The Multimodal Driver Monitoring Database: A Naturalistic Corpus to Study Driver Attention

Sumit Jha, *Graduate Student Member, IEEE*, Mohamed F. Marzban, *Graduate Student Member, IEEE*, Tiancheng Hu, *Student Member, IEEE*, Mohamed H. Mahmoud, *Graduate Student Member, IEEE*, Naofal Al-Dhahir, *Fellow, IEEE*, and Carlos Busso, *Senior Member, IEEE*

Abstract—A smart vehicle should be able to monitor the actions and behaviors of the human driver to provide critical warnings or intervene when necessary. Recent advancements in deep learning and computer vision have shown great promise in monitoring human behavior and activities. While these algorithms work well in a controlled environment, naturalistic driving conditions add new challenges such as illumination variations, occlusions, and extreme head poses. A vast amount of in-domain data is required to train models that provide high performance in predicting driving related tasks to effectively monitor driver actions and behaviors. Toward building the required infrastructure, this paper presents the *multimodal driver monitoring* (MDM) dataset, which was collected with 59 subjects that were recorded performing various tasks. We use the Fi-Cap device that continuously tracks the head movement of the driver using fiducial markers, providing frame-based annotations to train head pose algorithms in naturalistic driving conditions. We ask the driver to look at predetermined gaze locations to obtain accurate correlation between the driver's facial image and visual attention. We also collect data when the driver performs common secondary activities such as navigation using a smart phone and operating the in-car infotainment system. All of the driver's activities are recorded with high definition RGB cameras and a time-of-flight depth camera. We also record the *controller area network-bus* (CAN-Bus), extracting important information. These high quality recordings serve as the ideal resource to train various efficient algorithms for monitoring the driver, providing further advancements in the field of in-vehicle safety systems.

Index Terms—In-vehicle safety, driver visual attention, driver monitoring dataset.

I. INTRODUCTION

THE *National Highway Traffic Safety Administration* (NHTSA) reported that in 2019, around 36,120 people died in car accidents across the United States [1]. Driver distractions and human errors are the leading causes of these accidents. With the growing use of in-vehicle technologies and hand-held devices, drivers are exposed to more distractions, leading to more accidents. Artificial intelligence-based driver behavior monitoring systems can play a key role in modeling and understanding driver distractions, triggering timely alarms to save humans lives. In addition, they can decide when to take over the vehicle's control from the driver in levels 3 and 4 of autonomous driving. These driver monitoring systems need to be trained with an annotated dataset that mimics practical driving environments.

Recent advancements in the field of computer vision and deep learning have attracted increased research interests in the field of human-centered computing. Facial images can be used to extract multiple information such as identity [2], [3], emotions [4], and gaze [5], [6]. While in theory these solutions should be directly applicable for developing in-vehicle safety systems, in practice, most of these systems fail to work in a driving environment due to additional challenges, such as occlusions and illumination changes [7]. Since most recent technologies based on *deep learning* (DL) rely on in-domain data with ground truth annotations, naturalistic driving recordings collected from multiple drivers and with appropriate labels are needed to advance in-vehicle safety systems. This requirement is particularly important if the focus is to analyze the driver's visual attention in realistic driving environments. Driver's gaze is one of the most crucial features for modeling and classifying driver's distractions. However, estimating the driver's gaze in naturalistic driving environments is a challenging task because of regular illumination changes and common occlusions, especially from eye glasses. To tackle these challenges, previous studies have utilized the driver's head pose to obtain a coarse gaze estimation, which is relatively easy to estimate [8]–[11]. Other studies have included eye features to capture fine gaze changes [12], [13]. Understanding and capturing the interplay between the head pose and the eye movements is essential for any driver behavior monitoring system and any solid gaze estimation algorithm [14]. Studies have used gaze location to determine the driver behavior patterns while operating a car (e.g., what they look at when driving by intersections) [15], [16].

This study presents the *multimodal driver monitoring* (MDM) database, which was specifically designed to study the driver's visual attention. The study addresses some of the challenges of existing driver monitoring datasets, including their limited size and narrow scope focusing solely on either

the driver's head pose or the driver's gaze zones. The MDM corpus is a new multimodal dataset that considers both head pose and gaze directions to facilitate learning this complex interplay between the head pose, eye movements, and the resulting gaze. We achieve this goal by providing frame-based information about head pose, and spatial positions for target markers inside and outside the vehicle that were glanced at by the drivers. The corpus is multimodal, where we use multiple RGB cameras and a depth camera to record the drivers. We also have a camera recording the road. We have collected 59 different gender-balanced subjects spanning different ethnic groups. Our dataset captures primary (e.g., mirror checking, lane keeping, left/right lane changes, and left/right turns) and secondary activities (e.g., navigation using Google maps, and operating the radio). In addition, we ask the drivers to look at predefined targets inside and outside the vehicle, providing temporal gaze information. These target markers are placed on the windshield, mirrors, dashboard, radio, and gear shifter. We also ask the driver to look at other vehicles, billboards, buildings, and street signs. In addition, we collect continuous gaze annotations by asking the driver to look at a moving fiducial marker outside while the vehicle is parked. We design a systematic approach to obtain the driver's head pose per-frame annotation using the Fi-Cap helmet [11]. Fi-Cap is a helmet with multiple fiducial markers that are easy to detect. The helmet is worn on the back of the head, avoiding facial occlusions when the driver is recorded with frontal cameras. We provide manual annotations using the ELAN toolkit [17], describing gaze information and the secondary activities conducted during the recordings.

This study provides an initial analysis of the MDM corpus, showing that our dataset covers a wide range of head poses in all three rotation axes (pitch, yaw, and roll). The analysis reflects the diversity of the ground truth head pose obtained by including the primary and secondary activities in our protocol. Furthermore, we evaluate the accuracy of the head pose annotations provided by the Fi-Cap helmet as a function of the number of detected fiducial markers. Moreover, we present some preliminary analysis on head pose estimation using depth cameras, and gaze estimation using the RGB camera, demonstrating the potential of using multiple sensors for vision-based tasks inside the vehicle. In summary, the main features of this paper are:

- A multimodal database with drivers performing activities in naturalistic driving conditions.
- Continuous and discrete instances of drivers' faces associated with available ground truth gaze.
- Continuous tracking of the head pose of the driver with reliable ground truth using the Fi-Cap helmet.
- Multiple sensors to capture high quality RGB videos, point cloud data, CAN-Bus information, and audio recordings.

The rest of this paper is organized as follows. Section II describes previously published databases for in-vehicle driver monitoring systems. Section III presents the data collection protocol, describing the annotation process of various activities and gaze events. Section IV shows how we utilize the Fi-Cap helmet to annotate the driver's head pose on a frame-by-frame basis. This section also analyzes the reliability in the frame-based head pose annotations as a function of the number of fiducial markers. Section V analyzes the head pose and gaze angle distributions in our dataset. This analysis section also briefly describes a method to estimate the driver's head pose using the depth cameras, and a method to estimate gaze using RGB frames. Section VI concludes the paper by summarizing the main features of our corpus.

II. RELATED WORK

The research community has made several efforts to construct high-quality and diverse driving datasets, especially those focusing on road scenes. Examples include the BDD100K corpus [18], [19], Mapillary Vistas Dataset [20], Waymo Open Dataset [21], and nuScenes dataset [22]. However, the MDM dataset focuses on the drivers, not the road scenes. While there has been interest in different aspects of driver monitoring such as visual attention [23], emotions [24], and driving anomalies [25], [26], we will review relevant datasets that have focused on head pose estimation (Sec. II-A), and gaze estimation (Sec. II-B). This section also reviews recent efforts to collect long-term driving datasets (Sec. II-C).

A. Databases Focusing on Head Pose Estimation

Databases focusing on head pose estimation aim to provide frame-based annotations for the driver's head pose [27]–[32], which has a wide range of applications including coarse gaze estimation, driver behavior modeling, and *human-computer interaction* (HCI) for entertainment purposes. Table I summarizes the main databases created to estimate head pose of the driver.

Martin *et al.* [27] presented the Lisa-P dataset, which is a naturalistic driving dataset collected on the road. The recordings included a motion capture system (marker and camera) at the back of the driver to track the head movement. The authors placed a camera on the dashboard to capture the driver's face. The corpus provides annotations for roll, yaw, and pitch head rotation of the driver extracted with the motion capture system. Additionally, the corpus provides manual annotations for seven key facial features (eye corners, nose corners, and nose tips) for some video sequences every 5 to 10 frames. This dataset contains 14 video sequences with an average of 14,281 frames per video, at 30 *frames per second* (fps) (i.e., about 111 minutes of recordings in total). The resolution of the camera is 640 × 480.

The CoHMEt database [28] offers naturalistic driving recordings from three RGB cameras. The first camera was mounted on the car frame between the front mirror and the windshield. The second camera was placed on the front windshield. The third camera was mounted near the rear view mirror. The cameras collected images with a resolution of 640 × 480 at 30 fps. The corpus included two *inertial measurement units* (IMU), one on the driver's head and the other fixed on the car. The difference between the two IMU results was used as the label for the roll, yaw, and pitch head rotations. The IMUs are reset around every 10 seconds to

TABLE I

A COMPARISON OF EXISTING DRIVING DATABASES FOCUSING ON HEAD POSE ESTIMATION

Dataset Year	Recording Conditions	Sensors	Annotation provided	# Subjects M/F	Length of data # Frames
Lisa-P [27] 2012	Day and night driving	RGB-face(640x480,30fps) RGB-back(640x480,30fps) Mo-Cap	Head orientation 7 facial landmarks	14	1.85 hours 199,934
CoHMEt [28] 2014	Daytime, suburban street, highway	RGB-face(640x480,30fps) RGB-2sides(640x480,30fps) IMU	Head orientation	-	0.28 hours 30231
DriveAHead [29] 2017	Daytime, sunny, rainy and foggy small town, highway, and parking	depth(TOF) (512x424) IR camera (512x424) Mo-Cap	Head pose binary occlusions glasses, sunglasses, occlusion	20 16/4	- 1M frames
DD-Pose [30] 2019	Daytime, parked, highway, cities look at markers, make phone calls, read shop name, interact with pedestrians	RGB-back Depth(stereo)-face(2048x2048) Mo-Cap CAN-BUS	Head pose occlusion label steering wheel info vehicle motion info	27 21/6	- 330k
Pandora [31] 2017	Simulator constrained and unconstrained movement	RGB-face(1920x1080) depth-face(512x424) IMU	Head orientation shoulder skeleton joints shoulder pose	22 12/10	- 250k
AutoPose [32] 2020	Simulator look at marker constrained and unconstrained movement	RGB-full body(1920x1080,30fps) Depth-full body(512x424,30 fps) Infrared-full body(512x424,30fps) Infrared-face(512x424, 60fps) Mo-Cap	Head pose 6 gaze zone driver activity glasses yes/no glass type	21 10/11	4.63 hours(IR) + 1.46 hours(Kinect) 1M(IR) + 316k(Kinect)
MDM corpus 2021	Parked, daytime driving, look at markers, read road landmarks, change radio station, navigation on phone	RGB-face(1920x1080,60fps) RGB-back(1920x1080,60fps) RGB-mirror (1920x1080,60fps) RGB-road(1920x1080,60fps) depth(ToF)-face (171x224,45 fps) grayscale-IR LED(171x224,45 fps) Fi-Cap microphone array(5) CAN-BUS	Head pose 21 gaze targets vehicle info audio	59 27/32	50.23 hours

avoid drifts in the measurements. There are 30,231 frames in total in this dataset, which translates to roughly 17 minutes of recording. The dataset includes driving scenes on the streets and freeways near the University of California at San Diego.

