected eight basic categorical emotion categories as possible response values: 'happiness', 'anger', 'fear','surprise', 'neutral', 'contempt', 'disgust' and 'sadness'.

Similarly to Bethge et al. [6], we adopt the *in-situ* categorical emotion response (CER) rating [13] for labeling in-the-wild emotions. We opted against continuous emotion labeling (DER) in the form of valence-arousal ratings, as users would need to select their continuous emotional rating via touch on an in-cabin device. Touch interactions are shown to distract from first-level driving tasks and pose a risk factor in this study [35]. The free categorical emotion response method is found to have practical limitations as it can generate a large number of labels. Consequently, Dittrich et al. [13] recommend adopting in-situ categorical emotion ratings (CER) with an "appropriate number and naming of categories that cover a significant range of emotions".

Furthermore, it is challenging to find the optimal time interval between prompting the driver for their emotion. On the one hand, we do not want to distract, bias, and annoy the driver when asking too often for an emotion. On the other hand, we want to ask as frequently as possible to have a granular resolution of the emotional ground truth that helps to learn a better link between our features and emotions. Using this annotation procedure, we link the expressed emotions to contextual data within the window of the previous 60 seconds. However, this approach is deliberately oversimplified, as emotion transitions might not be correctly reflected in the annotated

data. We accept this trade-off in favor of a more realistic driving experience with the highest acceptable amount of interruptions. We further address the limitations of the experimental choice of defining emotional labeling in section 7.5.

5 DATASET CHARACTERISTICS

The following section presents an overview and necessary preprocessing steps of our data. The dataset consists of 48 sessions of different participants driving in the wild with our system explained in Section 3 and the data-gathering procedure described in Section 4.

5.1 Data Preprocessing

We experienced performance oscillations with the data-gathering application due to, e.g., battery state differences. These performance inconsistencies caused some session fragments to have a higher frame output than others, resulting in inconsistent data distribution data over time. Therefore, the dataset was down-sampled to 3 Hz, i.e., using three frames per second as data entries. We decided to drop the first minute of each session due to the time difference between starting our app and actual driving. Our preprocessing removes, on average, 7% of data per participant due to, e.g., invalid sensor information.

Some participants had difficulty adhering to the experiment's predefined emotions and used non-complying labels. We substituted some of these labels with synonyms. 'Scared' and 'concerned' were renamed to 'fear'. 'Annoyed' and 'frustrated' were replaced with 'anger'. Occurrences of 'stressed' were attributed to the label 'unknown' due to its ambiguity (could be interpreted as 'anger' or 'fear'). Other non-complying labels such as 'curious' and 'confused' were also changed to'unknown'. Emotion labels with 'unknown', duplicate values, and other rows with missing data were removed from the dataset.

5.2 Data Summary

We briefly give an overview of the preprocessed dataset in the following section and will recap the feature streams from our system thereafter.

After preprocessing, the dataset comprises 97020 samples (663.93 min) of labeled driving data from 48 sessions with 27 participants. We present an overview of our dataset in Figure 6. Due to the wide variety and depth of the acquired data, we only plot a subset of available features in Figure 6. We plot the Pearson-correlation matrix of all extracted features in the appendix Figure 11.

5.2.1 Perceived Emotion Labels. An overview of the perceived emotion per participant is shown in the upper left plot of Figure 6. Overall, we observe many real-world traffic conditions where drivers felt 'neutral' (57%), which is unsurprising given normal traffic conditions during many rides. The participants also perceived happy emotions 17% of the time. Negative emotional states did not occur frequently and were expressed primarily by a few participants. Drivers expressed numerous times to feel 'fear' which can be explained by some participants driving non-frequently and feeling nervous in complex traffic situations. Apart from sadness, all other pre-selected emotions (neutral, fear, happiness, anger, surprise, disgust, contempt) occurred throughout the sessions. In average, 2.79 distinct emotion categories were expressed per ride. There was only one participant who expressed a sad emotional state, while all other states were felt by multiple drivers.

5.2.2 Speed. The histogram in Figure 6 (right plot in the second row) shows the distribution of 'speed' (in km/h) across all sessions, revealing that in most rides, the range goes from 0 to 50 km/h. Two sessions have speeds surpassing the 50 km/h mark and going up to 200 km/h. Most sessions took place in cities, whereas two collected data on a motorway with no maximum speed limit.

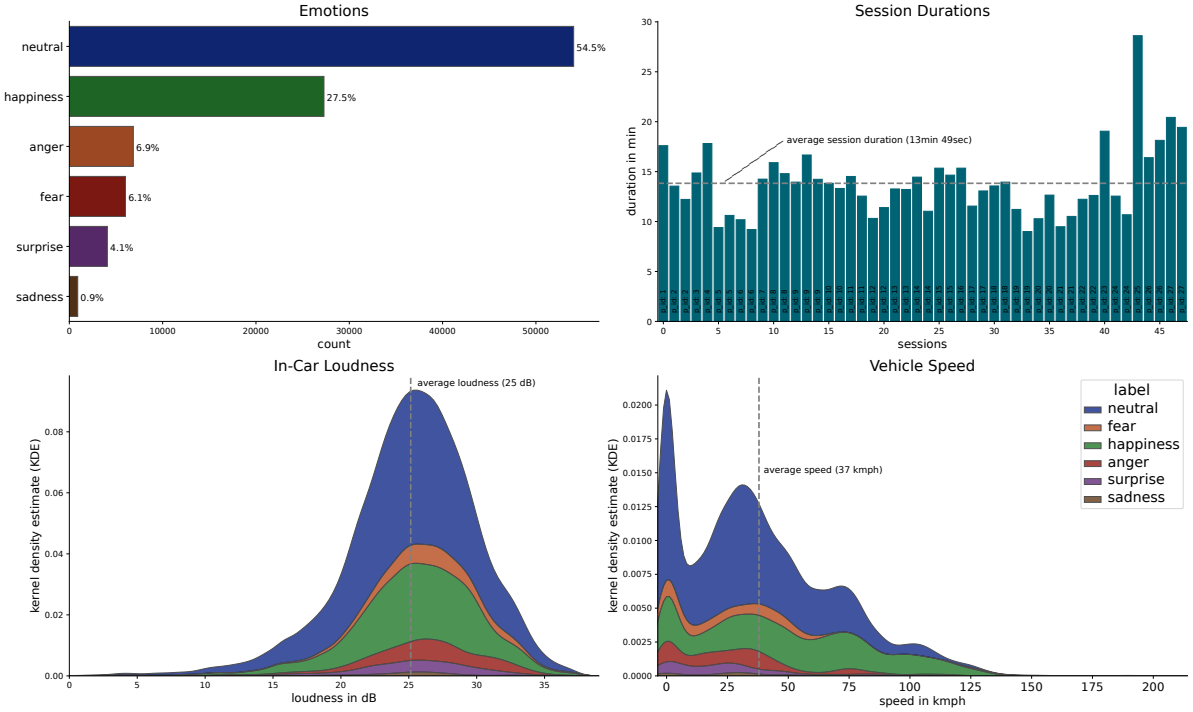


Fig. 6. Overview of the acquired data. While different personal meta-data statistics are shown in the figures in the top row, the bottom row shows the kernel density estimation (KDE) [47] of some contextual data including their label distribution.

5.2.3 Visual Object Detection. The computer vision module of our system tracks granular changes in environmental visual ongoings. Figure 7 shows an exemplary output of the sensor feed of the visual features 'num_cars', 'num_people' and 'num_traffic_lights' being computed. We observe that the computer vision module can detect ongoing traffic situations, e.g., pedestrians or traffic lights while driving.

6 RESULTS

The following section addresses how context, captured driving data, and environmental factors predict driver emotions. First, we analyze the features' importance for predicting driver emotions. After that, we evaluate the prediction performance of the different features in an extensive cross-validation setup and compare them against several baselines. At last, we compare the features' characteristics in terms of privacy and computational costs.

6.1 Emotion Classification Module

We set a Random Forest Ensemble Learning as a default classifier based on a 10-fold grid-search cross-validation[15]. The 'class_weight' parameter of the random forest is set to 'balanced' to ensure that the algorithm can handle an unbalanced label distribution. Further hyperparameters are found using a 10-fold hyperparameter tuning grid search (random state = 0, n_estimators = 50, max_features= log_2). This machine learning model is kept

[15]Using Support Vector Machines, KNeighbors, Decision Tree, Adaboost, and Random Forest classifier from scikit-learn with default parameters. The Random Forest achieved the highest average F_1 score.

