DGAZE: Driver Gaze Mapping on Road

by

Isha Dua, Thrupthi Ann John, Riya Gupta, C V Jawahar

in

2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
: 1
-8

Report No: IIIT/TR/2020/-1

Centre for Visual Information Technology
International Institute of Information Technology
Hyderabad - 500 032, INDIA
October 2020

DGAZE: Driver Gaze Mapping on Road

Isha Dua, Thrupthi Ann John, Riya Gupta and C.V.Jawahar

***Abstract*— Driver gaze mapping is crucial to estimate driver attention and determine which objects the driver is focusing on while driving. We introduce DGAZE, the first large-scale driver gaze mapping dataset. Unlike previous works, our dataset does not require expensive wearable eye-gaze trackers and instead relies on mobile phone cameras for data collection. The data was collected in a lab setting designed to mimic real driving conditions and has point and object-level annotation. It consists of 227,178 road-driver image pairs collected from 20 drivers and contains 103 unique objects on the road belonging to 7 classes: cars, pedestrians, traffic signals, motorbikes, auto-rickshaws, buses and signboards.**

We also present I-DGAZE, a fused convolutional neural network for predicting driver gaze on the road, which was trained on the DGAZE dataset. Our architecture combines facial features such as face location and head pose along with the image of the left eye to get optimum results. Our model achieves an error of 186.89 pixels on the road view of resolution 1920×1080 pixels. We compare our model with state-of-the-art eye gaze works and present extensive ablation results.

I. Introduction

Accurate eye gaze prediction is important in advanced driver assistance systems for ensuring the safety of the driver by determining the attention and fatigue levels of the driver. Studies have shown that the saccade patterns of a driver may be used to identify if the driver is fatigued [1]. Driver inattention is one of the leading causes of road accidents in the world [2]. According to National Highway Traffic Safety Administration (NHTSA), 15% of crashes in the U.S. in 2015 were due to driver inattention [3]. A 100-car naturalistic driving study shows that 80% of all crashes and 65% of near-crashes involved driver inattention due to distraction, fatigue or just looking away [4].

Previous approaches to eye gaze tracking rely on expensive eye tracking devices such as SMI Eye Tracking Glass, Tobii Pro(eye glass), Eye Link 1000 Plus and Tobii Eye Tracker(non-eye glass tracker). These devices are costly and cumbersome, hence they are not suited for monitoring driver gaze in day-to-day scenarios, as they require wearing of these devices during test time. In this work, we mitigate these problems by providing DGAZE, a driver eye gaze dataset which does not require any expensive hardware during deployment of models trained with it. Our dataset consists of image pairs showing the view inside the car and outside, which can be collected by a dashboard-mounted mobile phone. Our dataset is collected in a lab setting that matches real driving conditions. The outside-car scenes are captured from cars driving in the city and contain a good variation of lighting and driving conditions. Our dataset has more than 100,000 image pairs, each one annotated with the point and object the driver is looking at. We collect data using 20 drivers and annotate 103 unique objects on the road belonging to 7 different classes such as car, pedestrian, signboard, traffic signal, etc. To counter the lack of eye gaze trackers, we project the road video in front of the subject (the 'driver') and ask them to look at a point that is annotated on the projected video. We collect the dataset in the lab setting as it is not feasible or safe to ask drivers to look at specific points while driving on real roads.

Affiliated with KCIS, Centre of Visual Information Technology, IIIT Hyderabad. Email: isha.dua@research.iiit.ac.in

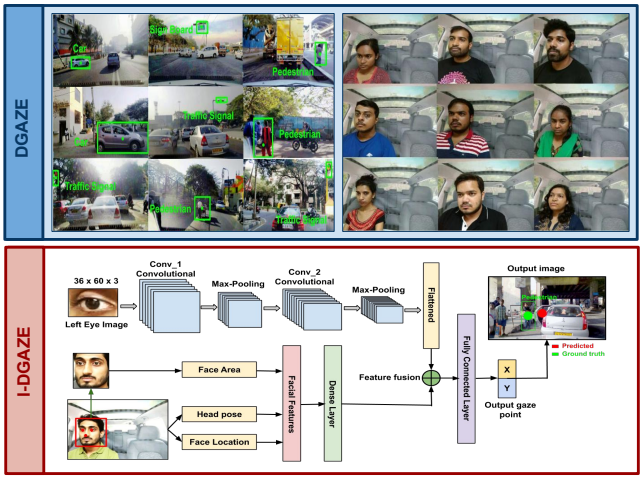

Fig. 1: In this work, we develop DGAZE the driver gaze mapping dataset that includes both driver and road view to capture driver gaze on the road using low-cost mobile phone cameras. Using DGAZE, we train I-DGAZE for prediction of gaze point on road.