The DriveAHead [29] is a more recent driving dataset. The recordings include videos using a Kinect V2 sensor, which provides infrared and depth captures at a resolution of 512 × 424. A 3D motion capture system was used to track the position and orientation of the driver's head. In addition, the authors provided binary annotation of three types of occlusions: hair, eye glasses, and self-occlusion. The database consists of 21 video sequences from 20 unique subjects (4 males, 16 females), each performing parking maneuvers, driving on highways and driving in a small town. The data collection took place in sunny, rainy, and foggy weather conditions. In total, this dataset has about 1 million frames.

The *Daimler TU Delft Driver Head Pose* (DD-Pose) Benchmark [30] utilizes one high-resolution stereo camera (2,048 × 2,048) capturing the driver's face, and a wide-angle RGB camera recording from the backside of the driver. The authors used a 3D motion capturing system and provide rotation and translation labels for the driver's head. The corpus also provides occlusion labels for each frame (none, partial, and full occlusion). In addition, the corpus provides steering

wheel and vehicle motion information. This dataset includes both naturalistic driving as well as induced driving conditions. Naturalistic driving conditions consist of driving recordings in highways and big German cities, with complex traffic scenarios. The drivers were asked to read names of the shops along the street while driving. During the induced driving conditions, the car was parked. The driver was instructed to look at a series of target points in the car. The DD-Pose dataset comprises of 27 subjects (21 males, 6 females) with a total of 330k frames.

Some of the databases for head pose estimation were recorded in physical simulators. The Pandora Database [31] uses a Kinect One camera mounted on the simulated dashboard. The resolution of the RGB camera was set to 1,920 × 1,080, while the resolution for the depth camera was set to 512 × 424. In this dataset, drivers were asked to perform driving-like maneuvers such as looking at the rear-view mirror and holding the steering wheel. Subjects were first asked to rotate their heads along one axis at a time. Then, they were instructed to freely move their heads. To create more variations in head pose and shoulder pose, subjects wore eye glasses, sun glasses, scarves, caps, and used smartphones. The corpus provides labels for the head and shoulder pose. An IMU on the back of the subject's head was used to record the ground

TABLE II

A COMPARISON OF EXISTING DRIVING DATABASES FOCUSING ON GAZE ESTIMATION

Dataset Year	Recording Conditions	Sensors	Annotation provided	# Subjects M/F	Length of data # Frames
Fridman et al. [33] 2016	Highway, daytime, voice control interface, select phone number	RGB-face(800x600,30fps)	6 gaze zone	50	17.23 hours 1.86M
DAD [25] 2019	Naturalistic driving suburban, urban and highway	RGB-road(1600x1200,30fps) RGB-face(1600x1200,30fps) CAN-BUS eye tracking glasses physiological sensors	4-layer driver behavior representation	4	250 hours
DR(eye)VE [23] 2019	Day and night different weather conditions countryside, downtown and highway	RGB-head mounted(1280x720,30fps) road(1920x1080,25fps)	Continuous gaze map GPS,vehicle speed,course	8 7/1	6 hours 0.56M
DG-Unicamp [34] 2019	Parked, day and night look at markers	RGB-face(320x240,30fps) depth(stereo)-face(320x240,30 fps) Infrared-face(320x240,30 fps)	18 gaze zone	45 35/10	12 hours 1M
DGW [35] 2020	Parked, day and night look at markers	RGB-face	9 gaze zone	338 247/91	N/A
AutoPose [32] 2020	Simulator look at marker constrained and unconstrained movement	RGB-full body(1920x1080,30fps) Depth-full body(512x424,30 fps) Infrared-full body(512x424,30fps) Infrared-face(512x424, 60fps) Mo-Cap	Head pose 6 gaze zone driver activity glasses yes/no glass type	21 10/11	4.63 hours(IR) + 1.46 hours(Kinect) 1M(IR) + 316k(Kinect)
MDM corpus 2021	Parked, daytime driving, look at markers, read road landmarks, change radio station, navigation on phone	RGB-face(1920x1080,60fps) RGB-back(1920x1080,60fps) RGB-mirror (1920x1080,60fps) RGB-road(1920x1080,60fps) depth(ToF)-face (171x224,45 fps) grayscale-IR LED(171x224,45 fps) Fi-Cap microphone array(5) CAN-BUS	Head pose 21 gaze targets vehicle info audio	59 27/32	50.23 hours

truth for the head pose. In total, the dataset has 250k frames. Another dataset collected in a car simulator is the AutoPose corpus [32]. This corpus relies on one dashboard IR camera running at 60 fps, and a Kinect V2 camera (RGB, Depth, and Infrared), mounted at the rear view mirror, running at 30 fps. There are 21 subjects in this dataset. The recording protocol included a series of tasks such as performing head rotations and looking at markers at various locations. Each task was performed in three conditions: without eye glasses, with eye glasses, and with sun glasses. Head rotations were performed with and without the subject wearing a scarf. Due to interference of the two cameras, each subject performed the entire protocol twice, recording the session with either the IR camera or the Kinect V2 camera. Head pose labels are provided using an OptiTrack Motion Capture System. Gaze labels associated with the markers are also provided. No primary driving activities were performed in this dataset due to its simple setup consisting of just a seat and a steering wheel, without a computer monitor simulating the road scene. In total, there are 1,018,885 IR images and 316,497 images from Kinect. We note that methods with physical simulation only partially replicate a naturalistic driving environment. Most systems lack feedback from the roads such as vibrations and inertia after maneuvers. Additionally, the illumination and background environment may not be realistic. Most importantly, the chances of making errors while driving are not high, regardless of how realistic the simulator is. Therefore, the behaviors of the drivers may not be exactly the same as the

ones observed in real driving conditions (e.g., the subject may look away for longer glances in physical simulators, without worrying about real accidents).

Compared to the aforementioned driver head pose estimation dataset, our proposed MDM has several features that make this corpus an ideal platform to build machine learning algorithms to track head pose in a vehicle: (1) it includes data when the driver is operating the car and when the car is parked, (2) it provides complementary multimodal sensors (four different RGB, depth, grayscale-IR LED, microphone array, CAN-Bus), (3) it has frame-based head pose information obtained with a simple setting based on fiducial markers, and (4) it is recorded from 59 gender balanced subjects.

B. Databases Focusing on Gaze Estimation

Driver gaze monitoring databases often provide ground truth location for regions where the drivers are asked to glance [23], [34], [36], which can be used to train different gaze estimation algorithms. Tables II summarizes the main databases created to estimate the gaze of the driver.

Fridman *et al.* [33] constructs a naturalistic driving dataset. The subjects were asked to drive on a highway, and perform secondary tasks such as making phone calls and entering addresses into a navigation system using a voice-control feature. The driver's face was recorded using a resolution of 800 × 600 at 30 fps. The dataset includes recordings of 1,860,761 frames from 50 subjects (17h, 14m). The ground

truth of the driver's gaze is manually labeled as one of six predefined gaze zones.

The DR(eye)VE Project [23] features data from both an ego-centric view of the road (from the eye-tracking glasses), and a car-centric view (a camera capturing the road). This dataset is somewhat different from many other gaze datasets in that it provides gaze data from the ego-centric view. The eye-tracking data was recorded at 60 fps, while the camera on the eye-tracking glasses acquired data at a resolution of 1280 × 720 at 30 fps. The road camera captured data at a resolution of 1920 × 1080 at 25 fps. In this dataset, eight subjects were asked to drive in various scenes including countryside, downtown, and highway, during different hours during the day (both day and night), and in different weather conditions. This dataset provides ground truth gaze in the form of a continuous gaze map, derived from the eye-tracking glasses. In addition, the corpus provides *global positioning system* (GPS) information and vehicle speed. In total, there are 555,000 frames in this dataset (roughly 6 hours).

Another interesting database is the *driving anomaly dataset* (DAD) [25]. The corpus was collected in a real car in an Asian city. The database involves two cameras, one capturing the road and the other capturing the driver's face. The database also includes eye tracking glasses to obtain the gaze of the driver. Additionally, the vehicle information was collected from the CAN-Bus. The drivers were asked to use wearable devices, which captured *heart rate* (HR), *breath rate* (BR), and *electrodermal activity* (EDA). One of the key strengths of this corpus is the detailed annotations, providing a 4-layer driver behavior representation: goal-oriented action (e.g., right turn), stimulus-driven action (e.g., stop), cause (e.g., a stopped car in front of ego car), and attention (e.g., pedestrian near ego lane). In total, the corpus has about 250 hours, where 120 hours have been annotated.

Ribeiro and Costa [34] presented the DG-Unicamp corpus, which is a driving dataset that focuses on gaze zone estimation. It uses an Intel Realsense R200 camera placed on top of the dashboard. The camera provides aligned RGB and depth recordings as well as two sources (from stereo) of infrared recordings at 30 fps, with a resolution of 320 × 240. This corpus divided the gaze direction into 18 gaze zones, providing the corresponding gaze labels. The recordings were collected in a parked car due to safety concerns. The recordings were collected in different locations, light conditions (both day and night) at different hours to ensure maximum diversity in the recordings. Each subject was asked to look at each gaze zone for 10s and then move to the next gaze zone. The corpus consists of 45 subjects (35 males, 10 females), where five subjects were recorded twice. In total, the corpus has about 1 million frames (12 hours). A similar corpus is the *Driver Gaze in the Wild* (DGW) Dataset [35], which is a gaze zone classification dataset that features 338 subjects with age between 18 and 63 years old. In this dataset, subjects were asked to look at markers in different locations in a parked car at different locations in a university. The recording was collected at different times of the day and at night. During the recording, the driver was asked to fixate on one of the nine gaze markers and say its number.

TABLE III
A COMPARISON OF LARGE-SCALE NATURALISTIC DRIVING DATABASES FOR VISUAL ATTENTION

Dataset Year	Sensors	# Subjects	Study Duration
NHTSA [37] 2005	5 Cameras CAN-BUS Doppler Radar IMU GPS	241	18 months
SHRP 2 [38] 2012	4 Cameras CAN-BUS Radar IMU GPS	2360	7 years
MIT-AVT [36] 2019	3/4 Cameras CAN-BUS IMU GPS	122	Ongoing

They directly labeled the gaze zone using an *automatic speech recognition* (ASR) system.