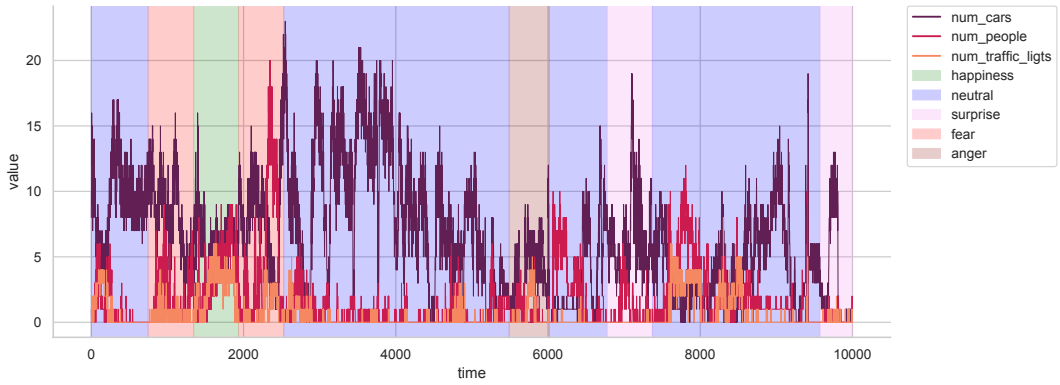


Fig. 7. Stream of outside-view visual complexity features for a sample session (participant 1). The y-axis shows the individual feature values, whereas the x-axis denotes the number of distinct, consequential data entries (sampled at 100 Hz). The object detection module recognizes the number of cars on the road, the number of people in the visual field, and the number of traffic lights at every moment of the drive. The prediction of the number of elements in the visual field varies as occlusion, shaky frames, and lighting conditions can occur.

identical to related work [6] to enable the comparison of results. We further evaluate the performance of a deep-learning-based feedforward neural network using all features. The neural network parameter settings are explained in detail in Appendix A.3. We explain the evaluation procedure in detail in the Section 6.4.

6.2 Importance of Features for In-The-Wild Emotion Recognition

We start by investigating how decisive each feature is for creating a classification model. Thus, we extracted the feature importance of the contextual variables: In a leave-one-participant-out situation, we assess the permutation importance for each variable, which is defined as the decrease in the balanced F_1 score of the classification algorithm. The permutation importance can be seen as a metric of how much performance we lose (here measured in F_1 score) if we do not have access to a specific system feature. This is done by randomly sampling the specific variable and thereby making the variable not-containing any meaningful information. The higher the permutation importance of a feature, the higher its predictive power, i.e., the more performance the classification decreases when it is unavailable. The feature importance does not provide information on which feature value contributes to a specific label prediction. We refer the reader to the concept of local feature importance computation, e.g., SHAP values, which could explain feature importances of a value range given an individual data object [40]. These importances are specific to a subject; therefore, this paper does not provide a local feature assessment. Figure 8 shows the calculated permutation importance for all features. In general, we observe that some features show very high importance, and most features do not. Overall, we detect a high importance of vehicle speed for the emotion classification decision. This finding overlaps with related work, which shows that free-flow highway driving and emotional happy states are tied. In contrast, low-speed values, combined with unforeseen traffic incidences such as traffic jams, have been associated with negative emotional feelings such as ‘anger’ and ‘contempt’ [61]. The available number of lanes on the road is a significant predictor in classifying the driver’s emotional states, which is unsurprising as a high number of lanes is weakly correlated with the traffic conditions, i.e., speed and acceleration behavior [33].

Furthermore, the felt temperature outside and the number of pixels associated with the sky (‘segment_sky’) also show high-importance measures. The high feature importance in both sky and environment temperature is

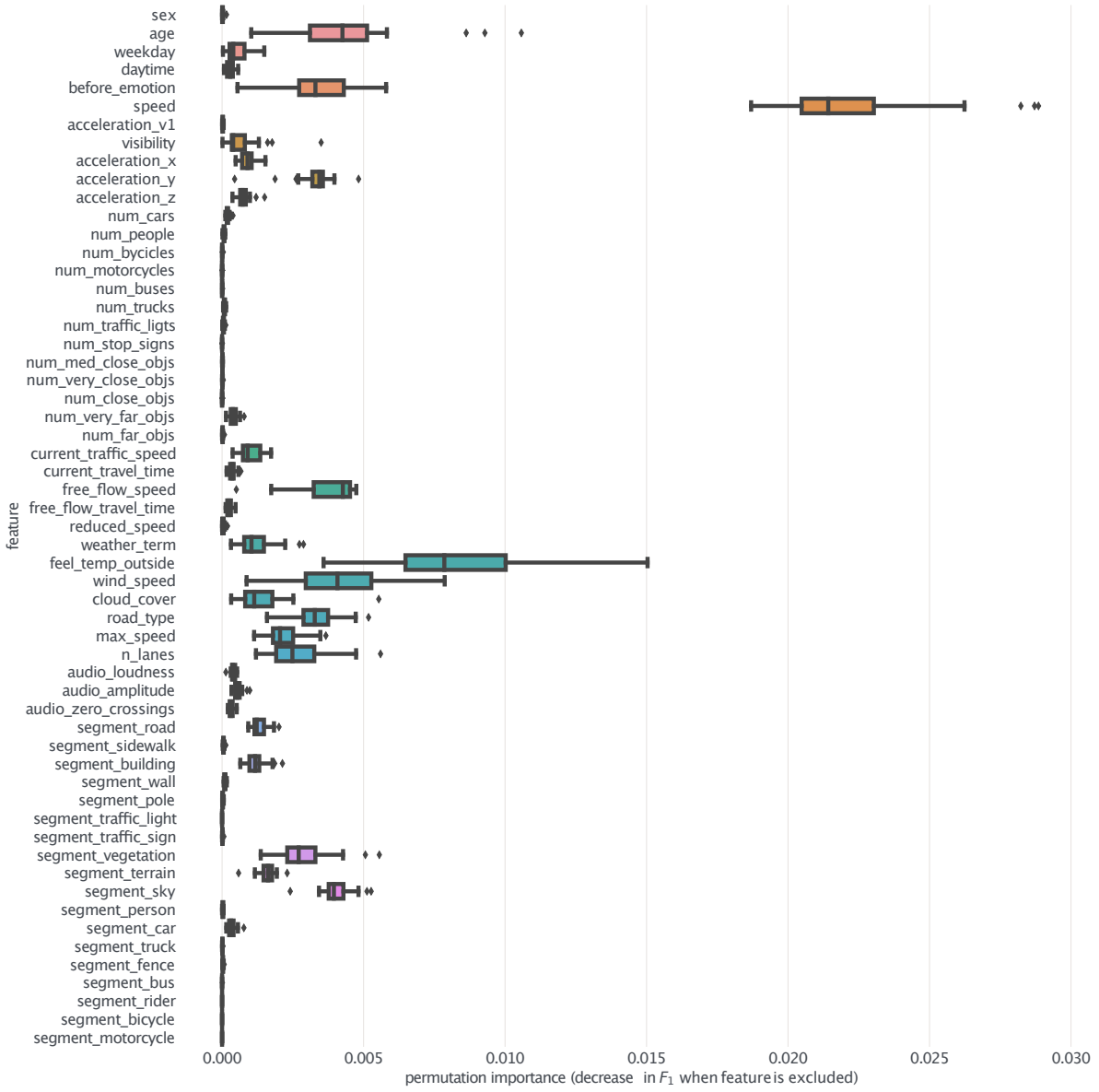


Fig. 8. Permutation importance of all available features. The permutation importance is characterized by a decrease in the model's F_1 score if the individual feature values are randomly permuted, i.e., made uninformative. Therefore, a high F_1 score indicates high feature importance to the emotion classification decision. Low permutation importance suggests that omitting the specific feature would not result in large prediction performance loss.

interesting, as related work has shown that greater sky exposure and air temperature tended to make drivers report lower stress levels and lower negative emotional states [3, 29].

Interestingly, the features representing drivers auditory complexity ('audio_loudness', 'audio_amplitude' and 'audio_zero_crossings') have low permutation importance scores making them not highly useful for classifying subjective emotional states. Driver speech analysis can predict human emotions successfully, and loud in-cabin sounds are potentially associated with driver distractions and the prevalence of annoyed and angry emotions [27, 49]. However, our results do not detect high audio feature importance, indicating that the other assessed contextual variables are more indicative of driver emotions. Although our in-the-wild results recommend omitting in-cabin audio recordings, further research is necessary to evaluate the impact of different audio features on emotions. For example, more advanced speech semantics, including tonality, pitch, or frequencies, can be more indicative of emotion recognition. Here, we expect individual differences in the audio data to mitigate the general classification performance at the cost of disclosing more privacy-sensitive data. In Section 6.5.2, we will further discuss the privacy-related concerns of in-cabin microphones.

6.3 Visual Driving Scene Features for Emotion Recognition In-The-Wild

The feature importance analysis provides insights into how indicative a feature is for predicting subjective emotions. Analyzing the driver's visual scene is beneficial for understanding the driving task's complexity and the environment's aesthetics [10]. This visual information can help to deduct the driver's well-being. The following section will further analyze the link between the extracted visual features and emotions in the wild.

We analyzed the driver's visual field in two ways: (1) via a computer-vision-based object detector for every outside view image frame of our system, and (2) via a visual segmentation engine. Figure 9 shows a boxplot of emotions over visual scene extracted features. Interestingly, compared to happy states, we see that a high number of detected cars in the visual field (high value of 'num_cars') is prevalent in conjunction with 'fear' driver states with a high degree of certainty ($p < .01$). Many cars in the visual driver scene are observed in dense traffic scenarios on the highway or in the city, often with traffic jams. Rural areas often have a lower incidence of cars. The median number of cars in happy states (2.0 - SD: 2.03) implies that fewer cars are prevalently observed in 'happy' states. We observe that the participants 'disgust' and 'contempt' states are in high traffic conditions, i.e., in conjunction with many cars. We note that the object detector also counts parked cars on the side of the road. However, the object detector only recognizes 2-3 parked vehicles in one frame due to occlusion, limiting visual scene object counting.