Our dataset may be used in two ways: as point or object prediction of driver eye gaze. In the first task, the exact point where the driver is looking is predicted. This is useful for accurate and fine-grained eye gaze determination. Through this task, we may infer the saccade patterns of the driver on which further analysis such as the tiredness, attention or drowsiness of the driver may be inferred. In the second task, the object on which the driver looks is predicted. This is beneficial for an advanced driver assistance system by answering questions such as whether a sign is observed or whether the driver notices a pedestrian.

We also present I-DGAZE, a model for point prediction of driver eye-gaze, which is trained on the DGAZE dataset. Our architecture is a late-fusion convolutional neural network that uses the left eye image and facial features to predict the point location of eye gaze on the road video. Our

algorithm gives an error of 186.29 pixels on the road view of resolution 1920×1080 pixels. We show how we use calibration for each driver to reduce the pixel error down. Our calibration images are provided with the dataset. We compare our algorithm with state-of-the-art eye gaze algorithms and present extensive ablation results.

Our contributions are twofold:

1) We introduce a novel large-scale dataset for predicting driver eye gaze on the road. Our dataset provides both the driver and road camera views and is annotated with point and object-level ground truth. The dataset may be used for multiple purposes from predicting driver gaze on road to analyzing the objects of critical attention on road.
2) We propose I-DGaze, a baseline method for point prediction of driver eye-gaze on the road. We demonstrate how we use calibration to improve the results of our model.

II. Related Work

Eye Gaze Datasets Several eye gaze tracking datasets are available for training appearance-based gaze estimators. It can be categorized as either real-world or synthetic based on the image generation process, and as either eye-only or full-face based on image content. Real-world datasets contain images of real people taken as they gaze at different points in the environment. ColumbiaGaze contains images of 56 subjects with a discrete set of head poses and gaze targets [5]. EYEDIAP contains videos of 16 subjects gazing at targets on a screen or floating in 3D [6]. MPIIGaze contains images of 15 subjects when using their laptops [7]. GazeCapture contains images of 1,474 subjects taken while they were using tablets [8]. RT-GENE contains images of 15 subjects behaving naturally [9]. Commercial eye tracking glasses were used to provide ground truth gaze direction, and a GAN was used to remove the eye tracking glasses from the images. These real datasets are collected by looking at different points in the environment and costly trackers and sensors are used to collect it. Otherwise the real dataset is collected by gazing at tablets or phones. In contrast, our dataset collection does not require eye gaze trackers and our dataset provides both the road view and driver view. Other synthetic eye gaze datasets include UT-multiview [10], UnityEyes [11] datasets etc to render eye-only images at arbitrary head poses and gaze directions.

Gaze Estimation Past approaches to appearance-based gaze estimation have included Random Forests [10], k-Nearest Neighbors [10], and Support Vector Regression [12]. More recently, the use of deep CNNs to appearance-based gaze estimation has received increasing attention. Zhang et al. proposed the first deep CNN for gaze estimation in the wild [7], which gave a significant improvement in accuracy. Considering regions of the face outside the eyes has further improved accuracy. For example, Krafka et al. proposed a CNN with multi-region input, including an image of the face, images of both eyes and a face grid [8]. Parekh et al. proposed a similar multi-region network for eye contact detection [13]. Zhang et al. proposed a network that takes the full face image as input and adopts a spatial weighting method to emphasize features from particular regions of the face [14]. Other work has focused on how to extract better information from eye images. Cheng et al. developed a multistream network to address the asymmetry in gaze estimation between left and right eyes [15]. Yu et al. proposed to estimate the locations of eye landmark and gaze directions jointly [16]. Park et al. proposed to learn an intermediate pictorial representation of the eyes [17]. Lian et al. proposed to use images from multiple cameras [18]. Xiong et al.proposed mixed effects neural networks to tackle the problem caused by the non i.i.d. nature of eye tracking datasets [19]. A Palazzi et al. [20] proposes multi-branch deep architecture that integrates three sources of information: raw video, motion and scene semantics(Semantic Segmentation). The architecture is expensive and requires quite lot of time to train and test the model.