We note that many gaze datasets are collected in parked vehicle conditions. We have observed that gaze behaviors are different from the patterns observed while the driver is operating a vehicle [14]. The driver is cognitively busy with primary driving tasks, resulting in more eye movements and shorter glances. Additionally, naturalistic driving conditions pose extra challenges due to sudden illumination changes and occlusion of the driver's face. These variations create a domain-shift problem for machine learning algorithms to be deployed in naturalistic driving. In the MDM corpus, we address these challenges by recording our database while the car is parked and while the driver is operating the vehicle. Likewise, compared to other gaze estimation datasets, our proposed MDM corpus includes multiple modalities (four different RGB, depth, infrared, microphone array, CAN-BUS), and provides fine-grained labels (21 gaze targets during driving condition; continuous label during parked condition).

C. Large-Scale Naturalistic Databases for Visual Attention

This section presents large-scale naturalistic databases that can be used to analyze the driver visual attention. Table III summarizes the main databases in this area. The NHTSA conducted a large-scale naturalistic driving study [37], which involved 241 drivers with 100 cars equipped with various sensors, conducted over the course of 18 months. In this study, the subjects were asked to drive naturally in cars with sensors. These sensors included five cameras capturing the driver's face, driver side road view, passenger side road view, road, and over-the-shoulder view for the driver's hands, and surrounding areas. The recordings also included the CAN-Bus, a Doppler radar, an accelerometer, a gyroscope, and a GPS. The equipment also included a toolbox for drivers to identify accidents. NHTSA collected approximately 43,000 hours of data, which included 82 total collisions. In each accident, Neale *et al.* [37] identified the pre-event maneuver, precipitating factor, event type, contributing factors, associative factors, and avoidance maneuver.

The SHRP 2 Naturalistic Driving Study [38] is the largest long-term naturalistic driving study. The goal of this study was to understand how drivers interact with and adapt to the environment, including the vehicle, the traffic environment, roadway characteristics, and traffic control devices. The database includes nearly 2,360 subjects from different parts of the U.S. The experimental vehicles had four cameras, recording at 15 fps, capturing the driver, road, instrument cluster, and road in the back. In addition, a single camera captures still images of passengers every few seconds. The database also had information about the CAN-Bus, GPS, radar, accelerometer, and rate sensor information.

The MIT *advanced vehicle technology* (AVT) Study [36] is a long-term driver behavior monitoring dataset. Subjects are asked to drive either their own vehicles or MIT-provided vehicles for a period of time between one month and over a year. The cars are equipped with sensors. This study aimed at capturing all aspects of the drivers' states as well as their interactions with technologies such as autopilot. They have recordings of 3 or 4 Logitech C920 RGB cameras, facing the driver's face, body, road, and occasionally the instrument cluster. The cameras record at $1,280 \times 720$ at 30 fps. In addition, they also have CAN-Bus data, audio from each of the cameras, IMU, and GPS data. The data are synchronized via a customized board.

While these efforts provide attractive platforms to build driver behavior models for in-vehicle systems, they do not focus on visual attention. Also, the data are proprietary and have important privacy constraints that prevent distributing the recordings to other research groups.

III. DATA RECORDING PROTOCOL

The goal of this study is to create a multimodal database recorded from multiple drivers targeting applications that model driver visual attention. For this purpose, we collect the *multimodal driver monitoring* (MDM) database, where our subjects are asked to follow a predefined protocol aiming to obtain labeled data to study gaze detection and head pose estimation. We perform our experiments in the UTDrive vehicle [39]–[41], which is a 2004 Toyota RAV4 SUV equipped with multiple on-board sensors and equipment (Fig. 1). We use various devices and sensors to set up baselines and record the data. The data is collected from a total of 59 subjects (32 females, 27 males). During the data collection, 31 of them wore prescription eye glasses, two wore sun glasses, and 26 did not wear any eye glasses. The drivers were all affiliated with the University of Texas at Dallas. We collected a total of 50 hours and 14 minutes of data, which captures the drivers performing a variety of activities as defined in our data collection protocol. We present in details the protocol followed during the data collection (Sec. III-A), the sensors used to record the driver, car, and road information (Sec. III-B), the synchronization of the sensors (Sec. III-C), the methods used to spatially calibrate the orientation of different sensors in the environment (Sec. III-D), and the annotations provided in the corpus (Sec. III-E).

Fig. 1. The UTDrive vehicle used for the data collection. The car is equipped with multiple sensors.

Fig. 2. Step 1 in the data collection protocol. The driver looks at a target board with a fiducial mark held by one of the researchers outside the car. This part in the protocol provides continuous gaze information.

A. Protocol

Our protocol is an improved version of the data collection protocol used in our previous work [14], where we made several additions and improvements, learning from our previous experience. Our goal is to obtain (1) continuous gaze data with reliable ground truth, and (2) naturalistic data when the driver is performing driving related tasks. We ask the drivers to perform multiple tasks while being at the driver seat in the UTDrive car. We provide a detailed description of each of the tasks included in this protocol.

1) Parked Car: Looking at a Target Outside the Vehicle: The first step in our protocol aims to collect continuous gaze information. We start the recordings while the vehicle is parked. Then, we ask the driver to sit in the driver seat while a researcher slowly moves a board outside the vehicle (Fig. 2). The board has a big fiducial marker with fixed pattern that can be easily tracked using basic image processing algorithms. We use an AprilTag [42] for this purpose such that we can easily track the 3D position of the marker (Fig. 2). We ask the driver to follow the target board with her/his gaze as the researcher moves the board in front of the car. We collect data in three to five sessions with short breaks in between. Each session is approximately one minute long.

2) Parked Car: Looking at Target Markers Inside the Car: As mentioned before, our goal is to have gaze direction information. The second step in the protocol is to ask the driver to look at predefined markers placed inside the vehicle. We placed 21 different target markers inside the car (Fig. 3). The first 13 markers are placed on the windshield. The rest

Fig. 3. Location of the target markers placed inside the car. Markers 1-13 are in the windshield. The rest of the markers are in the mirrors (14-16), side of the windows (17-18), speedometer panel (19), radio (20), and gear (21).

of the markers are placed on the mirrors (14-16), side of the windows (17-18), the speedometer panel (19), radio (20), and gear shifter (21). We ask the driver to look at each of these markers two to three times, calling the number in random order. The driver is asked to look at each marker as naturally as possible. This part of the protocol is conducted while the car is parked. The drivers get familiar with the location of the markers in a safe environment.

3) Driving Car: Looking at Target Markers Inside the Car: After the subject is familiar with the task of looking at target markers, we repeat this step while the driver is operating the vehicle. We ensure that this step is conducted when the subject is driving on a straight road, with reduced demand for driver maneuvering. We ask the subject to quickly glance at the target marker. We ask the subject to look at all the markers three times in a random order. The investigator points to the location of the target markers to reduce the cognitive load associated with searching for the right marker.

4) Driving Car: Looking at Targets Outside the Car: We ask the subjects to locate and identify targets such as landmarks, street signs, buildings, and other cars on the road to complement the gaze data obtained while looking at the markers inside the car. We ask questions to the subjects that prompt them to naturally look at the target locations. Examples of these questions are: *can you tell me the name of the store in the corner?* and *can you read the license plate of the car in front?* We also ask the subjects to search for multiple targets (e.g., can you find red cars? and can you read the license plates of as many cars as possible?). This part of the data collection creates implicit gaze responses that can be easily associated with the corresponding gaze directions. We ask each subject to identify between 20 to 30 landmarks on the road.

5) Driving the Car: Navigation: The next step of the data collection starts with the secondary tasks. We setup a multi-destination navigation route on a smartphone. The directions are chosen to have multiple maneuvers. We ask the drivers to wait for the navigation instructions to reroute in situations when they miss an instruction. The audio of the navigation is turned off and the drivers follow navigation purely by looking at the phone screen. Once the drivers reach a destination, they hit the continue button to obtain the navigation instructions for the next stop in the trip. We set four destinations in total.

6) Driving Car: Operating the Radio: The next step in the protocol is also a secondary driving task. We ask the drivers to operate the radio in the vehicle. We ask the drivers to turn on the car radio and set it to a given station. Since the UTDrive

(a) Navigation

(b) Radio

Fig. 4. Examples of subjects performing secondary tasks: (a) Following navigation on a smart phone, and (b) operating the in-car radio.

vehicle does not have radio controls on the steering wheel, the driver needs to operate the radio using the knobs in the central console. We repeat this step by calling out random stations for the drivers to find. 5-10 different radio channels were called out at random and the driver turned the radio to the closest existing channel.

7) Naturalistic Driving: During multiple instances in the data collection, we ask the subjects to drive with minimal intervention from the researcher. Only navigation instructions are provided to the driver and no additional task is carried out. These recordings aim to provide a baseline for natural driving conditions without secondary tasks.

B. Multimodal Sensors

To collect high quality data and relevant information from the driver, vehicle, and road environment, we use multiple devices during the data collection. Figure 5 shows the setup used to place different devices inside the car. This section discusses the sensors used to collect the data.

1) Fi-Cap: Providing reliable labels for head pose estimation is important for modeling the driver's visual attention. Our primary objective was to have a head pose label for each of the frames in our corpus. Alternative systems to track head pose in actual vehicle include motion capture systems [27], [30], and IMU measures [28]. The IMU systems are sensitive to drifts in the measurement and, hence, the recording of more than 10 minutes is not possible without significant drop in the reliability. Passive *Motion capture* (Mo-Cap) systems with retroreflective markers have gained popularity in recent studies [27], [30]. The reliability depends on the number of markers, the frame rate and resolution of the camera, and environmental factors such as background illumination and occlusion. A Mo-Cap system requires a sophisticated setup,
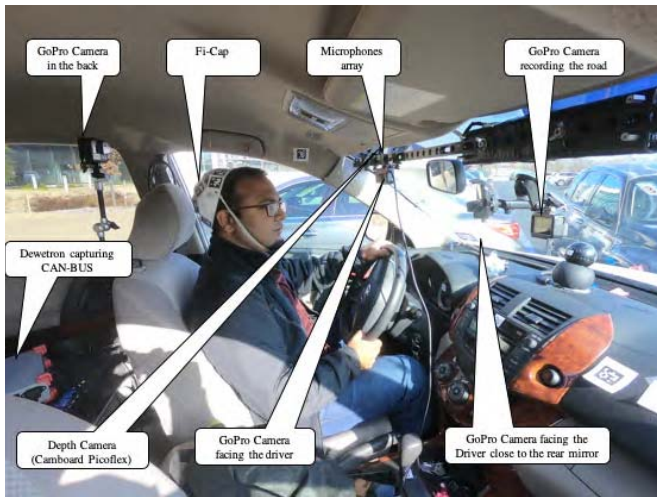
Fig. 5. Data collection setup showing the placement of the sensors in the UTDrive car. The setting includes four HD cameras, a depth camera, which simultaneously records grayscale images with IR illumination, microphone array, and a Dewetron that records the CAN-Bus data.

so it is more appropriate for data collections in car-simulators (e.g., AutoPOSE corpus [32]). There are very few examples of motion capture systems in naturalistic data collections in real road [27], [29], [30]. Vibrations in the car and movements of the markers can have a big impact on the recordings. Furthermore, most Mo-Cap systems do not work in outdoor environments. The placement of the markers can also occlude the driver's face, which can pose difficulty if we are interested in studying visual attention of the drivers. Additionally, Mo-Cap systems are usually expensive. A simpler alternative relying on a similar optical tracking is using fiducial markers that can be tracked using regular cameras and simple image processing techniques.