Looking at the visual segmentation features 'segment_car', we detect a higher degree of car scene pixels in the frame in negative emotional states. Similar to the 'num_car' feature, a lower percentage of car presence in the visual scene links to 'happy' emotional states. This observation validates previous research in driver stress recognition, which showed that high stress levels often happen in highway and city driving conditions [10]. Interestingly, the 'segment_vegetation' feature is not increased for happy emotional states compared to neutral states. The degree of sky presence ('segment_sky') for the driver visual field is highly relevant for increased for 'sad' emotional states, e.g., the presence of blue skies has been shown to affect personal well-being positively [58]. However, only the number of pixels associated with the sky is non-complete for defining a specific emotional state, as, e.g., rainy weather conditions in combination with the presence of the sky can induce negative emotional states [2]. As a result, the segmentation features of the outside-view can be regarded as a non-complete feature set for predicting subjective driver emotions in the wild. Further studies, including a broader range of study participants, should validate this visual scene object detection and segmentation findings.

6.4 Comparison of Recognition Performances

We perform an extensive evaluation setup to assess the feasibility of using diverse contextual, audio, and visual data streams for recognizing emotions in the wild. We compare the performance of the machine learning classifier system using different sensor stream inputs and evaluate their prediction performance. Table 2 provides the

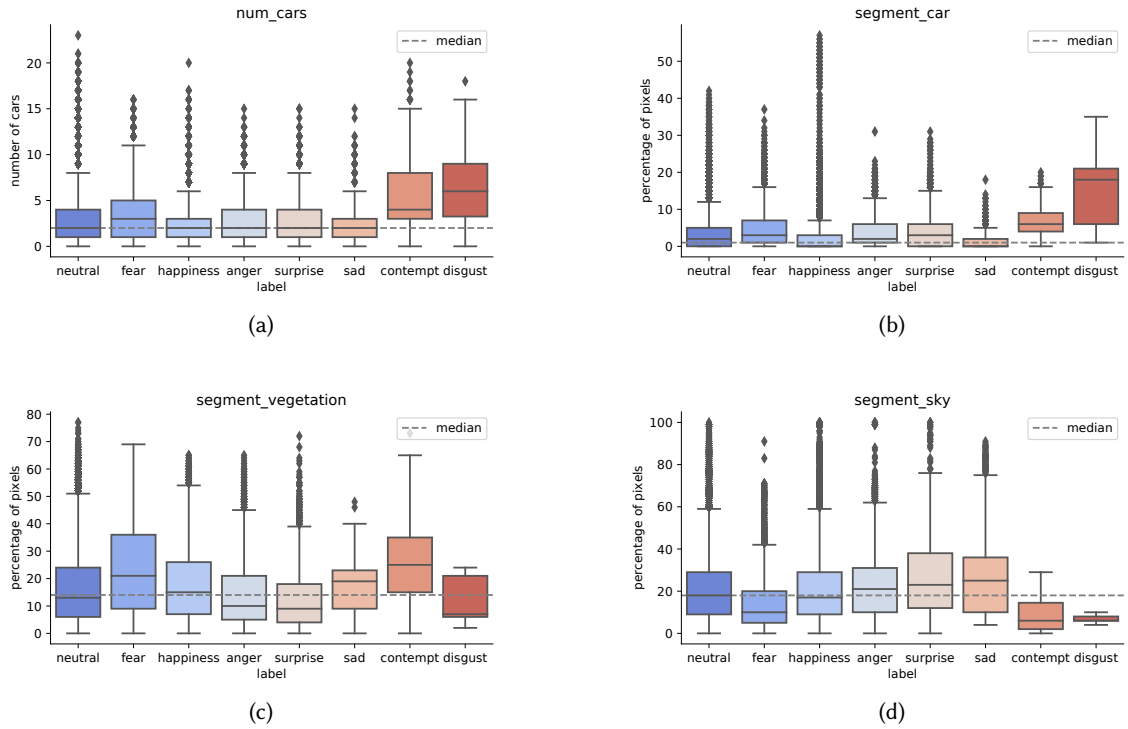


Fig. 9. Visual segmentation and object detection features relation to expressed emotional state. We analyze the mean differences of the features between the emotional states via a one-sided t-test controlled for sample size differences. **(a)** Boxplot of 'num_cars' feature. We observe a significant higher mean number of cars of the visual field in 'fear' ($p < 0.01$) emotional states than when people felt 'happiness'. The difference in observed number of cars between 'neutral' and 'happiness' is significant. **(b)** Boxplot of 'segment_car' showing the differences in percentage of pixels in visual fields showing cars per emotion. **(c)** Boxplot of 'segment_vegetation'. **(d)** Boxplot of 'segment_sky'.

prediction performance scores for a leave-one-participant-out evaluation procedure based on our system and baseline approaches.

6.4.1 Baseline Approaches. We compare our classification system against several state-of-the-art baselines. As a common remote sensing technology for emotion recognition in-vehicle, we set facial expression recognition as a baseline. FERPlus is a facial recognition classifier (VGG16) learned on the popular FERPlus dataset, and Facial Expressions (Azure) defines the classifying system via the Microsoft Face Recognition API. Furthermore, we compare our system to VEmotion proposed by Bethge et al. [6], a machine learning classifier system based on GPS-based context and driver-related metadata.

6.4.2 Emotion Extraction from Facial Expressions. Appendix Figure 10a shows a confusion matrix giving a comparison between the output from the facial expression recognition (FER) model against the true labels (i.e. emotions expressed by the participants during the experiment). FER outputs contained all of the possible labels[16], while the true labels had all but 'sadness'. FER achieved an overall mean accuracy of 43%, most of it owning to

[16]'anger', 'contempt', 'disgust', 'fear', 'happiness', 'neutral', 'sadness', and 'surprise'.

Table 2. Random Forest Evaluation Results. Table 2 shows averaged evaluation results for each of the feature groups in a global classifier learning leave-one-participant-out evaluation. We compute accuracy (Acc.), class-weighted precision (Prec.), unweighted average recall (UAR) and F_1 scores. Best values are indicated in bold.

	Global (Leave-One-Participant-Out)			
Feature Group	**Accuracy**	**Precision**	**UAR**	F_1
Facial Expressions (FERPlus)	.43 ± .12	.46 ± .13	.13 ± .03	.41 ± .13
Facial Expressions (Azure)	.55 ± .13	.48 ± .15	.17 ± .03	.47 ± .13
VEmotion (VE)	.52 ± .16	.38 ± .17	.26 ± .09	.42 ± .17
Visual Complexity Segmentation (VC-Seg.)	.54 ± .14	.48 ± 0.12	.18 ± .03	.44 ± 0.16
Visual Complexity - Object Detection (VC-ObjD.)	.50 ± .11	.32 ± .13	.09 ± .02	.37 ± .13
VC-Seg. + Audio	.55 ± .14	.49 ± .09	.17 ± .03	.44 ± .17
VC-ObjD. + Audio	.48 ± .12	**.50 ± .11**	.12 ± .01	.45 ± 0.13
Audio only	.37 ± .06	.49 ± .1	.10 ± .01	.40 ± .09
Audiovisual (OjbD. + Seg. + Audio)	.56 ± .14	.48 ± .15	.19 ± .03	.44 ± .17
GPS-infered features only	.57 ± .16	.44 ± .19	.29 ± .11	**.46 ± .18**
All features	**.59 ± .15**	.43 ± .17	**.29 ± .08**	.45 ± .17
Neural Network (All features)	.43 ± .14	.52 ± .15	.18 ± .06	.43 ± .14

the model predicting 'neutral' correctly while failing to do so in a significant way on any of the other classes. 'Anger', for instance, is often miss-predicted as 'sadness' (26%) or 'neutral' (64%). 'Fear' emotional states are never correctly predicted as such, and even predicted as 'happiness' in 4% of the cases.

Although our results show that facial expressions poorly predict facial expressions, we acknowledge that our used API (i.e., Microsoft Azure) and training dataset (i.e., FERPlus) are not optimized for in-vehicle use. Training a participant-dependent model about the contextual vehicle data can improve classification performance. However, this would require prior individual data collection. Other platforms, such as Affectiva[17], offer car-specific classifications but are costly to deploy. Comparing different facial expression classification platforms for in-vehicle use is a research topic for future work. In summary, off-the-shelf facial expression classification substantially over-predicts 'neutral' and under-predicts all other emotional states, showing a worse prediction performance driver emotions in-the-wild.

6.4.3 Global Modeling: Leave-One-Participant-Out. We evaluate the feasibility of a general classification model using all participant data except for one for training and using the last participant for evaluation. Semantically, this approach learns a model without knowing anything about the driver in advance and predicts the drivers' emotions independent from individual context emotion preferences. In production, such a model could be trained once on a set of participants and then shipped to the customer's vehicle without retraining. By using this

[17] www.affectiva.com

cross-validation setting, the chance of overfitting to participants or specific road properties is very low. Besides accuracy, class-weighted recall, precision, and F_1 score, we address the issue of unbalanced emotional class labels by reporting the unweighted average recall (UAR). UAR calculates the recall for each label and finds its unweighted mean.

Looking at the accuracy of our sensory system, all features combined achieve the highest average prediction accuracy of 59%. This is significantly better than using facial recognition engines alone which achieve an accuracy of 43% (FERPlus) and 55% (Azure). The difference of the all features model to VEmotion is 7 percentage points showing that incorporating additional features for global modeling is favorable. Furthermore, visual complexity segmentation features alone can achieve a high emotion recognition performance of 54%. The high performance of using only visual segmentation shows that outside-view information only by camera systems can already predict emotions for unknown drivers. This result offers the chance of using the already in-the-car integrated segmentation results of some autonomous driving control units for scene understanding to infer visual complexity and possible subjective emotions.