Driver Gaze Prediction Previous works in driver gaze prediction approached the problem in two ways. Works such as [21], [22], [23], [24], [25] focused on the interior of the car. They divide the inside of the car into separate zones and predicted which zone the driver was looking at. The head pose and gaze of the driver were collected using a video camera installed on the steering wheel column of the car. The system computes the gaze using the track landmark and a 3-D model [26]. Another work in this class is Deep Learning-Based Gaze Detection System for Automobile Drivers Using a NIR Camera Sensor [27]. The paper proposes a deep learning-based gaze detection method using a near-infrared (NIR) camera sensor considering driver head and eye movement that does not require any initial user calibration.

The other approach focuses solely on the regions outside the car. DR(eye)VE [20] predicts the driver's focus of attention on the road. The goal is to estimate what a person would pay attention to while driving, and which part of the scene around the vehicle is more critical for the task.They introduce large dataset of 500,000 frames consisting of driving scenes and corresponding eye-tracking annotations. The driving scenes are made by matching egocentric views (from glasses worn by drivers) and car-centric views (from roof-mounted camera), further enriched by other sensors measurements. In contrast to these approaches, our dataset is unique in providing both inside car view and outside car view.

III. DGAZE: Driver Gaze Mapping Dataset

A. Data Collection Set-up

In this section, we present DGAZE, a new dataset for mapping the driver's gaze onto the road. This is an important task as its output can be further processed to determine where the driver is focusing, whether the driver is inattentive (when the driver looks off the road for example) and to determine the fatigue level of the driver by observing their eye movements [1]. Currently, driver gaze datasets are collected using eye-tracking hardware which are expensive and cumbersome,

Fig. 2: DGAZE Collection Setup: Dataset is collected in lab setting which matches closely with real driving setting. A mobile phone camera is attached to tripod stand at a similar height and distance as a mobile phone camera mounted on the wind shield of a car. The phone uses its front and back cameras to collect both the driver view and projected road view at the same frame rate.

and thus unsuited for use during testing. Thus, our dataset is designed so that no costly equipment is required during test time. Models trained using our dataset requires only a dashboard-mounted mobile phone during deployment, as our data is collected using mobile phones. We collect the data in a lab setting with a video of a road projected in front of the driver. We overcome the limitation of not using eye trackers by annotating points on the road video and asking the drivers to look at them.

The task of driver eye-gaze can be solved as point-wise prediction or object-wise prediction. We provide annotation so that our dataset can be used for point-wise as well as object-wise prediction. Both types of eye-gaze prediction are useful. Predicting the object which the driver is looking at is useful for higher-level ADAS systems. This may be done by getting object candidates using an object detection algorithm and using the eye gaze to predict which object is being observed. Object prediction can be used to determine whether a driver is focusing on a pedestrian or if they noticed a signboard for example. Point-wise prediction is much more fine-grained and are more useful for nearby objects, as they show which part of the object is being focused on. They can be used to determine the saccade patterns of the eyes or to create a heatmap of the attention of a driver. They may even be converted into object-wise prediction. Our dataset allows both types of analyses to be conducted. We discuss the set-up used for collecting our dataset and the dataset statistics below.

The DGAZE dataset is collected in a lab which is set up to mimic real driving conditions. The layout of the lab is shown in Figure 2. A participant (henceforth referred to as 'driver') sits in front of a backdrop of the interior of a car. A video of a road having resolution 1920×1080 is projected in front of the driver. The video is collected using dash-board mounted cameras of cars driving on roads at different times of the day and differing traffic conditions. We mount a mobile phone on a tripod in front of the driver at a height and distance from the driver similar to that of a dash-board mounted phone. We use the front and rear cameras to simultaneously collect the videos of the driver and the projected video using the same frame-rate. The application used for recording is the HAMS application, previously used in [28], [29] for data collection. At every step, care is taken to make the set-up as realistic as possible. (See Figure 7 for examples of the driver-side videos.) In place of using expensive eye-gaze trackers that are cumbersome, we annotated a single object on each frame of the projected video with a bounding box and marked the center of it. (See Figure 3 for example frames). We asked the driver to look at the center of each object. Thus, the dataset can be used for gaze prediction at a point or object level.