Our solution for estimating the frame-based labels for the head pose is based on fiducial markers placed on a solid helmet with a predefined structure with highly contrasting black and white patterns. The position and orientation of multiple fiducial points are easily tracked with simple algorithms, which are then used to estimate the head pose of the drivers. Our first prototype was a headband with 17 fiducial markers used in the forehead [14]. The main problem of this setting was the facial occlusion caused while wearing the headband. Our second prototype is the Fi-Cap helmet [11] (Fig. 6), which is used in this data collection. The Fi-Cap is a helmet that contains 23 different AprilTags [42]. The size of each of them is 3.2cm × 3.2cm. The high number of fiducial points, the size of each marker, and the 3D structure of the helmet lead to robust estimation of head movement, which can be established for a wide range of head poses. Another advantage of the Fi-Cap device is that the helmet is worn on the back of the driver's head, so the facial occlusion is minimal (see Fig. 7). We request the drivers to avoid touching and adjusting the Fi-Cap helmet during the recordings to avoid drifts. However, the drift that may still exist can be compensated over time during the calibration. The readers are referred to Jha and Busso [11] for more details about the Fi-Cap helmet. Notice

Fig. 6. The Fi-Cap helmet [11], which we use to obtain head pose labels for each frame. A HD camera is placed behind the driver to record the fiducial points in the helmet (Fig. 7(d)).

(a) Face camera

(b) Road camera

(c) Mirror camera

(d) Back camera

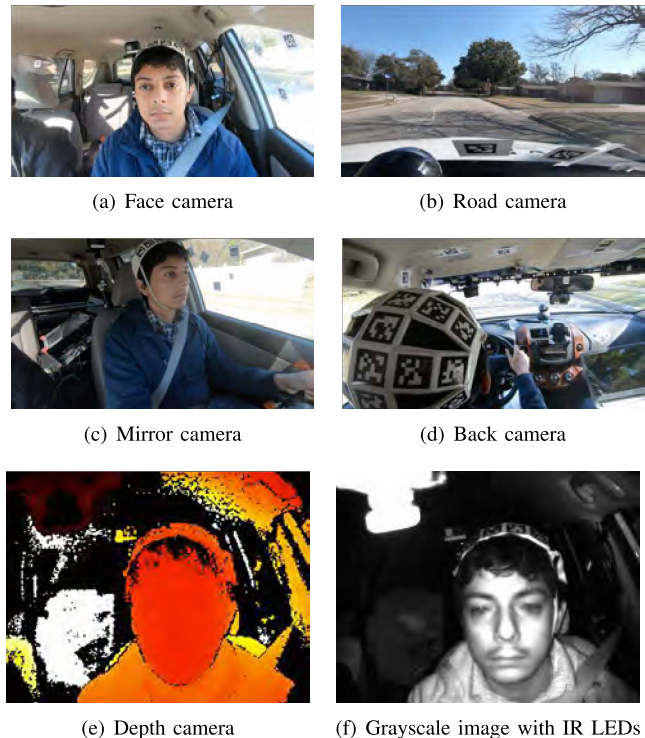(e) Depth camera

(f) Grayscale image with IR LEDs

Fig. 7. Examples of images collected with different sensors. All the recordings are synchronized with a clapboard.

that with the addition of the Fi-Cap helmet, we need an additional camera that can record the back of the driver's head. The additional camera is the only extra sensor in our setting, which makes this solution a very convenient approach for estimating frame-based labels for head poses. These advantages were the main reasons for using the Fi-Cap over other alternative methods.

2) RGB Cameras: We use four GoPro Hero 6 Black cameras in our data collection. All the cameras are set to record in full HD $(1,920 \times 1,080)$ in linear mode. The purpose and placement of each of the four cameras are listed as follows:

- Face Camera - This camera is placed in front of the driver to record the frontal view of the driver's face (Fig. 7(a)).
- Road Camera - This camera is placed in the center of the dashboard facing the road. This camera captures road related information (Fig. 7(b)).

- Mirror Camera - This camera is placed under the rearview mirror to obtain a profile view of the driver's face. Several cameras installed on commercial cars are placed on the rearview mirror, so this camera provides realistic views that can be expect in deploying in-vehicle technology (Fig. 7(c)).
- Back Camera - This camera is placed behind the driver to record the Fi-Cap helmet (Fig. 7(d)). The camera is attached to the driver seat.

3) Depth Camera: A sensor that can be particularly useful in a vehicle is a depth camera, which has gained popularity in various computer vision applications. For our database, we place a camboard picoflex camera close to the face camera. This camera records the point cloud data using time-of-flight technology, providing robust estimation of the depth map in varying illumination conditions (Fig. 7(e)). This camera provides reliable information even during night-time. The camera also records grayscale images which are illuminated using IR LEDs (Fig. 7(f)) and, hence, is immune to ambient lighting variations. The camboard picoflex camera records at 45fps, with a resolution of 224×171.

4) CAN-Bus: The UTDrive vehicle records the CAN-Bus information during the recordings. From the CAN-Bus, we obtain the information about accelerator, brakes, steering, and speed of the vehicle. The UTDrive vehicle also has a gas and brake pressure sensors. The information provided in the CAN-Bus is very useful, for example, to analyze the vehicle state, and obtain driver maneuver information.

5) Microphone Array: The UTDrive vehicle is also equipped with a microphone array with five microphones. While our protocol does not include any task that elicits speech, the audio information is useful for understanding potential auditory distractions.

6) Dewetron: The CAN-Bus data and the microphone array are connected to a Dewetron system, model DA-120, which stores and synchronizes the modalities.

C. Synchronization of Sensors

The Dewetron system can record up to two cameras. Since we are using multiple sensors, each with its own independent clock, we use a clapboard to synchronize the recordings. This is an easy, simple, and effective approach to synchronize audiovisual recordings, which is used in movie productions and other popular audiovisual databases [43]. Figure 8 shows the process. The clapboard provides a reference time to synchronize these devices. We perform two claps to synchronize all the cameras. The first clap is performed inside the car such that it is visible to the face camera, mirror camera, back camera, and the depth camera. The second clap is performed outside the car such that it is visible to the back camera and the road camera. The claps can be precisely detected in the recording using the video and the audio. Using the video, the clap is manually analyzed by obtaining the first frame when the two bars of the clapping board touch each other. In the audio, we manually mark the peak observed when the clap happens. Using these timings, we calculate the time offset (t_{offset}) by obtaining the difference between two
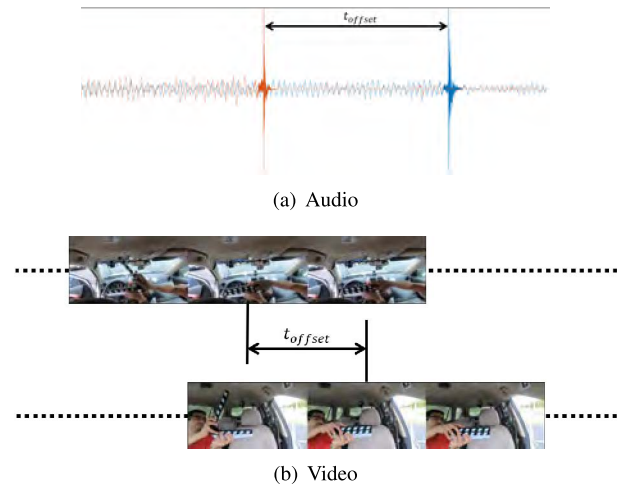

(a) Audio

(b) Video

Fig. 8.   Synchronization of audiovisual sensors using a clapping board. The peak in the audio and the videos' frames when the two bars of the clapping board touch each other are manually annotated.

streams of data. We did not observe any drift in the recordings provided by the GoPro cameras. The sampling rate for the picoflexx camera is not constant. However, the camboard picoflex camera records the time information of each frame, which is used to synchronize the point cloud with RGB data.

D. Cameras Calibration

We perform camera calibration in two steps. In the first step, we establish a reference position and orientation for each camera. This step was performed once before we started our data collection. For this step, we place all the cameras at their desired locations. We place multiple AprilTags [42] of size $3.2\text{cm} \times 3.2\text{cm}$ around the car such that all the cameras can see some of the tags. Figure 9 shows example of these markers. Then, we use an additional camera to capture static images inside the vehicle from multiple perspectives. We rely on the Kabsch algorithm [44] to calculate the transformation between the coordinate systems associated with different cameras. We establish the location of the tags from multiple cameras, which are used to obtain the relative position between the cameras. The procedure is similar to the one proposed in Jha and Busso [45]

We also place AprilTags at each of the target marker (Fig. 3) as shown in Figure 9(h) to estimate the 3D positions of the markers. Using these tags, we obtain the 3D coordinates of the markers with respect to a common coordinate system (we arbitrarily choose the back camera as our reference).

The second step in the calibration process is performed to compensate for small camera placement variations between sessions. The driver seat is adjusted affecting the location of the back camera. In addition, all the cameras are removed from the vehicle after the session. After recharging the equipment, they are reinstalled back in the vehicle to their original positions. This process can result in slight changes in the positions and orientations of the cameras. For this purpose, we fix multiple AprilTag markers in different locations in the car, including the ceiling, windshield, dashboard, and side-windows. These markers are left in the car for all the sessions.