Interestingly, using the audio-visual complexity measures (visual object detection, visual segmentation, and in-cabin audio features) seems to increase the performance of the classifying system by only 2%, so that acquiring inside-cabin audio information does not improve results significantly. We also explored the possibility of learning participant-dependent models i.e., models that are trained only on an individual participant's data. We report the evaluation results in the Appendix Table 4.

6.4.4 Conclusion of Model Performances. In general, the hierarchy of prediction performances for specific feature sets remains constant across the evaluation settings, i.e., visual complexity only features show high recognition performances. A promising alternative is using only GPS-inferred features. We observe the highest performance of 59% to predict subjective emotions on unknown participants. Overall, a participant-independent classifier is able to predict emotions on unknown participants confidently and enables a promising alternative to e.g., facial expression detection. Furthermore, the global model can be employed as is and enables the possibility to be retrained on individual participants context preferences.

6.5 Technical Design Considerations

In-car real-time applications such as driver emotion monitoring are developed under strong computing power constraints [55]. The number of extracted features can strongly affect the computing time of the algorithm. Furthermore, features that require intensive computation and may only perform equally well as other features are often not used. An additional constraint to be respected is the degree of privacy erosion caused by each feature. With that in mind, in the next section, we discuss the importance and relevance of all features regarding their computational cost, privacy impact, and influence on model performance. We present the results of this trade-off in Table 3.

6.5.1 Computational Cost Factors. We choose two factors as the base for specifying computational complexity: (1) local computability and (2) third-party-API dependence [39]. These factors are crucial for the time delay caused by a single feature. Differences in computing time are insignificant as long as the feature can be computed locally. Equal differences in computing time are negligible as soon as the feature needs to be inferred externally. Further notable increases occur when a third-party API is required, as described in the second parameter. Also, the dependence of a third-party API negatively impacts privacy [17].

6.5.2 Privacy-Eroding Factors. We treat Stark et al.'s [53] work as a starting point defining the emotional context of information privacy. One factor to characterize the degree of privacy erosion is the type of sensors required to collect a particular feature. For example, in-cabin audio or video recording may contribute to a feeling of surveillance more than just an accelerometer or GPS tracking. Next, a sensitive topic is whether user data had to

be transferred over the internet, which, e.g., can be a potential data leak and, therefore, potentially privacy erosive. Finally, another erosive privacy aspect is the need for private and/or sensitive data as defined by Zainab et al. [60]. Overall, emotion prediction is a highly personal prediction decision that should be treated with caution.

Table 3. Trading off ubiquitous feature stream importance. We show different cost computational and privacy eroding factors of features while trading them off against their influence on performance in the form of F_1 decrease.

		Feature Acquisition Factors				Privacy Factors				Prediction Importance
Context	**Feature**	**Required Sensors**	**Complex Preprocessing**	**Locally Computable**	**Third-Party API Dependent**	**Transfer of User Data over the Internet**	**Personal Data**	**Sensitive Data**	**In-Cabin Recording**	F_1 **Decrease**
Personal	*sex* *car_model* *age* *before_emotion*	–	✗	✓	✗	✗	✓	✗	✗	0.0 0.0 0.001 0.003
Session Time	*weekday* *daytime*	–	✗	✓	✗	✗	✓	✗	✗	0.0 0.001
Motion	*acceleration_x* *acceleration_y* *acceleration_z* *vemotion_acceleration*	*Accelerometer*	✗	✓	✗	✗	✓	✗	✗	0.002 0.005 0.001 0.0
GPS	*speed* *latitude* *longitude*	*GPS*	✗	✓	✗	✗	✓	✗	✗	0.025 – –
Traffic Data	*current_travel_time* *free_flow_travel_time* *current_traffic_speed* *free_flow_speed* *reduced_speed*	*GPS*	✗	✗	✓	✓	✗	✗	✗	0.0 0.0 0.001 0.008 0.0
Weather Data	*wind_speed* *precipitation_24_hours_mm* *feel_temp_outside* *cloud_cover* *weather_term*	*GPS*	✗	✗	✓	✓	✗	✗	✗	0.001 0.0 0.009 0.001 0.0
Road Data	*road_type* *max_speed* *num_lanes*	*GPS*	✗	✗	✓	✓	✗	✗	✗	0.006 0.002 0.011
Facial Expression Pred.	*facial_expression_label*	*In-Cabin Camera*	✓	✓	✗	✗	✓	✓	✓	–
Audio	*audio_amplitude* *audio_loudness*	*In-Cabin Microphone*	✓	✓	✗	✗	✓	✓	✓	0.001 0.001
Visual Complexity (Object Detection)	*buses, bicycles, cars, close_objs, far_objs, med_close_objs, motorcycles, people, trucks, stop_signs, traffic_ligts, very_far_objs, very_close_objs*	*Outside-View Camera*	✓	✓	✗	✗	✗	✗	✗	0.0, 0.0, 0.0, 0.0 0.0, 0.0 0.0, 0.0, 0.0, 0.0 0.0, 0.0 0.001
Visual Complexity (Segmentation)	*bicycle, building, bus, car, fence, motorcycle, person, pole, terrain, rider, road, sidewalk, vegetation, sky, traffic_light, truck, traffic_sign, wall*	*Outside-View Camera*	✓	✓	✗	✗	✗	✗	✗	0.0, 0.005, 0.0, 0.001, 0.0, 0.0, 0.0, 0.0, 0.0, 0.001, 0.0, 0.009, 0.001, 0.0, 0.0, 0.0, 0.002, 0.0

6.5.3 Prediction Importance. The final factor is the influence on the model performance, which is evaluated by the decrease in F_1 score if the respective feature is removed.

Based on Table 3 and the above-described evaluation parameters, use-case-oriented feature sets can be built. In the setting of a production car, the manufacturer should focus on an easy-to-compute and privacy-preserving set of features. Hence we recommend a feature set without sensitive data and where all features are preferred to be locally computable and third-party API independent. We propose a performance-oriented feature set designed to allow more privacy erosion to have higher accuracy while maintaining a low computational cost that the user can manually select. This feature set would neglect the privacy protection to improve performance and use all of our proposed features. For research purposes, we propose to use computer vision extracted features,

i.e., object detection and visual scene segmentation features. First, these features are not directed at the drivers themselves and offer unobtrusive sensing of emotions benefiting driver privacy factors. Second, they show high feature importance while also offering possible local computability. Third, research on driver-view context affecting driver's well-being is underexplored [10]. Overall, we do not recommend empathic application designers to acquire emotions through facial expression analysis due to their non-robust detection and privacy-related concerns [54]. Still, many car companies employ driver facial monitoring software as driver-facing cameras are already equipped in-car, and facial expression software is easy to integrate [7, 11].

7 DISCUSSION

We propose a technical design space that extracts a large bandwidth of streams, giving information about the contextual in- and outside cabins ongoings using a consumer smartphone only. Our results show that contextual features are highly informative for recognizing the driver's emotional states in the wild. The approach inhibits several limitations and ethical considerations. The following section will discuss our findings and propose endeavors for future work.

7.1 Context and Audio-Visual Features Predict Emotions

We show that a consumer smartphone paired with machine learning modeling and computer vision can predict emotions in the wild. The capabilities and variety of sensors in our smartphones will increase in the future, and head-worn devices such as augmented reality glasses are already in development for large-scale use. This poses a challenge for future remote sensing systems, as small ubiquitous devices can infer context from little sensory information to predict emotional states. Our results show that driver emotions can be classified with up to 59% when using contextual and audio-visual features, an improvement of 7% over emotion detection using facial expressions. Our work confirms previous results using contextual data as a reliable classification input for emotions [6, 36], where adding environmental data streams (i.e., the outside and inside view) can improve the overall emotion classification performance. This conforms with previous work that showed how fast-paced changing driving situations influence the state of drivers, such as stress [10, 52]. Our results show that this concept can also be translated to emotions: environmental conditions are indicative of emotions and improve the overall classification accuracy when analyzed together with contextual data.

7.2 A Technical Framework to Prototype Empathic Car Interfaces

We separated and investigated different data streams for their influence on the overall classification performance. Including all features (i.e., contextual and audio-visual) provided the most efficient classification performance. We derived a technical design framework (see Table 2), separating the influence of the different data streams on the overall accuracy. On one side, designers and developers of empathic car interfaces can choose which data streams are available on the hardware or which data streams are necessary to achieve a particular classification performance. On the other side, users can enable or disable specific data streams to their preferences and desired classification accuracy. For example, users can opt-in for contextual data only and leave out the environmental data in case of privacy concerns. Developers and users can suit their sensing preferences according to the use case. Since the results of our study are obtained using a smartphone only, we envision that developers and designers can inexpensively prototype novel empathic car interfaces using the evaluated data streams. We are confident that our work will encourage researchers to investigate additional data streams for emotion-sensing while fostering rapid prototyping of novel empathic car interfaces.