The projected video has special calibration frames at the start and end, which serves to synchronise the projected video with the videos taken by the mobile phone. We align the videos by carefully dropping frames of the longer video. The following equation is used to compute the frame drop:

$$fd = \frac{max(fc(V_r), fc(V_d))}{|fc(V_r) - fc(V_d)|} \quad (1)$$

where fd denotes the number of frames after which a frame is dropped., fc denotes the frame count, V_r is the road video, and V_d denotes the driver video.

B. Dataset Collection

We collected road videos using mobile phones mounted on the dash-boards of cars driven in the city. We combined the road videos to create a single 18-minute video that had a good mix of road, lighting and traffic conditions. The road images have varied illumination as the images are captured from morning to evening in the real cars on actual roads. For each frame, we annotated a single object belonging to one of the classes: car, bus, motorbike, pedestrian, auto-rickshaw, traffic signal and sign board. We also marked the center of each bounding box to serve as the groundtruth for the point detection challenge. We annotated objects that typically take up a driver's attention such as relevant signage, pedestrians and intercepting vehicles. See Figure 3 for examples of annotated objects belonging to each class. In total, the collated video has 103 different objects annotated. The objects in the video are annotated using the dlib [30] correlation tracker implementation based on Accurate Scale Estimation for Robust Visual Tracking [30]. The tracking algorithm works in real-time by tracking the objects that change in both translation and scaling throughout a video sample.

We collected videos of 20 drivers in the 20-30 age group. The drivers were both male and female and included some people wearing eye glasses. Each driver was seated in front of the car back drop and was asked to gaze at the marked points of the 18-minute video. All drivers are given the same sequence of videos in the same order. In addition, we collected the videos of each driver looking at 9 calibration points spread equally across the screen.

Fig. 3: Samples from DGAZE dataset corresponding to seven unique objects annotated on road. The samples are collected such that there is significant variation in the size of the object, distance of the object from the driver and illuminance variation on road.

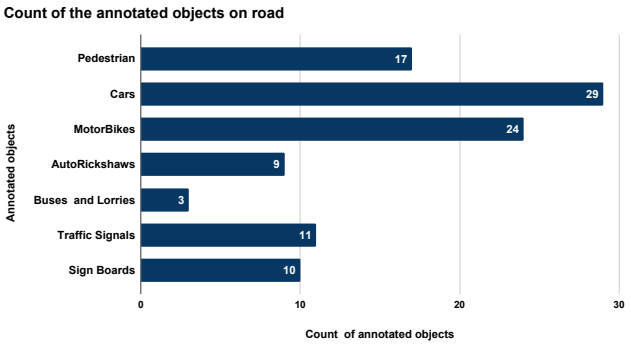

Fig. 4: Number of annotated objects in views corresponding to each unique object on road

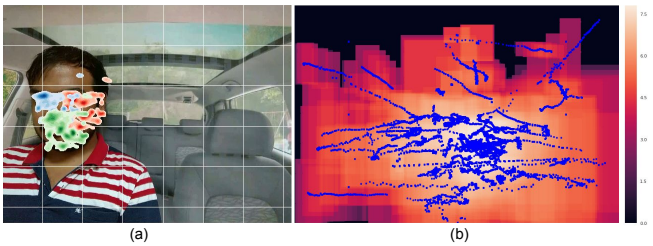

Fig. 5: Heatmaps depicting the spatial distribution of facial features and annotated objects in the dataset. (a) The spatial distribution of the left eye, right eye and mouth for the entire dataset is shown as red, blue and green heatmaps respectively. (b) The heatmap shows the distribution of annotated objects and the blue dots depict the ground truth point distribution.