(a) Face camera

(b) Back camera

(c) Road camera

(d) Mirror camera

(e) Moving camera view 1

(f) Moving camera view 2

(g) Moving camera view 3

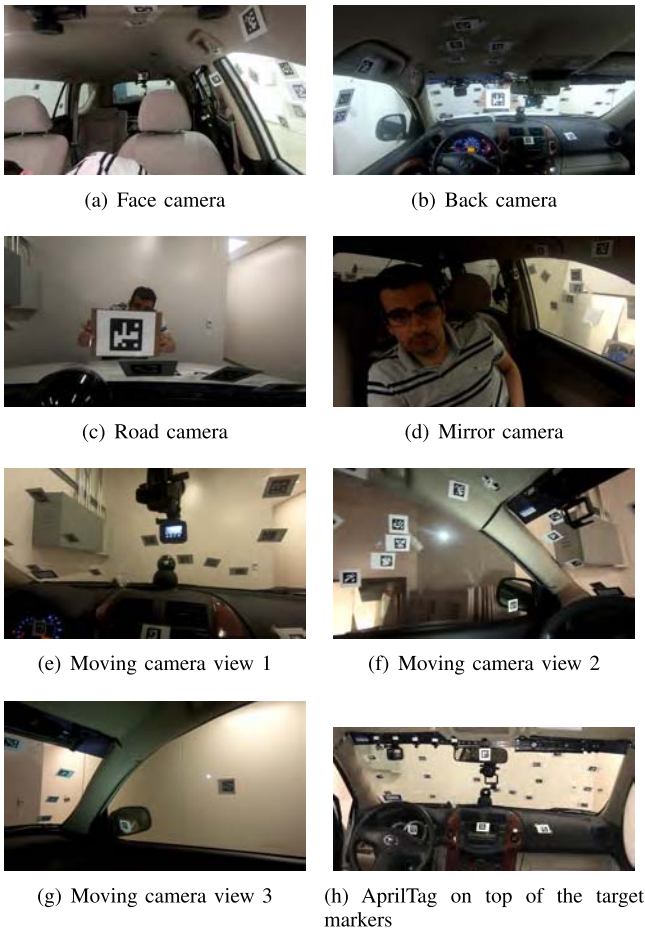(h) AprilTag on top of the target markers

Fig. 9.   AprilTags placed at different locations inside and outside the car to calibrate the cameras.

We ensure that each camera has at least two AprilTag markers within its field of view. By estimating the corners and the center of each AprilTag, we estimate at least 10 reference points for every camera view. We calculate the variations in the setup of the cameras for each session using the changes in the positions of these tags from the original reference setup using the Kabsch algorithm.

E. Annotations

We use the ELAN tool [17] to annotate events in the videos (Fig. 10). The videos from the four GoPro cameras are synchronized by manually finding the precise timings of the claps in each video. The annotation process has multiple channels with relevant information that are currently provided in the database:

- *Activity:* This channel splits the data collection into the stages defined in our protocol (Section III).
- *Continuous_Gaze:* This channel identifies the segments when the subjects are following the target board outside the car while the car is parked (Section III-A.1).
- *Gaze_Markers:* This channel locates the times when subjects are looking at different markers inside the vehicle (Sections III-A.2 and III-A.3). The numbers of the markers are also provided in the annotations.

- *Gaze_Road:* This channel locates the times when subjects are looking at different landmarks on the road (Sec. III-A.4). It also provides information about the actual target objects glanced at by the driver.

Since ELAN provides a convenient way of adding multiple tier of annotations, more information can be added in the future based on specific target applications (e.g., mirror-checking actions [46], driver maneuvers [47]).

## IV. HEAD POSE ESTIMATION WITH FI-CAP

A. Head Pose Estimation

The Fi-Cap contains multiple AprilTags that are tracked using the back camera. The locations of the fiducial points determine the frame-based labels for the head poses of the drivers. The first step to track the head pose is to obtain the position and the orientation of the cap with respect to a standard reference given a visible subset of tags on the cap. We follow the steps proposed in Jha and Busso [11]. We obtain a mesh of the Fi-Cap by finding the corners of each of the AprilTags in the structure by conducting continuous recording of the cap from different perspectives. Using this mesh as a reference, we obtain the transformation of all the visible corners in a given frame using the Kabsch algorithm [44]. We use a method inspired by the *iterative closest point* (ICP) algorithm [48] to reduce the effect of outliers in estimating the pose. We reconstruct the mesh from the estimated transformation and remove the points that have high reconstruction error. Then, we recalculate the transformation matrix based on the reduced set of points. We repeat this step until the error is below a threshold. We consider this final transformation as the pose of the Fi-Cap helmet for each frame, referred to as R_{FiCap}.

The next step is to find the head orientation from the Fi-Cap pose. We use the multiple *local reference frame* (LRF) calibration approach proposed by Hu *et al.* [49]. This transformation is subject-specific because each subject might wear the Fi-Cap helmet differently. The transformation ensures that the ground truth for the head pose is defined uniformly across subjects. We use the ICP algorithm [48] to achieve this goal. First, we manually select one frontal frame from one subject to be the *global reference* (GR) frame (Fig. 11(b)). We calculate its head pose estimate using OpenFace2.0 [50] by processing its corresponding gray-scale image. For every subject, we choose a number of frames with the closest rotation angles as the GR frame (Fig. 11(a)). We refer to these selected frames as *local reference* (LR) frames, obtaining the rotation matrix R_{LRF} for each of them. For each LR frame, we crop the face region in the point cloud data for the LR and GR frames. Then, we run the ICP algorithm between the two point cloud data to find a transformation between them, which compensates for the differences in head pose rotation between the LR and GR frames (matrix $R_{LocalGlobal}$ in Fig. 11(b)). Therefore, the LR and GR frames do not need to have the exact head rotation. Then, the calibration matrix is defined by $R_c = R_{LocalGlobal}R_{LRF}$. For frame t, with rotation R_{FiCap}, we estimate the final head rotation with $R_t = R_c^{-1}R_{FiCap}$. During long trips, the Fi-Cap helmet may move causing errors

Fig. 10.   The corpus uses ELAN as the interface to annotate the recordings. The annotation includes the segments associated with the seven steps in the protocol. It also provides the temporal information for instances when the subject looked at the target markers or objects inside and outside the vehicle.
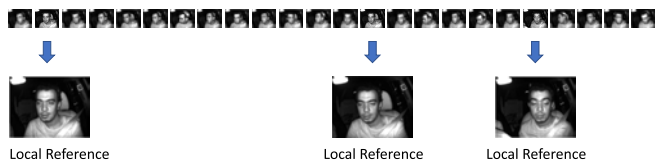
(a) Selection of multiple *local reference frames* to avoid drifts during a session


$$R_c = R_{LocalGlobal}R_{LRF}$$
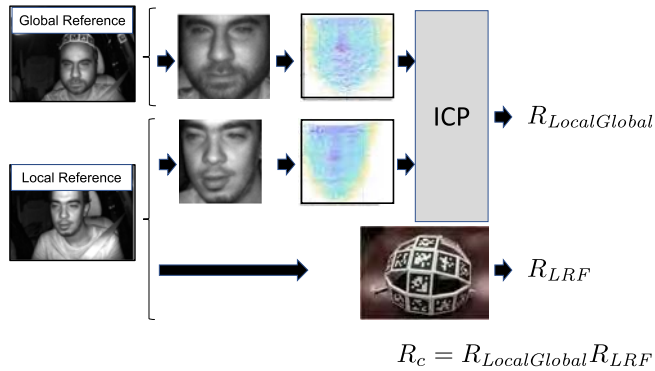
(b) Transformation from local to global reference frames

Fig. 11.   Approach proposed in Hu *et al.* [49] to estimate the head pose labels using multiple local reference frames.

in the calibration. We maintain multiple transformations for each subject to ensure maximum accuracy against any drift of the Fi-Cap helmet. We use the closest LRF in time from frame t for this head pose compensation. Section V-E presents an analysis of the benefits of using multiple LRF.

B. Reliability of the Fi-Cap Helmet

The Fi-Cap helmet provides the position and orientation of the driver's head. In Jha and Busso [11], we estimate the reliability of the Fi-Cap using a virtual animation, where an exact copy of the helmet was rendered and placed in a virtual agent. By controlling the rotation, resolution of the image and illumination, we demonstrated that the median angular error was less 1.62°, and that the 95 percentile error was less than 2.88°. We also validated the approach in a laboratory setting,
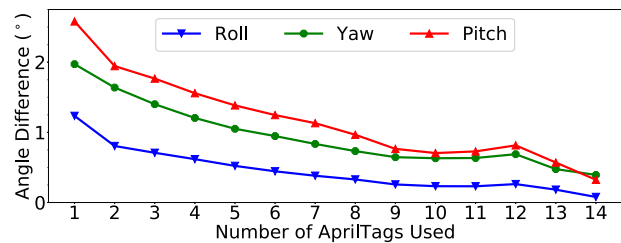
(a) Mean of the rotation angle of matrix R
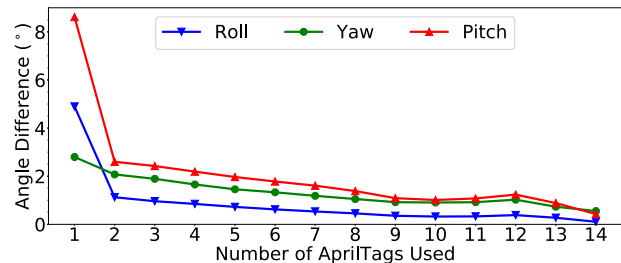
(b) Standard deviation of the rotation angle of matrix R

Fig. 12.   Mean and standard deviation of the rotation angle of matrix R capturing the variability of head rotation angles estimated as a function of the number of detected fiducial points.

using a laser mounted on glasses worn by a subject. We projected the laser into a white screen, measuring the angular distance between the head pose direction, as determined by the laser marker, and our prediction. The results showed median angular errors that were less than 2.31°, and 95 percentile angular errors that were less than 6.93°. The estimation of head position is more reliable when more fiducial markers are detected. This section evaluates the reliability of the ground truth for head pose provided by the Fi-Cap helmet as a function of the number of detected fiducial markers.

Our approach to evaluate the reliability of the Fi-Cap helmet consists of comparing the rotation matrix estimated with all the fiducial markers, and with a subset of the markers. Let n denotes the number of tags detected in one particular frame. We first calculate the global head pose using all the n tags. We refer to this rotation matrix as R_{all}. For every combination, we evaluate the rotation matrix with different

(a) Driver's head: yaw-pitch angles   (b) Driver's head: yaw-roll angles   (c) Driver's head: roll-pitch angles
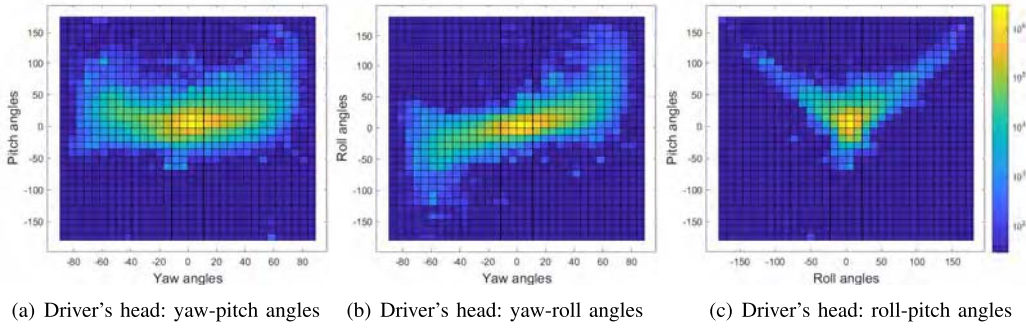
Fig. 13.   Distribution of the driver's head orientation in our database estimated with the Fi-Cap helmet. The results are projected into the yaw-pitch, yaw-roll and roll-pitch spaces.

subsets of markers using only the selected tags:

$$\binom{n}{1}, \binom{n}{2}, \binom{n}{3} \cdots \binom{n}{n-2} \qquad (1)$$

We refer to one of these matrices as R_{subset}. Then, we calculate the rotation difference between R_{all} and R_{subset}, namely $R = R_{subset} R_{all}^{-1}$. We aggregate all the matrices calculated in the previous step, obtaining their mean absolute value and standard deviation. Notice that we use the combinations of up to $n-2$ to avoid having few combinations to estimate these statistics (Eq. 1).