7.3 Ethics and Privacy

We emphasize an ethical and cautious use of the context features explored in this study. Emotion-related information is highly personal and, thus, this sensitive data must be handled appropriately. The proposed approach uses the external-view camera stream, which may capture other people's information. Nevertheless, it offers a potentially more driver-privacy-preserving and discomfort-reducing alternative to the driver for measuring emotions in the wild than using facial expression or voice analysis, which requires in-cabin audio and video recordings. Filming outside, however, affects other people's privacy so that obfuscating faces are necessary, which is discussed thoroughly in related work [1, 37]. Moreover, our system can further alleviate privacy concerns by running object detection locally. This on-device run approach would make the system completely independent from an internet connection or GPS coordinates and third-party APIs, enabling broader coverage, e.g., in tunnels or remote country roads.

The analysis shows that visual features alone can predict emotions with reasonably high accuracy of 54%, outperforming facial expression analysis significantly. The robust performance provides the designer of affective in-car systems with new possibilities that do not involve cameras directed at the driver, which might raise a feeling of surveillance. Instead, our approach may only require an image representing what the driver sees. Furthermore, current driver assistance systems already obtain fine-grained outside-view information from sensors attached to the vehicle, which could directly serve as input for a potential in-cabin emotion classifying system based on visual features.

7.4 Reproducibility

We gain many insights by recording a fine-grained picture of the driver, its surroundings, and possible influences on emotion in a noisy real-world environment. Our work equips automotive user interface designers with an additional tool to design unobtrusive empathic car interfaces deployed in real-world scenarios. Furthermore, to encourage research in this area, we enable other researchers to access the data, reproduce our results and use the smartphone sensing architecture on their own by making our source code publicly available at https://github.com/msatiya/unobtrusive_driver_emotion_ds/.

7.5 Limitations

7.5.1 Emotion Annotation. Emotions are complex psycho-physiological phenomena and, as such, are difficult to study, especially in experiments in the wild. Several participants raised concerns regarding the emotion representation model, expressing that the predefined set was hard to memorize, had a priming effect, and did not allow them to truly express their emotions. Besides, they mentioned not being able to differentiate between 'contempt' and 'disgust' or, in some cases, did not even know what 'contempt' meant. This raises transparency concerns about the functionality and accuracy of AI-related classifications [32]. Furthermore, the difficulties of tracking emotions in driving contexts and, in particular, the pitfalls of using discrete emotions have been discussed thoroughly in related work [13, 62]. We recognize that this methodology is prone to noise and renders diminished nuance, but it is practically viable for in-the-wild driving contexts.

Our emotional annotation process is designed to collect ground-truth emotional labels of drivers in the wild. Due to the individual subjective nature of the expressed emotion, the label's robustness is based on trustworthiness of the participants to provide their true subjective feelings and cannot be verified by outsiders. Furthermore, the acquired emotional labels exhibit that the driver's emotions are heavily class-imbalanced. The emotion distribution has high support for neutral and happy classes, whereas there is little data support for, e.g., surprise emotions. This heavy class-tailed emotion distribution reflects the true underlying distribution of driver emotions in the wild and is not an effect of the annotation process.

7.5.2 Feature Importance Interpretations. Regarding the feature importance interpretation, we do not have sufficient data to make explicit statements about whether a specific visual feature can increase the probability of a particular emotion. Related work has shown that emotions can be assessed in various ways and a variety of factors can induce subjectively felt emotions. In our study, we used various environmental, visual, auditory, and contextual data, however, this subset is non-complete, and further efforts on, e.g., cultural aspects influencing subjective emotions can be evaluated. In our study, we could show that a high number of traffic participants in the driver's visual field are indicative of negative emotional states, however, this visual complexity assessment could be different in, e.g., India, where dense traffic is the norm.

7.5.3 Generalizability and Real-World Applicability. Our dataset contains a preliminary study of in-the-wild driver emotions. Our dataset contains multiple caveats that affect the model's generalizability: imbalanced emotion class labels, not all registered participants drove multiple sessions, and heterogeneous in-the-wild data acquisition setting. To not overfit specific participants, we decided to report the results of a leave-one-participant-out cross-validation and were able to show that the model outperforms baselines. Furthermore, the average session duration is 13.83 minutes, while the general daily usage of vehicles in the US is 27.6 minutes. Therefore the gathered dataset is acquired under realistic circumstances, but longer commute times are unavailable.

7.5.4 Sensor Data Quality. Our work presents results based on the current state-of-the-art gathering and analysis of smartphone data captured in the wild. The model's performances and sensor data quality thus should be regarded as a pillar of what is currently possible, however, some sensor data quality limitations still exist. The dataset contains features with high variability (e.g., speed), many of which cannot be controlled in an in-the-wild setting. Therefore, a more extensive dataset could also lead to better global models by covering more situations than currently represented in our dataset. This could also have a positive impact on the performance of a global model.

The back-facing camera frames present considerable variance across sessions due to different camera positioning and dashboard settings in different cars, resulting in inconsistent angles relative to the road. The visual field differences may result in uneven representations of the driving context. A camera-calibration step could be introduced in the app to add consistency to the collected data. The classification of relative distance to camera fails when a portion of an object is occluded, resulting in it being classified as further than it is. Too many distance classes create irrelevance for the less frequent ones. Some vehicles with lower incidence, like motorcycles, should be included with similar ones (e.g., buses and trucks as large vehicles). Our visual segmentation engine shows good recognition performance for subjective emotions, however, using a mounted camera inhibits several limitations. The camera frame quality (shaky video streams, low-resolution frames, and low frame rates) and occlusion due to, e.g., a truck occluding the other ongoing traffic participants affects the segmentation performance. Ongoing advancements in the smartphone camera quality render some of the issues mentioned above unimportant. Furthermore, reverse-geocoding might fail to recognize the exact road type for every geolocation due to imprecise GPS or nearer pedestrian road elements.

7.6 Future Work

Our study can be extended using a more extensive database of rides with a wider variety and distinction of emotions and more extended personal driving history. We propose a longitudinal study including more participants and longer sessions. Including a broader range of traffic scenarios while addressing the previously mentioned labeling issue by grouping emotions and adding other possible categories likely to arise in traffic contexts (e.g., stressed, confused).

To address the approach's limitations, we recommend revisiting the choice of features. Emulating relevant non-visual features would certainly help increase performance. For example, image segmentation could easily

extract the number of lanes from the camera stream, with the advantage of more precision and independence than using GPS and third-party APIs. The modeling approach does not learn time-dependent information across multiple GPS traces and image frames. Future work can address this time-relationship learning of emotional context by fitting a time-aware model, e.g., recurrent neural networks.

Furthermore, several features used in the study can provide a more detailed emotion assessment. For example, advanced audio features, including pitch, frequencies, or lexical density, can provide more insights into driver emotions. However, this requires recording the voice and environmental sounds, impacting the driver's privacy. In future work, we will investigate how advanced audio features relate to emotions and how privacy-preserving emotion prediction through audio can be implemented.

7.6.1 Additional Sensor Stream Integration. Additional sensor streams can be easily integrated to infer a broader range of environmental ongoings [5]. For example, newer smartphone generations offer a light intensity sensor which could be an informative feature for explaining emotional feelings in the wild. Furthermore, we envision a non-remote sensing scenario in the future, where additional physiological information from a wearable smartwatch is connected to the smartphone. For example, physiological signals such as galvanic skin response and heart-rate-variability have shown to be a good predictor of arousal levels.

7.6.2 Additional Application Scenarios. The proposed sensor system stream can infer contextual, environmental, and visual-auditory scene understanding for various in-the-wild application scenarios. Furthermore, our easy-to-integrate smartphone app can be used in bike studies to infer emotions for bike riders, e.g., urban areas, to provide infrastructure planers feedback of bike riders' emotions. This approach can be combined with advanced immersive technologies [31] to obtain more accurate emotion assessment results.

8 CONCLUSION

This paper presents a novel technical design space using contextual and audio-visual data for unobtrusive driver emotion detection. We show that by analyzing the audio-visual complexity of the outer-car ongoings, driver emotions can be predicted with 59% accuracy in a leave-one-participant-out cross-validation using a smartphone only. In contrast, only-outside view information using the smartphone's camera stream on the road offers a recognition accuracy of 54% while providing a less driver-privacy intrusive sensing system. Our smartphone-based sensor fusion implementation is uncomplicated to integrate into other ubiquitous sensor streams with GPS or camera functionality. We make our implementation and data publicly available to foster research in this area. We encourage the research community to participate in improving on-the-road emotion classifications and discuss the ethical implications of using empathic car interfaces.

ACKNOWLEDGMENTS

This work has been partly funded by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A (MCML).

REFERENCES

[1] Prachi Agrawal and PJ Narayanan. 2011. Person de-identification in videos. *IEEE Transactions on Circuits and Systems for Video Technology* 21, 3 (2011), 299–310. https://doi.org/10.1109/TCSVT.2011.2105551

[2] Liisi Ausmees, Anu Realo, and Jüri Allik. 2011. The Influence of the Weather on Affective Experience: An Experience Sampling Study. *Journal of Individual Differences* 32 (01 2011), 74–84. https://doi.org/10.1027/1614-0001/a000037

[3] Francisco Benita and Bige Tunçer. 2019. Exploring the effect of urban features and immediate environment on body responses. *Urban Forestry & Urban Greening* 43 (2019), 126365. https://doi.org/10.1016/j.ufug.2019.126365

[4] David Bethge, Lewis Chuang, and Tobias Grosse-Puppendahl. 2020. Analyzing Transferability of Happiness Detection via Gaze Tracking in Multimedia Applications. In *ACM Symposium on Eye Tracking Research and Applications* (Stuttgart, Germany) *(ETRA '20 Adjunct)*. Association for Computing Machinery, New York, NY, USA, Article 34, 3 pages. https://doi.org/10.1145/3379157.3391655

[5] David Bethge, Philipp Hallgarten, Tobias Grosse-Puppendahl, Mohamed Kari, Ralf Mikut, Albrecht Schmidt, and Ozan Özdenizci. 2022. Domain-Invariant Representation Learning from EEG with Private Encoders. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1236–1240.