C. *Dataset Statistics*

We collected 18-minute videos for each of the 20 drivers in our study. The frames were extracted to obtain approximately 100,000 dataset samples. Each sample consists of an image of the road with an annotated object, and an image of the driver looking at the center of the annotated object. We also provide samples corresponding to the 9 calibration points. Table I shows the comparison of DGAZE dataset with other existing datasets. Works such as [31], [32], [5], [6], [10], [7], [33], [8] collect eye gaze data by displaying predefined gaze points on a monitor display, mobile phone or tablet screen with the aim of predicting user gaze on these device screens. On the other hand, the proposed dataset can be used for driver gaze mapping on the road.

We have annotated 103 unique objects belonging to 7 classes in the 18-minute road video. Figure 4 shows the number of annotated objects in the views corresponding to each class. We show a few examples in Figure 3. The annotated objects are of various sizes on the screen. We have annotated objects that are very near as well as objects that are quite far on the road. Figure 5(b) shows a heat-map of the position of objects on the road video, as well as the position of points. As we can see, the objects cover a good portion of the video, except the top part (as the sky realistically does not contain many objects of interest.) The objects also move and leave realistic trails, which means the dataset may be used as a video dataset also.

Figure 5(a) shows how the position of the eyes and mouth vary in the dataset samples, We observe that there is a good

	Target Type	#People	Poses	Targets	Images
[31]	Monitor display	20	1	16	videos
[32]	Fixed Gaze Target	20	19	2-9	1,236
[5]	Fixed Gaze Target	56	5	21	5,880
[6]	Monitor display	16	cont.	cont.	videos
[10]	Monitor display	50	8+synth.	160	64,000
[7]	Laptop Screen	15	cont.	cont.	213,659
[33]	Mobile tablets	51	cont.	35	videos
[8]	Mobile and Tablet	1474	cont.	13+cont.	2,445,504
DGAZE	Projected Road View	20	cont.	9+cont.	100,000

TABLE I: Comparison of DGAZE with other eye gaze mapping dataset

variation of eye and mouth positions, but they remain within realistic bounds for drivers.

IV. I-DGAZE: Gaze Prediction Architecture

We present I-DGAZE, a model for point prediction of driver eye gaze trained on the DGAZE dataset. Accurate gaze estimation requires the knowledge of the position and angle of the head and the direction of gaze of the eye. Taking this into consideration, we design our model as a two-branch late-fusion network. The first branch takes as input facial features such as face location and head pose. The second branch takes in the left eye image as input. The output of the network is the x, y location of the eye gaze on the road video. In Section IV-A, we describe the facial features required for I-DGAZE and the algorithms used to extract these features. In section IV-B, we describe second branch of our network. Section IV-C describes the I-DGAZE architecture.

A. Feature Branch

This branch of the network takes as input several facial features so that the network may work out the position of the head. The features are computed with dedicated deep models trained for the corresponding tasks and fed as input to the feature branch. The face of the driver is first detected using DLIB [34] which gives the bounding box, as well as the facial landmarks, including the pupil locations. We codify the location using the coordinates of the face bounding box and the nose. The distance of the head from the screen is another important variable. Since we cannot measure it directly, we use the area of the facial bounding box as an approximation, with the assumption that the facial area does not vary much between drivers. Thus a large facial area may be considered to mean that the head is close to the camera and vice-versa for a small facial area. We find the head pose using [35]. We use the yaw, pitch and roll angles as input to indicate the head pose. The location of the pupils of the eye is also used as input to the feature branch, as it helps to determine the gaze direction. The final input vector has 10 elements which consist of the face area, roll, pitch and yaw of the head and the x,y coordinates of both the pupils and the nose.

B. Eye Branch

This branch takes in the cropped image of the left eye as input. The image is cropped and resized to size $36 \times 60 \times 3$. The eye region is obtained from the facial key landmarks

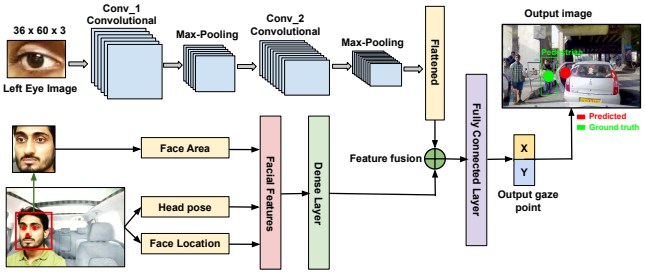

Fig. 6: I-DGAZE Architecture to predict driver gaze on road. The network is a two-branch late fusion convolutional neural network with input to one branch as eye image and input to the other branch as facial features like head pose, face location and distance of the driver's face from the mobile phone camera.