Figure 12 shows the mean and standard deviation of the rotation angle of matrix R as a function of the number of fiducial points used in the estimation. Lower angles indicate small deviations from the angles estimated with all the markers. This analysis shows the uncertainty in the ground truth labels when few fiducial points are used in the estimation. We observe that we can obtain a mean difference of less than $1°$ with sufficient number of visible tags (2 for roll, 5 for yaw and 8 for pitch). We observe the largest difference for pitch angle, which is the most difficult rotation to estimate. The standard deviation represents the variability in estimating the Fi-Cap pose when selecting different subsets of visible tags. The standard deviation drops significantly when more than two tags are detected. As we add more fiducial points, the estimation of the Fi-Cap position becomes more consistent. In fact, Figure 12 shows that with more tags, the rotation estimates are closer to R_{all}. As a reference, the average number of detected fiducial points per frame is 6.16. We have five or more fiducial points in 57.93% of the frames with detected tags. These results show that our ground truth labels for head pose are very close to the hypothetical scenario where every single tag is visible and detected. This analysis demonstrates the reliability of our ground truth labels estimated with the Fi-Cap helmet.

V. ANALYSIS

The MDM database provides a diverse resource to design algorithms to model the visual attention of the driver. This section analyzes the database to discuss its potential use.

A. Head Pose Distribution

We start our analysis by showing the distribution of the driver's head orientation during the entire recording.

Figure 13 shows the distribution, projected on the yaw-pitch, yaw-roll, and roll-pitch spaces. We notice that our dataset covers a large range of head poses along all three rotation axes due to the large number of subjects included, and the variety of primary and secondary driving activities considered during the data acquisition. Figure 13(a) shows a wide symmetric yaw angle range around the origin spanning between $-80°$ to $80°$, which reflects practical driving scenarios where the driver looks forward most of the time, but frequently checks the mirrors, dashboard, windshield, and side windows. In contrast, pitch angles have an asymmetric range spanning from $-50°$ to $100°$, reflecting high degrees of freedom at high pitch angles as the driver looks at the dashboard or gear shifter compared to low pitch angles, which almost vanishes at the edge between the windshield and the car ceiling.

B. Driver's Gaze Distribution

We also analyze the distribution of the annotated gaze directions. We represent the driver's gaze using the elevation (vertical) and azimuth (horizontal) gaze angles and report the distribution of these angles. First, we pick the frames where the driver is looking at target markers inside the car while driving (Fig. 3). The locations of these markers are estimated during the calibration phase for our reference. Second, we take the frames where the driver is looking at the fiducial markers outside the car. The locations of these markers are estimated with respect to the road camera. We transform the points into the back camera coordinate using the calibration data (Sec. III-D). Third, we approximate the location of the head with the location of the Fi-Cap helmet. With the head location, we define the driver's gaze vector ($\mathbf{g}$) as the line joining the head's location with the target marker and calculate the elevation angle (ϕ_{gaze}) and azimuth angle (θ_{gaze}) of this line in the 3D space. Hence, we can obtain absolute gaze directions that are agnostic of the camera locations and calibrations. Equations 2-5 show the expressions of these angles, where $\mathbf{t_{gaze}}$ is the 3D location of the target gaze, $\mathbf{h_{pos}}$ is the head location, and $\hat{\mathbf{g}} = (\hat{g}_x, \hat{g}_y, \hat{g}_z)$ is a unit vector in the direction of the gaze.

$$\mathbf{g} = \mathbf{t_{gaze}} - \mathbf{h_{pos}} \qquad (2)$$

$$\hat{\mathbf{g}} = \frac{\mathbf{g}}{\|\mathbf{g}\|} \qquad (3)$$

(a) Looking at in-vehicle markers   (b) Looking at the moving board
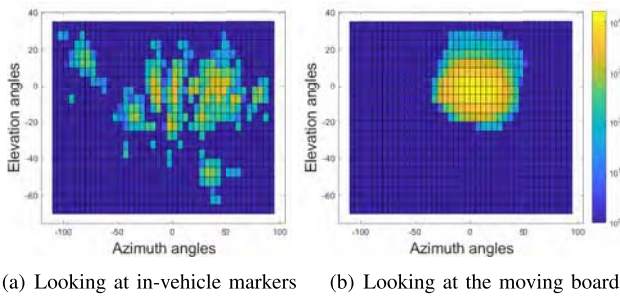
Fig. 14. Distribution of the driver's gaze angles in our dataset: (a) distribution of the drivers' gaze as she/he looks at target markers inside the car (step 3 in our protocol), and (b) distribution of the driver's gaze angles as the driver looks at the moving board outside the car (step 1 in our protocol).

$$\theta_{gaze} = arctan\left(\frac{\hat{g}_x}{-\hat{g}_z}\right) \qquad (4)$$

$$\phi_{gaze} = arctan\left(\frac{\hat{g}_y}{\sqrt{\hat{g}_x^2 + \hat{g}_z^2}}\right) \qquad (5)$$

Figure 14(a) reports the distribution of the gaze angles when the driver looks at markers inside the car while driving (step 3 in our protocol - Sec. III-A). Figure 14(b) reports the distribution of the gaze angles as the driver looks at the moving board during the parking phase (step 1 in our protocol - Sec. III-A). As expected, the gaze angle distribution for the in-vehicle markers has a higher azimuth and elevation angles ranges since these markers span more locations in the car in contrast to the moving board gaze distribution, which only spans the windshield area. However, the moving board gaze angles have a more balanced and smooth distribution due to their continuous nature. In addition, they can help in training time-based deep learning models (e.g., recurrent neural networks) as they provide continuous gaze annotations for successive frames.

The MDM database can be used for other gaze representations. For example, the driver's gaze can be represented in terms of normalized pixel location in the road camera scene. The 3D location of the target marker can be mapped into any camera as the reference frame, since all cameras are calibrated. This alternative representation can be useful if the goal is to map the driver's gaze directly to targets on the road.

C. Head Pose Estimation With Depth Data

This section demonstrates one of the potential uses of the MDM database for *head pose estimation* (HPE). In Hu *et al.* [51], we showed that we can effectively estimate head pose from a depth camera by directly processing point cloud data. In this work, we achieved better performance than state-of-the-art HPEs based on RGB images, especially in frames with large rotations. Figure 15 shows the model structure inspired by the work of Qi *et al.* [52], [53], where we formulate this task as a regression problem. We have three basic building blocks for this model: sampling, grouping, and PointNet. In sampling, we use the iterative farthest point sampling algorithm to get "anchor points." This step reduces the redundancy of the point cloud while maintaining its

Fig. 15. Diagram of the algorithm proposed by Hu *et al.* [51] to estimate head pose of the driver using point cloud data. The framework directly uses the point cloud data without projecting the 3D points into 2D spaces.

TABLE IV

MSE OF THE POINT-CLOUD BASED ALGORITHM AND OPENFACE 2.0. WE REPORT RESULTS OF THE PROPOSED MODEL ON THE ENTIRE TEST SET AND ALSO ON THE SET WHERE OPENFACE 2.0 PROVIDES ESTIMATIONS

Method	Errors			Percentage Frame Missed
	Roll(°)	Yaw (°)	Pitch (°)	
Hu et al.	5.77	5.84	6.33	0%
	OpenFace set			
Hu et al.	**5.68**	5.63	**6.27**	0%
OpenFace 2.0	7.63	**5.06**	7.20	3.08%

structure as much as possible. In the grouping building block, we group points within a radius R of the anchor points, capturing the relationship between each anchor point and its neighbors. In PointNet, we adopt multi-layer perceptrons to learn from the data features that are more discriminative for our task. We repeat the sampling-grouping-PointNet layer set five times, aggregating features from different resolutions. Finally, we acquire a high-level feature vector that represents the input point cloud. We use a fully-connected layer to derive the head rotation from this feature vector. We retrain the models using data from 39 subjects for training, 10 for validation, and 10 for testing.

We estimate the *mean square error* (MSE) of this model for all the frames in the test set, reporting the average results in Table IV. The average angular error for our estimation was between 5.77° and 6.33° on the entire test set. We compare our point cloud HPE model with a state-of-the-art HPE algorithm based on RGB images. We use OpenFace 2.0 [54], estimating the head pose with the $1,920 \times 1,080$ face camera (GoPro camera – Fig. 7(a)). To ensure that the definition of head pose in our dataset is consistent with OpenFace 2.0, for each subject, we find an average transformation between the head pose ground truth and OpenFace 2.0. We observe that OpenFace 2.0 fails at giving predictions in 3.08% of the frames in the test set. The performances on the frames with predictions are 7.63° for roll, 5.06° for yaw, and 7.20° for pitch angles. In contrast, the proposed approach produces a head pose estimate for every frame. For fair comparison, we evaluate our approach only on the frames with predictions from the RGB-based HPE algorithm, which we refer to as the *OpenFace set*. Table IV shows clear improvements in using our point cloud algorithm, with absolute differences of 1.95° for roll and 0.93° for pitch angles. In the yaw direction, openface works marginally better with the absolute error difference of 0.57°.

D. Driver Gaze Estimation From Eye Appearance

This section demonstrates another potential research of the MDM dataset in predicting the gaze direction of the driver
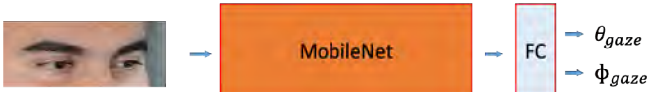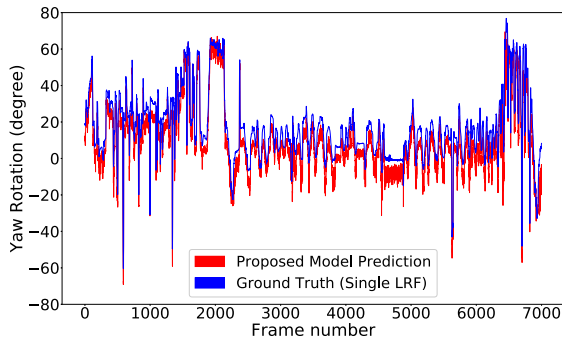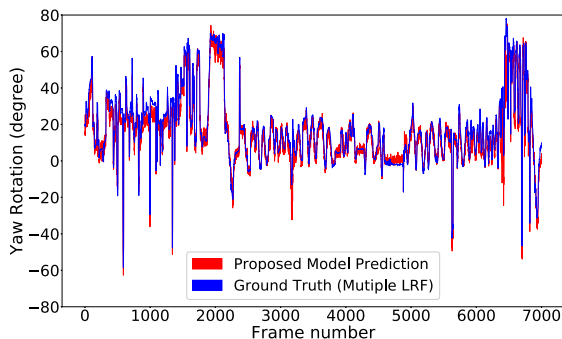
Fig. 16. Diagram of the model using the MobileNet network to predict the driver's gaze from the eye appearance.