[6] David Bethge, Thomas Kosch, Tobias Grosse-Puppendahl, Lewis L. Chuang, Mohamed Kari, Alexander Jagaciak, and Albrecht Schmidt. 2021. *VEmotion: Using Driving Context for Indirect Emotion Prediction in Real-Time*. Association for Computing Machinery, New York, NY, USA, 638–651. https://doi.org/10.1145/3472749.3474775

[7] Fariz Redzuan bin Monir, Rusyaizila Ramli, and Nabilah Rozzani. 2021. Driving Alert System Based on Facial Expression Recognition. In *2021 IEEE International Conference on Automatic Control & Intelligent Systems (I2CACIS)*. IEEE, 104–109. https://doi.org/10.1109/I2CACIS52118.2021.9495910

[8] Michael Braun, Jonas Schubert, Bastian Pfleging, and Florian Alt. 2019. Improving Driver Emotions with Affective Strategies. *Multimodal Technologies and Interaction* 3, 1 (March 2019), 21. https://doi.org/10.3390/mti3010021

[9] Michael Braun, Florian Weber, and Florian Alt. 2021. Affective Automotive User Interfaces–Reviewing the State of Driver Affect Research and Emotion Regulation in the Car. *ACM Comput. Surv.* 54, 7, Article 137 (sep 2021), 26 pages. https://doi.org/10.1145/3460938

[10] Cristina Bustos, Neska Elhaouij, Albert Sole-Ribalta, Javier Borge-Holthoefer, Àgata Lapedriza, and Rosalind Picard. 2021. Predicting Driver Self-Reported Stress by Analyzing the Road Scene. 1–8. https://doi.org/10.1109/ACII52823.2021.9597438

[11] Silvia Ceccacci, Maura Mengoni, Andrea Generosi, Luca Giraldi, Giuseppe Carbonara, Andrea Castellano, and Roberto Montanari. 2020. A Preliminary Investigation Towards the Application of Facial Expression Analysis to Enable an Emotion-Aware Car Interface. 504–517. https://doi.org/10.1007/978-3-030-49108-6_36

[12] Monique Dittrich. 2021. Why Drivers Feel the Way they Do:An On-the-Road Study Using Self-Reports and Geo-Tagging. In *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '21)*. Association for Computing Machinery, New York, NY, USA, 116–125. https://doi.org/10.1145/3409118.3475130

[13] Monique Dittrich and Sebastian Zepf. 2019. Exploring the Validity of Methods to Track Emotions Behind the Wheel. 115–127. https://doi.org/10.1007/978-3-030-17287-9_10

[14] Xinyu Du, Yue Shen, Ruosong Chang, and Jinfei Ma. 2018. The exceptionists of Chinese roads: The effect of road situations and ethical positions on driver aggression. *Transportation Research Part F: Traffic Psychology and Behaviour* 58 (Oct. 2018), 719–729. https://doi.org/10.1016/j.trf.2018.07.008

[15] Paul Ekman. 1992. Are there basic emotions? *Psychological Review* 99, 3 (1992), 550–553. https://doi.org/10.1037/0033-295X.99.3.550

[16] Paul Ekman. 1999. Basic Emotions. *Handbook of Cognition and Emotion* (1999).

[17] Benjamin Eriksson, Jonas Groth, and Andrei Sabelfeld. 2019. On the Road with Third-party Apps: Security Analysis of an In-vehicle App Platform. 64–75. https://doi.org/10.5220/0007678200640075

[18] Vérane Faure, Régis Lobjois, and Nicolas Benguigui. 2016. The effects of driving environment complexity and dual tasking on drivers' mental workload and eye blink behavior. *Transportation Research Part F: Traffic Psychology and Behaviour* 40 (2016), 78–90. https://doi.org/10.1016/j.trf.2016.04.007

[19] Tara E Galovski and Edward B Blanchard. 2004. Road rage: a domain for psychological intervention? *Aggression and Violent Behavior* 9, 2 (2004), 105–127. https://doi.org/10.1016/S1359-1789(02)00118-0

[20] G. M. Hancock, P. A. Hancock, and C. M. Janelle. 2012. The impact of emotions and predominant emotion regulation technique on driving performance. *Work* 41, Supplement 1 (Jan. 2012), 3608–3611. https://doi.org/10.3233/WOR-2012-0666-3608 Publisher: IOS Press.

[21] Neska El Haouij, Jean-Michel Poggi, Sylvie Sevestre-Ghalila, Raja Ghozi, and Mériem Jaïdane. 2018. AffectiveROAD system and database to assess driver's attention. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*. 800–803. https://doi.org/10.1145/3167132.3167395

[22] Mariam Hassib, Michael Braun, Bastian Pfleging, and Florian Alt. 2019. Detecting and Influencing Driver Emotions Using Psycho-Physiological Sensors and Ambient Light. In *Human-Computer Interaction – INTERACT 2019*, David Lamas, Fernando Loizides, Lennart Nacke, Helen Petrie, Marco Winckler, and Panayiotis Zaphiris (Eds.). Vol. 11746. Springer International Publishing, Cham, 721–742. https://doi.org/10.1007/978-3-030-29381-9_43 Series Title: Lecture Notes in Computer Science.

[23] Douglas Heaven. 2020. Why faces don't always tell the truth about feelings. *Nature* 578, 7796 (Feb. 2020), 502–504. https://doi.org/10.1038/d41586-020-00507-5 Bandiera_abtest: a Cg_type: News Feature Number: 7796 Publisher: Nature Publishing Group Subject_term: Psychology, Society, Computer science.

[24] Megan E. Hempel, Joanne E. Taylor, Martin J. Connolly, Fiona M. Alpass, and Christine V. Stephens. 2017. Scared behind the wheel: what impact does driving anxiety have on the health and well-being of young older adults? *International Psychogeriatrics* 29, 6 (June 2017), 1027–1034. https://doi.org/10.1017/S1041610216002271

[25] Myounghoon Jeon. 2016. Don't Cry While You're Driving: Sad Driving Is as Bad as Angry Driving. *International Journal of Human–Computer Interaction* 32, 10 (Oct. 2016), 777–790. https://doi.org/10.1080/10447318.2016.1198524

[26] O. Karaduman, H. Eren, H. Kurum, and M. Celenk. 2013. An effective variable selection algorithm for Aggressive/Calm Driving detection via CAN bus. In *2013 International Conference on Connected Vehicles and Expo (ICCVE)*. 586–591. https://doi.org/10.1109/ICCVE.2013.6799859

[27] Costas I. Karageorghis, Garry Kuan, William Payre, Elias Mouchlianitis, Luke W. Howard, Nick Reed, and Andrew M. Parkes. 2021. Psychological and psychophysiological effects of music intensity and lyrics on simulated urban driving. *Transportation Research Part F: Traffic Psychology and Behaviour* 81 (2021), 329–341. https://doi.org/10.1016/j.trf.2021.05.022

[28] Mohamed Kari, Tobias Grosse-Puppendahl, Alexander Jagaciak, David Bethge, Reinhard Schütte, and Christian Holz. 2021. *SoundsRide: Affordance-Synchronized Music Mixing for In-Car Audio Augmented Reality.* Association for Computing Machinery, New York, NY, USA, 118–133. https://doi.org/10.1145/3472749.3474739

[29] Won Hee Ko, Stefano Schiavon, Hui Zhang, Lindsay T. Graham, Gail Brager, Iris Mauss, and Yu-Wen Lin. 2020. The impact of a view from a window on thermal comfort, emotion, and cognitive performance. *Building and Environment* 175 (2020), 106779. https://doi.org/10.1016/j.buildenv.2020.106779

[30] Thomas Kosch, Mariam Hassib, Robin Reutter, and Florian Alt. 2020. *Emotions on the Go: Mobile Emotion Assessment in Real-Time Using Facial Expressions.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3399715.3399928

[31] Thomas Kosch, Andrii Matviienko, Florian Müller, Jessica Bersch, Christopher Katins, Dominik Schön, and Max Mühlhäuser. 2022. NotiBike: Assessing Target Selection Techniques for Cyclist Notifications in Augmented Reality. *Proc. ACM Hum.-Comput. Interact.* 6, MHCI, Article 197 (sep 2022), 24 pages. https://doi.org/10.1145/3546732

[32] Thomas Kosch, Robin Welsch, Lewis Chuang, and Albrecht Schmidt. 2022. The Placebo Effect of Artificial Intelligence in Human-Computer Interaction. *ACM Trans. Comput.-Hum. Interact.* (mar 2022). https://doi.org/10.1145/3529225 Just Accepted.

[33] Michael Kyte, Zaher Khatib, Patrick Shannon, and Fred Kitchener. 2000. Effect of environmental factors on free-flow speed. In *Fourth International Symposium on Highway Capacity*. Citeseer, 108–119.

[34] Tuan Le Mau, Katie Hoemann, Sam H Lyons, Jennifer Fugate, Emery N Brown, Maria Gendron, and Lisa Feldman Barrett. 2021. Professional actors demonstrate variability, not stereotypical expressions, when portraying emotional states in photographs. *Nature communications* 12, 1 (2021), 1–13. https://doi.org/10.1038/s41467-021-25352-6

[35] John D Lee and David L Strayer. 2004. Preface to the special section on driver distraction. *Human factors* 46, 4 (2004), 583–586. https://doi.org/10.1518/hfes.46.4.583.56811

[36] Shu Liu, Kevin Koch, Zimu Zhou, Simon Föll, Xiaoxi He, Tina Menke, Elgar Fleisch, and Felix Wortmann. 2021. The empathetic car: Exploring emotion inference via driver behaviour and traffic context. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–34. https://doi.org/10.1145/3478078

[37] Sascha Löbner, Frédéric Tronnier, Sebastian Pape, and Kai Rannenberg. 2021. Comparison of De-Identification Techniques for Privacy Preserving Data Analysis in Vehicular Data Sharing. In *Computer Science in Cars Symposium*. 1–11. https://doi.org/10.1145/3488904.3493380

[38] Marshall Long. 2014. 3 - Human Perception and Reaction to Sound. In *Architectural Acoustics (Second Edition)* (second edition ed.), Marshall Long (Ed.). Academic Press, Boston, 81–127. https://doi.org/10.1016/B978-0-12-398258-2.00003-9

[39] Andre Luckow, Ken Kennedy, Fabian Manhardt, Emil Djerekarov, Bennie Vorster, and Amy Apon. 2015. Automotive big data: Applications, workloads and infrastructures. In *2015 IEEE International Conference on Big Data (Big Data)*. IEEE, 1201–1210. https://doi.org/10.1109/BigData.2015.7363874

[40] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*. 4768–4777.

[41] Jolieke Mesken, Marjan P Hagenzieker, Talib Rothengatter, and Dick De Waard. 2007. Frequency, determinants, and consequences of different drivers' emotions: An on-the-road study using self-reports,(observed) behaviour, and physiology. *Transportation research part F: traffic psychology and behaviour* 10, 6 (2007), 458–475. https://doi.org/10.1016/j.trf.2007.05.001

[42] Mimma Nardelli, Gaetano Valenza, Alberto Greco, Antonio Lanata, and Enzo Pasquale Scilingo. 2015. Recognizing Emotions Induced by Affective Sounds through Heart Rate Variability. *IEEE Transactions on Affective Computing* 6, 4 (2015), 385–394. https://doi.org/10.1109/TAFFC.2015.2432810

[43] Meital Navon and Orit Taubman Ben-Ari. 2019. Driven by emotions: The association between emotion regulation, forgivingness, and driving styles. *Transportation Research Part F: Traffic Psychology and Behaviour* 65 (Aug. 2019), 1–9. https://doi.org/10.1016/j.trf.2019.07.005

[44] Michael Oehl, Felix W. Siebert, Tessa-Karina Tews, Rainer Höger, and Hans-Rüdiger Pfister. 2011. Improving Human-Machine Interaction – A Non Invasive Approach to Detect Emotions in Car Drivers. In *Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments (Lecture Notes in Computer Science)*, Julie A. Jacko (Ed.). Springer, Berlin, Heidelberg, 577–585. https://doi.org/10.1007/978-3-642-21616-9_65

[45] Rajesh Paleti, Naveen Eluru, and Chandra R. Bhat. 2010. Examining the influence of aggressive driving behavior on driver injury severity in traffic crashes. *Accident; Analysis and Prevention* 42, 6 (Nov. 2010), 1839–1854. https://doi.org/10.1016/j.aap.2010.05.005

[46] Pablo E. Paredes, Francisco Ordonez, Wendy Ju, and James A. Landay. 2018. Fast & Furious: Detecting Stress with a Car Steering Wheel. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–12. https://doi.org/10.1145/3173574.3174239

[47] Emanuel Parzen. 1962. On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics* 33, 3 (1962), 1065 – 1076. https://doi.org/10.1214/aoms/1177704472

[48] Esther Ramdinmawii, Abhijit Mohanta, and Vinay Kumar Mittal. 2017. Emotion recognition from speech signal. In *TENCON 2017-2017 IEEE Region 10 Conference*. IEEE, 1562–1567.
[49] Alicia F Requardt, Klas Ihme, Marc Wilbrink, and Andreas Wendemuth. 2020. Towards affect-aware vehicles for increasing safety and comfort: recognising driver emotions from audio recordings in a realistic driving study. *IET Intelligent Transport Systems* 14, 10 (2020), 1265–1277. https://doi.org/10.1049/iet-its.2019.0732
[50] James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161–1178. https://doi.org/10.1037/h0077714 Place: US Publisher: American Psychological Association.
[51] B. Scott-Parker. 2017. Emotions, behaviour, and the adolescent driver: A literature review. *Transportation Research Part F: Traffic Psychology and Behaviour* 50 (2017), 1–37. https://doi.org/10.1016/j.trf.2017.06.019
[52] Maria Seitz, Thomas J. Daun, Andreas Zimmermann, and Markus Lienkamp. 2013. Measurement of Electrodermal Activity to Evaluate the Impact of Environmental Complexity on Driver Workload. In *Proceedings of the FISITA 2012 World Automotive Congress*. Springer Berlin Heidelberg, Berlin, Heidelberg, 245–256. https://doi.org/10.1007/978-3-642-33838-0_22
[53] Luke Stark. 2016. The emotional context of information privacy. *The Information Society* 32 (01 2016), 14–27. https://doi.org/10.1080/01972243.2015.1107167
[54] Luke Stark. 2019. Facial recognition is the plutonium of AI. *XRDS: Crossroads, The ACM Magazine for Students* 25, 3 (2019), 50–55. https://doi.org/10.1145/3313129
[55] Rainer Steffen, Richard Bogenberger, Joachim Hillebrand, Wolfgang Hintermaier, Andreas Winckler, and Mehrnoush Rahmani. 2008. Design and realization of an ip-based in-car network architecture. In *Proceedings of the First Annual International Symposium on Vehicular Computing Systems (ISVCS 2008)*. https://doi.org/10.4108/ICST.ISVCS2008.3543
[56] Alejandro A. Torres-García, Omar Mendoza-Montoya, Marta Molinas, Javier M. Antelis, Luis A. Moctezuma, and Tonatiuh Hernández-Del-Toro. 2022. Chapter 4 - Pre-processing and feature extraction. In *Biosignal Processing and Classification Using Computational Learning and Intelligence*, Alejandro A. Torres-García, Carlos A. Reyes-García, Luis Villaseñor-Pineda, and Omar Mendoza-Montoya (Eds.). Academic Press, 59–91. https://doi.org/10.1016/B978-0-12-820125-1.00014-2
[57] Xiaoyuan Wang, Yongqing Guo, Jeff Ban, Qing Xu, Cheng-Lin Bai, and Shanliang Liu. 2020. Driver Emotion Recognition of Multiple-ECG Feature Fusion based on BP Network and D-S Evidence. *IET Intelligent Transport Systems* 14 (03 2020). https://doi.org/10.1049/iet-its.2019.0499
[58] Stephen Westland, Yuan Li, Dabo Guan, Yanni Yu, P Wang, Xuejun Wang, Kebin He, Shu Tao, and Jing Meng. 2019. A psychophysical measurement on subjective well-being and air pollution. *Nature Communications* 10 (11 2019). https://doi.org/10.1038/s41467-019-13459-w
[59] Changsheng Xu, Namunu C Maddage, Xi Shao, Fang Cao, and Qi Tian. 2003. Musical genre classification using support vector machines. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, Vol. 5. IEEE, V–429.
[60] Syeda Sana e Zainab and Tahar Kechadi. 2019. Sensitive and Private Data Analysis: A Systematic Review. In *Proceedings of the 3rd International Conference on Future Networks and Distributed Systems* (Paris, France) *(ICFNDS '19)*. Association for Computing Machinery, New York, NY, USA, Article 12, 11 pages. https://doi.org/10.1145/3341325.3342002
[61] Sebastian Zepf, Monique Dittrich, Javier Hernandez, and Alexander Schmitt. 2019. Towards Empathetic Car Interfaces: Emotional Triggers while Driving. *CHI EA '19: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–6. https://doi.org/10.1145/3290607.3312883
[62] Sebastian Zepf, Javier Hernandez, Alexander Schmitt, Wolfgang Minker, and Rosalind W. Picard. 2020. Driver Emotion Recognition for Intelligent Vehicles: A Survey. *Comput. Surveys* 53, 3 (July 2020), 1–30. https://doi.org/10.1145/3388790

A APPENDIX

A.1 Confusion Matrices

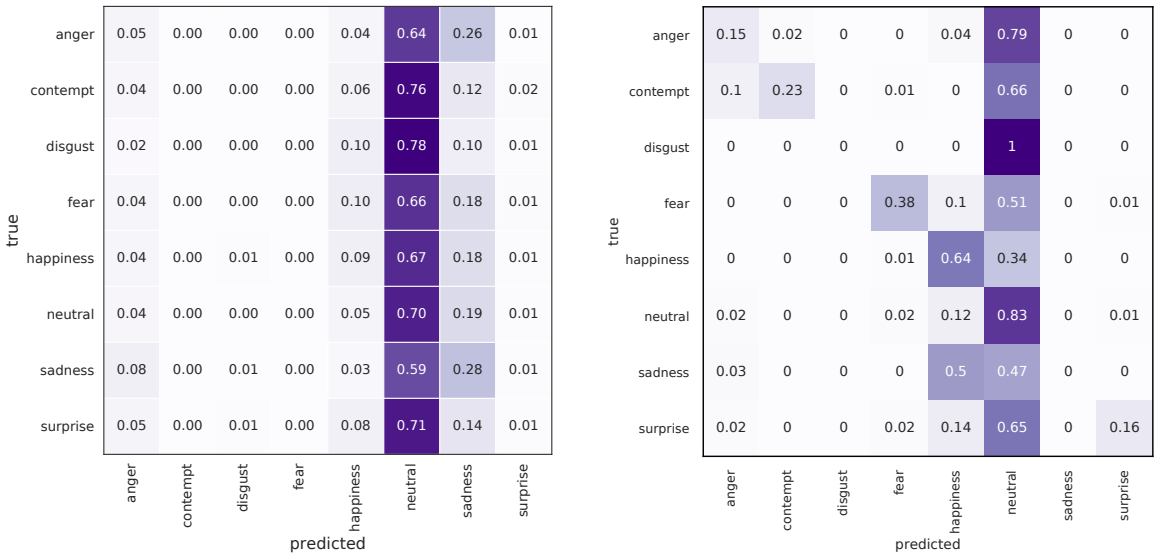
The confusion matrices of the facial expression baseline and the Random Forest predictor is shown in Fig 10.