TABLE V
MSE OF GAZE ESTIMATION ON THE TEST SET, USING THE CONTINUOUS GAZE SET

Method	Errors	
	θ_{gaze} (°)	ϕ_{gaze} (°)
Mobilenet	13.90	5.10

(a) Single Local Reference Frame

(b) Multiple Local Reference Frame

Fig. 17. Illustration of the effect of drifts in the Fi-Cap placement on predicted versus ground truth head orientations (yaw angles). The figure shows the results using a single LRF or multiple LRF for head pose calibration.

from the facial image. We evaluate the model shown on Figure 16, which is a simple approach based on the MobileNet architecture [55]. This model uses the FAN algorithm [56] to extract a patch from the face camera image that includes both eyes. This image is sent through the MobileNet architecture, which is trained from scratch. The output of the MobileNet block is fed to a fully connected layer with two outputs for the horizontal and vertical gaze angles (Fig. 16). We create a subject independent partition with data from 39 subjects for the train set, data from 10 subjects for the test set, and data from 10 subjects for the development set. We train and test the model with the continuous gaze data (e.g., drivers looking at the board held by the researcher outside the car when the car is parked). Table V shows the results. We observe that the test

error is 13.90° in the horizontal direction (θ_{gaze}) and 5.10° in the vertical direction (ϕ_{gaze}). We expect that these results can be improved with more sophisticated models that compensate for variations in appearance across subjects.

E. Drift Compensation Using Multiple Reference Frames

We analyze the benefits of using the multiple LRF approach, presented in Section IV-A, to compensate for potential Fi-Cap drifts during the data collection. We use the point cloud-based HPE model discussed in Section V-C for the analysis. Figure 17 shows the prediction of yaw angles when using a single LRF and when using multiple LRF. We observe that while the pattern of the estimated head pose follows the ground truth when using a single LRF, there are random drifts in different sections of the plots (Fig. 17(a)). In contrast, using multiple LRF alleviates this problem (Fig. 17(b)).

VI. CONCLUSION

This paper presented the MDM database, which provides naturalistic recordings that are ideal for studying visual attention in realistic driving conditions. The database includes multimodal sensors including four HD cameras, depth sensors, grayscale camera with IR illumination, and a microphone array. The database also includes the CAN-Bus information to obtain vehicle related information. The database provides ground truth for both head pose and gaze information. The Fi-Cap helmet is used to provide continuous annotation of the driver's head pose. The placement of the helmet in the back of the driver's head avoids occlusions for frontal cameras recording the face of the subject. This setup provides head pose information for each of the millions of frames included in the corpus. We obtained the driver's reference gaze in a continuous setting, where we moved a board with a fiducial point outside the car while the car is parked. We also collected the event based gaze directions by asking drivers to look at particular locations, both inside and outside the vehicle. In addition, we collected data when the driver performs natural secondary activities such as changing the radio station and following instructions. Collectively, the information included in the corpus can be useful in training efficient algorithms to study driver visual attention in naturalistic driving conditions.

This study also provided an analysis of the MDM corpus. We introduced a novel experimental framework to quantify the robustness of our head pose labels by studying the estimated head pose as a function of the number of detected fiducial points in the Fi-Cap helmet. By analyzing the distribution of gaze and head pose information, we showed that our dataset covers a wide range of angles in all rotation axes, which reflects the variety of primary and secondary activities included in the data collection protocol.

A. Potential Research Relevant to the MDM Corpus

The collected data provides a great resource to analyze the driver behavior. The corpus's size, the driver diversity, the naturalistic nature of the recording, the presence of multiple sensors, and the detailed annotations provided in the

corpus are ideal to train sophisticated deep learning solutions to monitor driver visual attention for in-vehicle systems. The MDM corpus is also an ideal resource to explore various research questions in the field of driver monitoring systems. The dataset contains continuous benchmarking of head pose for the entire duration of the recording, which can be used to train machine learning models for head pose estimation, using RGB cameras, depth information, or multimodal frameworks combining both modalities. Additionally, the ground truth gaze is collected in varying conditions. These data points can be used to train models for driver visual attention that are robust to road conditions observed in naturalistic recordings. Another feature of this corpus is the multiple camera setup, which can be ideal for designing multi-view models. This setting is also ideal to study the effect on the performance of visual attention models when the perspective of the camera is not perfect. For example, we can analyze the compromise in accuracy by using the mirror camera (profile view) instead of the face camera (frontal view). The road cameras can be used to extract information about the environment, which can be associated with the driver activity and behavior. Likewise, we can estimate the vehicle state by using the CAN-Bus information, which can be correlated with the driver's head pose as well as his/her visual attention. The corpus also offers an ideal platform to study driver behavior during secondary tasks such as following the navigation system or operating the radio. Techniques such as driving anomaly detection [25], [26] can also be studied by analyzing the data collected from various sensors during different activities.

B. Limitations of the Corpus

Our database has some limitations. The *Institutional Review Board* (IRB) approved by our university required the data to be collected only during low traffic conditions and during daylight hours. Therefore, we do not have data collected during the night or in rush hour traffic. Challenging lighting conditions can be simulated using data augmentation strategies, which can be leveraged to train robust models against low illumination. The secondary tasks selected in the data collection were also constrained by the IRB approved by our institution. For example, we were not able to include tasks that require the driver to use a cellphone. Furthermore, the corpus is not ideal to study important research problems such as fatigue detection. Finally, we do not provide continuous gaze annotations for all frames, similar to head pose. Some studies have attempted to collect continuous gaze labels using commercial glasses with cameras. Unfortunately, these systems occlude the face so they were not appropriate for our data collection. Nevertheless, we expect that the gaze data provided in our corpus include representative images collected in varied settings to train efficient algorithms for naturalistic gaze estimation.

C. Access to Corpus

The multimodal driver monitoring dataset is licensed free of cost to academic institutions under a *Federal Demonstration Partnership* (FDP) Data Transfer and Use Agreement. We have also established a licensing program for commercial entities interested in our corpus.The instructions for obtaining the corpus are available at https://ecs.utdallas.edu/research/researchlabs/msp-lab/MDM.html.

REFERENCES

[1] (2019). *The National Highway Traffic Safety Administration (NHTSA)*. Accessed: Aug. 5, 2020. [Online]. Available: https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety

[2] A. M. Bronstein, M. M. Bronstein, and R. Kimmel, "Expression-invariant 3D face recognition," in *Audio-Video-Based Biometric Person Authentication* (Lecture Notes in Computer Science), vol. 2688, J. Kittler and M. Nixon, Eds. Guildford, U.K. Berlin, Germany: Springer, Jun. 2003, pp. 62–70.

[3] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, Swansea, U.K., Sep. 2015, pp. 1–12.

[4] L. De Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition using multi-modal information," in *Proc. Int. Conf. Inf., Commun. Signal Process. (ICICS)*, Singapore, vol. 1, 1997, pp. 397–401.

[5] S. Jha and C. Busso, "Estimation of gaze region using two dimensional probabilistic maps constructed using convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 3792–3796.

[6] N. Li and C. Busso, "Calibration free, user-independent gaze estimation with tensor analysis," *Image Vis. Comput.*, vol. 74, pp. 10–20, Jun. 2018.

[7] S. Jha and C. Busso, "Challenges in head pose estimation of drivers in naturalistic recordings using existing tools," in *Proc. IEEE Int. Conf. Intell. Transp. (ITSC)*, Yokohama, Japan, Oct. 2017, pp. 1624–1629.

[8] S. Jha and C. Busso, "Probabilistic estimation of the driver's gaze from head orientation and position," in *Proc. IEEE 20th Int. Conf. Intell. Transp. Syst. (ITSC)*, Yokohama, Japan, Oct. 2017, pp. 1630–1635.

[9] S. Jha and C. Busso, "Probabilistic estimation of the gaze region of the driver using dense classification," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Maui, HI, USA, Nov. 2018, pp. 697–702.

[10] S. Vora, A. Rangesh, and M. M. Trivedi, "Driver gaze zone estimation using convolutional neural networks: A general framework and ablative analysis," *IEEE Trans. Intell. Veh.*, vol. 3, no. 3, pp. 254–265, Sep. 2018.

[11] S. Jha and C. Busso, "FI-CAP: Robust framework to benchmark head pose estimation in challenging environments," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, San Diego, CA, USA, Jul. 2018, pp. 1–6.

[12] A. Tawari, K. H. Chen, and M. M. Trivedi, "Where is the driver looking: Analysis of head, eye and iris for robust gaze zone estimation," in *Proc. IEEE Conf. Intell. Transp. Syst.*, Qingdao, China, Oct. 2014, pp. 988–994.

[13] B. Vasli, S. Martin, and M. M. Trivedi, "On driver gaze estimation: Explorations and fusion of geometric and data driven approaches," in *Proc. 19th IEEE Int. Conf. Intell. Transp. Syst.*, Rio de Janeiro, Brazil, Nov. 2016, pp. 655–660.

[14] S. Jha and C. Busso, "Analyzing the relationship between head pose and gaze to model driver visual attention," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC)*, Rio de Janeiro, Brazil, Nov. 2016, pp. 2157–2162.

[15] G. Li *et al.*, "Drivers' visual scanning behavior at signalized and unsignalized intersections: A naturalistic driving study in China," *J. Saf. Res.*, vol. 71, pp. 219–229, Dec. 2019.

[16] G. Li *et al.*, "Influence of traffic congestion on driver behavior in post-congestion driving," *Accident Anal. Prevention*, vol. 141, Jun. 2020, Art. no. 105508.

[17] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "ELAN: A professional framework for multimodality research," in *Proc. Int. Conf. Lang. Resour. Eval. (LREC)*, Genoa, Italy, May 2006, pp. 1556–1559.

[18] F. Yu *et al.*, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 2633–2642.

[19] Y. Xia, D. Zhang, J. Kim, K. Nakayama, K. Zipser, and D. Whitney, "Predicting driver attention in critical situations," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, in Lecture Notes in Computer Science, vol. 11365, C. Jawahar, H. Li, G. Mori, and K. Schindler, Eds., Perth, WA, Australia. Berlin, Germany: Springer, Dec. 2018, pp. 658–674.