A.2 Input Feature Correlation

The Pearson correlation of the input features of our system is shown in Figure 11.

A.3 Neural Network Architecture

The specification of the used neural network architecture is explained in the following section. We employ a feedforward fully-connected neural network, with two hidden layers and one output layer. The first layer contains 100 neurons, the second layer contains 50 neurons, and both use a relu activation function. The output layer



(a) Confusion matrix comparing facial expression recognition output (FERPlus model) against true labels.

(b) Confusion matrix comparing the results of the Random Forest prediction model against true labels.

Fig. 10. Confusion matrices of the facial expression baseline (a) and (b) the Random Forest trained on the participants multi-domain contextual data. The values in the matrix are normalized on the true emotion class occurrences. (a) Only 5% of 'anger' emotional states are recognized by the facial expression classifier and 64% of all true 'anger' emotions are falsely predicted as being 'neutral'. (b) The model is participant-dependent trained model. Although, the model overpredict 'neutral' states, the random model performs significantly better in predicting 'fear' (38%), 'happiness' (64%), and 'neutral' (83%) states.

outputs one-hot-encoded emotion labels using a sigmoid function. The network is trained with a batch size of 64 for 3000 epochs using the adam optimizer for backpropagation. We employ early stopping criteria from avoiding overfitting after a waiting period of 50 epochs. To counterbalance the imbalance of classes, we assign a loss weight according to the inverse frequency of class observation to the categorical cross-entropy loss optimization function. We report the neural network performance in Table 2.

A.4 Person-Dependent Modeling

We analyzed participant-dependent modeling using a participant-dependent Leave-One-of-10-Road-Segments-Out cross-validation. This setting denotes that we are training a participant-dependent model and validating the participant's holdout set using a 10-fold cross-validation scheme. In general, participant-dependent models can adapt to specific persons and provide a possibly more privacy-aware and personal emotion predictor as data is not shared globally. However, each person-dependent model has only limited training data available, so longer drive durations are needed to reach a satisfactory recognition performance. We acquired multiple sessions for some participants to circumvent the issue of having too little data on individual participants. We did not employ a leave-one-session cross-validation as not every participant acquired driving data in multiple sessions. From the last four rows in the emotion recognition performance table, we see that the overall recognition performance of the models using our features is best. The combined classification model with all input features reaches an accuracy of 66% and an F_1 score of 67%. Therefore this model is significantly better than the baselines and the

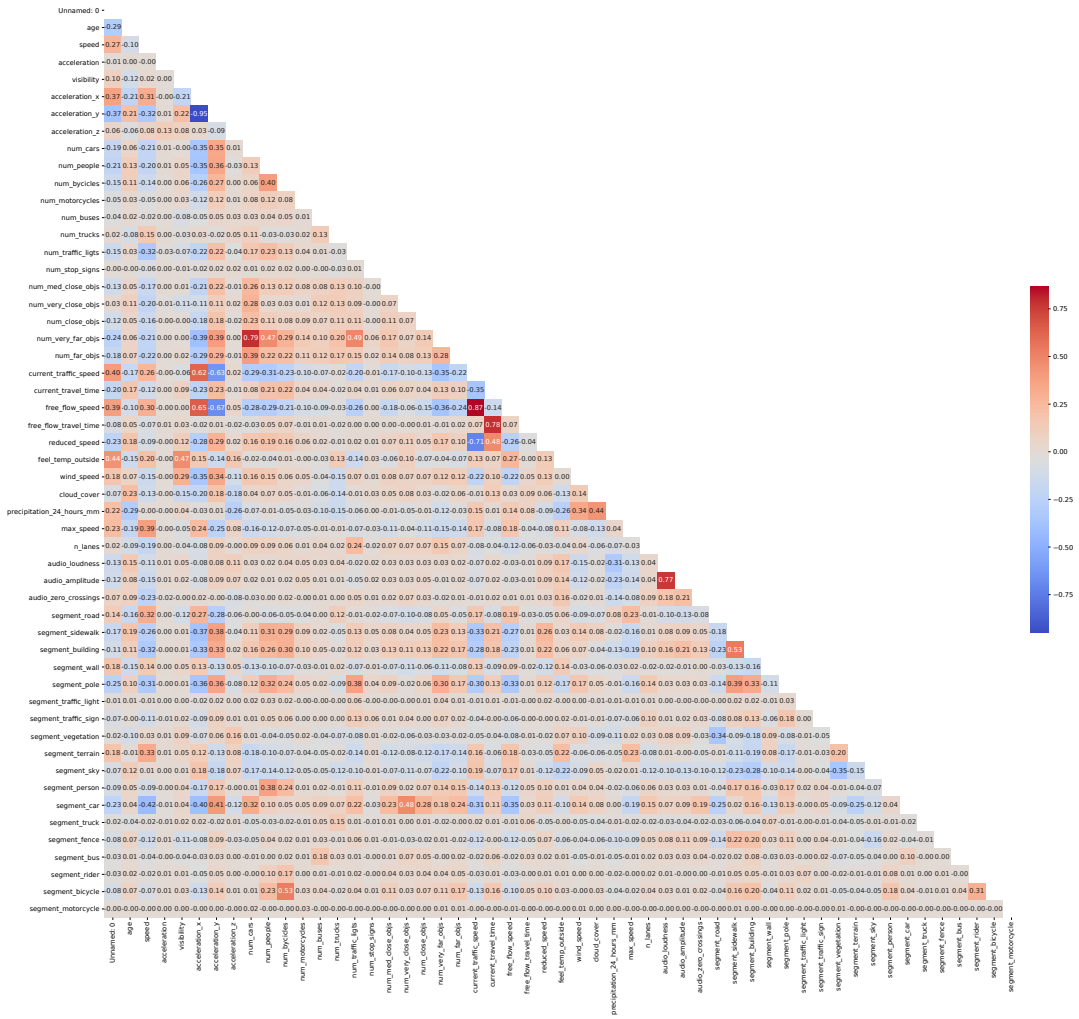


Fig. 11. Correlation matrix of the available features of our system. High positive correlations are depicted in red and high negative correlations are shown in deep blue. The feature 'segment_train' has no variance and is therefore left blank, since the segmentation module has not detected any trains in the in-the-wild driving.

performances of global modeling procedures. The audio-visual feature set predicts emotions confidently with 62% accuracy (F_1: 62%), whereas GPS-sensor-only extracted features are able to predict subjective emotions with 65% accuracy (F_1: 57%).

Table 4. Random Forest Evaluation Results. We report the averaged evaluation results for each of the feature groups across all evaluation steps: global classifier learning leave-one-participant-out evaluation and learning participant-dependent models. Accuracy, class-weighted precision, unweighted average recall (UAR) and F_1 scores. Values are averages from the 10-fold cross-validation (best values are indicated in bold).

	Participant-Dependent (Leave-One-Road-Segment-Out)			
	Accuracy	**Precision**	**UAR**	F_1
Facial Expressions (FERPlus)	.33 ± .0	.41 ± .09	.23 ± .04	.36 ± .09
Facial Expressions (Azure)	.39 ± .04	.42 ± .04	.27 ± .0	.38 ± .03
VEmotion (VE)	.65 ± .04	.63 ± .04	.59 ± .01	.64 ± .03
Visual Complexity Segmentation (VC-Seg.)	.6 ± .04	.50 ± .04	.38 ± .01	.62 ± .03
Visual Complexity - Object Detection (VC-ObjD.)	.51 ± .06	.51 ± .06	.27 ± .01	.56 ± .03
VC-Seg. + Audio	.61 ± .04	.51 ± .04	.39 ± .01	.61 ± .03
VC-ObjD. + Audio	.58 ± .04	.42 ± .04	.32 ± .01	.61 ± .02
Audio only	.52 ± .02	.39 ± .02	.28 ± .01	.57 ± .02
Audiovisual (OjbD. + Seg. + Audio)	.62 ± .03	.52 ± .03	.43 ± .01	.62 ± .02
GPS-infered features only	.65 ± .04	.63 ± .04	.57 ± .11	.57 ± .01
All features	**.66 ± .03**	**.65 ± .03**	**.6 ± .01**	**.67 ± .03**