[20] G. Neuhold, T. Ollmann, S. R. Bulò, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proc. ICCV*, Venice, Italy, Oct. 2017, pp. 5000–5009.

[21] P. Sun *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 2443–2451.

[22] H. Caesar *et al.*, "Nuscenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Aug. 2020, pp. 11618–11628.

[23] A. Palazzi, D. Abati, S. Calderara, F. Solera, and R. Cucchiara, "Predicting the driver's focus of attention: The DR(eye)VE project," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1720–1733, Jul. 2019.

[24] W. Li *et al.*, "A spontaneous driver emotion facial expression (DEFE) dataset for intelligent vehicles: Emotions triggered by video-audio clips in driving scenarios," *IEEE Trans. Affect. Comput.*, early access, Mar. 3, 2021, doi: 10.1109/TAFFC.2021.3063387.

[25] Y. Qiu, T. Misu, and C. Busso, "Driving anomaly detection with conditional generative adversarial network using physiological and CAN-bus data," in *Proc. Int. Conf. Multimodal Interact.*, Suzhou, China, Oct. 2019, pp. 164–173.

[26] Y. Qiu, T. Misu, and C. Busso, "Use of triplet-loss function to improve driving anomaly detection using conditional generative adversarial network," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, Rhodes, Greece, Sep. 2020, pp. 1–7.

[27] S. Martin, A. Tawari, E. Murphy-Chutorian, S. Y. Cheng, and M. Trivedi, "On the design and evaluation of robust head pose for visual user interfaces: Algorithms, databases, and comparisons," in *Proc. 4th Int. Conf. Automot. User Interfaces Interact. Veh. Appl. (AutomotiveUI)*, Portsmouth, NH, USA, Oct. 2012, pp. 149–154.

[28] A. Tawari, S. Martin, and M. M. Trivedi, "Continuous head movement estimator for driver assistance: Issues, algorithms, and on-road evaluations," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 2, pp. 818–830, Apr. 2014.

[29] A. Schwarz, M. Haurilet, M. Martinez, and R. Stiefelhagen, "DriveAHead—A large-scale driver head pose dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Honolulu, HI, USA, Jul. 2017, pp. 1165–1174.

[30] M. Roth and D. M. Gavrila, "DD-pose—A large-scale driver head pose benchmark," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Paris, France, Jun. 2019, pp. 1176–1182.

[31] G. Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara, "POSEidon: Face-from-depth for driver pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 5494–5503.

[32] M. Selim, A. Firintepe, A. Pagani, and D. Stricker, "AutoPOSE: Large-scale automotive driver head pose and gaze dataset with deep head orientation baseline," in *Proc. 15th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, Valletta, Malta, Feb. 2020, pp. 599–606.

[33] L. Fridman, P, Langhans, J, Lee, and B. Reimer, "Driver gaze region estimation without use of eye movement," *IEEE Intell. Syst.*, vol. 31, no. 3, pp. 49–56, May/Jun. 2016.

[34] R. F. Ribeiro and P. D. P. Costa, "Driver gaze zone dataset with depth data," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Lille, France, May 2019, pp. 1–5.

[35] S. Ghosh, A. Dhall, G. Sharma, S. Gupta, and N. Sebe, "Speak2Label: Using domain knowledge for creating a large scale driver gaze zone estimation dataset," Apr. 2020, *arXiv:2004.05973*. [Online]. Available: http://arxiv.org/abs/2004.05973

[36] L. Fridman *et al.*, "MIT advanced vehicle technology study: Large-scale naturalistic driving study of driver behavior and interaction with automation," *IEEE Access*, vol. 7, pp. 102021–102038, 2019.

[37] V. Neale, T. Dingus, S. Klauer, J. Sudweeks, and M. Goodman, "An overview of the 100-car naturalistic study and findings," Nat. Highway Traffic Saf. Admin., Washington, DC, USA, Tech. Rep. 05-0400, Jun. 2005.

[38] A. Blatt *et al.*, "Naturalistic driving study: Field data collection," Transp. Res. Board, Tech. Rep. S2-S07-RW-1, Jul. 2014.

[39] J. H. L. Hansen, C. Busso, Y. Zheng, and A. Sathyanarayana, "Driver modeling for detection and assessment of driver distraction: Examples from the UTDrive test bed," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 130–142, Jul. 2017.

[40] P. Angkititrakul, M. Petracca, A. Sathyanarayana, and J. Hansen, "UTDrive: Driver behavior and speech interactive systems for in-vehicle environments," in *Proc. IEEE Intell. Vehicles Symp.*, Istanbul, Turkey, Jun. 2007, pp. 566–569.

[41] P. Angkititrakul *et al.*, "Getting start with UTDrive: Driver-behavior modeling and assessment of distraction for in-vehicle speech systems," in *Proc. Interspeech*, Antwerp, Belgium, Aug. 2007, pp. 1334–1337.

[42] E. Olson, "AprilTag: A robust and flexible visual fiducial system," in *Proc. IEEE Int. Conf. Robot. Automat.* Shanghai, China: IEEE, May 2011, pp. 3400–3407.

[43] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008.

[44] W. Kabsch, "A discussion of the solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Sect. A*, vol. 34, no. 5, pp. 827–828, Sep. 1978.

[45] S. Jha and C. Busso, "Estimation of driver's gaze region from head position and orientation using probabilistic confidence regions," Dec. 2020, *arXiv:2012.12754*. [Online]. Available: http://arxiv.org/abs/2012.12754

[46] N. Li and C. Busso, "Detecting drivers' mirror-checking actions and its application to maneuver and secondary task recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 4, pp. 980–992, Apr. 2016.

[47] T. Hulnhagen, I. Dengler, A. Tamke, T. Dang, and G. Breuel, "Maneuver recognition using probabilistic finite-state machines and fuzzy logic," in *Proc. IEEE Intell. Vehicles Symp.*, San Diego, CA, USA, Jun. 2010, pp. 65–70.

[48] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-D point sets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-9, no. 5, pp. 698–700, Sep. 1987.

[49] T. Hu, S. Jha, and C. Busso, "Temporal head pose estimation from point cloud in naturalistic driving conditions," *IEEE Trans. Intell. Transp. Syst.*, early access, May 6, 2021, doi: 10.1109/TITS.2021.3075350.

[50] T. Baltrušaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Xi'an, China, May 2018, pp. 59–66.

[51] T. Hu, S. Jha, and C. Busso, "Robust driver head pose estimation in naturalistic conditions from point-cloud data," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Las Vegas, NV USA, Oct. 2020, pp. 1176–1182.

[52] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA, Dec. 2017, pp. 5099–5108.

[53] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jun. 2017, pp. 77–85.

[54] T. Baltrusaitis, L. Li, and L.-P. Morency, "Local-global ranking for facial expression intensity estimation," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, San Antonio, TX, USA, Oct. 2017, pp. 111–118.

[55] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," Apr. 2017, *arXiv:1704.04861*. [Online]. Available: http://arxiv.org/abs/1704.04861

[56] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (And a dataset of 230,000 3D facial landmarks)," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, Oct. 2017, pp. 1021–1030.

Sumit Jha (Graduate Student Member, IEEE) received the B.Tech. degree in electronics and communication engineering from the National Institute of Technology (NIT), Trichy, India, in 2012, and the M.S. degree in electrical engineering from The University of Texas at Dallas (UTD), Richardson, TX, USA, in 2016, where he is currently pursuing the Ph.D. degree. Since 2015, he has been a part of the Multimodal Signal Processing (MSP) Laboratory, UTD. His research interest includes machine learning computer vision solutions for driver monitoring and in-vehicle safety systems.

Mohamed F. Marzban (Graduate Student Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from Cairo University, Egypt, in 2013 and 2016, respectively. He is currently pursuing the Ph.D. degree in electrical and computer engineering with The University of Texas at Dallas, Richardson, TX, USA. From 2013 to 2014, he was a Systems Research and Development Engineer with Intel Labs, Cairo, Egypt. From 2015 to 2016, he was a Software Engineer with Avidbeam Inc., Cairo. In 2018, he joined Qualcomm Technologies, Santa Clara, CA, USA, as a Connected Vehicles Intern. His research interests include machine/deep learning applications, advanced driver assistance systems, and wireless communications. He was a recipient of the Betty and Gifford Johnson Graduate Studies Award, Louis Beecherl Graduate Fellowship, and the Jan Van der Zeil Graduate Fellowship in 2017, 2018, and 2019, respectively.

Tiancheng Hu (Student Member, IEEE) received the B.Sc. degree in electrical engineering from The University of Texas at Dallas, Richardson, TX, USA, in 2020. From 2018 to 2019, he was the Assistant Director of Tutoring of the IEEE Student Branch at UTD. During his undergraduate studies, he also worked as an Undergraduate Researcher with the Multimodal Signal Processing (MSP) Laboratory. His current research interests include computer vision and perception for advanced driver-assistance systems. He was a recipient of the University-Wide Undergraduate Research Scholarship and the Jonsson School Undergraduate Research Award in 2019 and 2020, respectively.

Mohamed H. Mahmoud (Graduate Student Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from Cairo University, Egypt, in 2007 and 2016, respectively. He is currently pursuing the Ph.D. degree in electrical engineering with The University of Texas at Dallas, Richardson, TX, USA. From 2010 to 2016, he was with Wasiela, Cairo, where he involved in the development of the LTE-UE and DVB. From 2016 to 2019, he joined RIOT-MICRO, Cairo, Egypt, as a Principle Systems Engineer, where he involved in the development of the NB-IoT and LTE-M. His current research interests include machine/deep learning application, advanced driver assistance systems, and wireless communications.

Naofal Al-Dhahir (Fellow, IEEE) received the Ph.D. degree from Stanford University. He is currently a Erik Jonsson Distinguished Professor and the ECE Associate Head with The University of Texas at Dallas, Richardson, TX, USA. He was a Principal Member of Technical Staff at GE Research and the AT&T Shannon Laboratory. He is a co-inventor of 43 patents and coauthor of 470 articles. He is a fellow of the National Academy of Inventors. He was a co-recipient of the four IEEE best paper awards. He received the 2019 IEEE SPCE Technical Recognition Award, and he served as the Editor-in-Chief for IEEE TRANSACTIONS ON COMMUNICATIONS.

Carlos Busso (Senior Member, IEEE) is currently a Professor with the Department of Electrical Engineering, The University of Texas at Dallas (UTD), Richardson, TX, USA. His current research interests include human-centered multimodal machine intelligence and applications, the broad areas of affective computing, nonverbal behaviors for conversational agents, and machine learning methods for multimodal processing. He was a recipient of the NSF CAREER Award. In 2014, he received the ICMI Ten-Year Technical Impact Award. He also received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain), and the Best Paper Award at the AAAC ACII 2017 (with Yannakakis and Cowie).