Left eye branch		
Layer	**Kernel**	**Output Shape**
Conv2D_1	3×3	$34 \times 58 \times 20$
MaxPool2d_1	2×2	$17 \times 29 \times 20$
Dropout		$17 \times 29 \times 20$
Conv2D_2	3×3	$15 \times 27 \times 50$
MaxPool2d_2	2×2	$7 \times 13 \times 20$
Flatten_1		4550

Feature branch		
Layer	**Kernel**	**Output Shape**
Dense_1		16

Fused branch		
Layer	**Kernel**	**Output Shape**
Merge_1		4566
Dense_2		500
Dense_3		2

TABLE II: The table shows the architecture of the I-DGAZE network, as shown in Figure 6.

detected around the eyes using the algorithm explained above. Here, the assumption is that the movement of the pupils of both eyes is correlated, so we need to use only one eye image.

C. Architecture and Training Details

The overall structure of the I-DGAZE model is given in Figure 6 and table II. We have used a modified version of the architecture proposed in MPIIGaze [7] The eye branch takes as input the image of the left eye of size $36 \times 60 \times 3$ which is passed through LeNet style model used in [36] where the image is first passed through a first convolutional layer with 20 channels followed by max-pooling and then passed to another convolutional layer with 50 channels followed by max-pooling and then the output from the max pool layer is flattened to get 4550-dimensional feature vector. The feature branch takes a 10-dimensional face feature vector as input which then goes through one fully-connected layers of output size 16. The feature branch and eye branch are merged into the common branch resulting in 4566 dimensions. This is then passed through one fully connected layers of output dimensions 500. The final 2-dimensional output vector cor-

responds to the x,y coordinates of the driver's eye gaze on the road view.

We use mean absolute error as our loss function. We train the network using Adam [37] with a learning rate of 1e-3 and weight decay of 1e-5. The network was trained for 60 epochs with a batch size of 32.

D. Calibration

The I-DGAZE model learns a relationship between the driver view and the road view. However, the position of the camera may change between sittings. Also the height of the driver and the positions of the eyes also vary. Thus, we perform an additional calibration step for each driver to create driver-specific models. We use 9 additional calibration points for each driver, which is included in the dataset. This comes to about 1000 image pairs for each driver.

The procedure for calibration is as follows. We first train a generic I-DGAZE model using the training procedure described in Section IV-C.We then freeze the left eye branch and the feature branch and finetune only the fused branch using the calibration frames of the specific user. The fine-tuning is performed using Adam optimizer with a learning rate of 1e-4 and weight decay of 5e-4 per epoch. and run for ten epochs to obtain the user-specific model. In Section V, we show that calibration greatly reduces the error of the model.

V. Experimental Evaluation and Results

In this section, we thoroughly evaluate the performance of I-DGAZE for point prediction task on the DGAZE dataset. I-DGAZE outperforms other state-of-the-art models in predicting driver gaze. The pixel error for driver gaze prediction is 186.89 pixels on average without calibration and is reduced to 182.67 pixels on calibration. We also present the qualitative results for the I-DGAZE architecture and an ablation study to evaluate the various components of I-DGAZE.

We split the dataset into train, validation and test sets as follows. Of the 20 drivers, 16 were used for training, 2 for validation and 2 were used for testing. The 18-minute road video consists of 103 smaller video sequences, each containing a single tracked object. Of these 60% sequences were used for training, 20% for validation and 20% for testing. Thus, our test data contains 2 unseen drivers and 20 unseen object tracks. In total, 98306 image pairs were used for training, 4779 for validation, and 3761 were used for testing. In addition, we used image-pairs from 9 sequences corresponding to the calibration points for each driver to be used in the calibration procedure.

A. Evaluation Metrics

We evaluate our results using mean absolute error between the ground truth gaze point and the predicted gaze point as used previously in [33], [8]. The error between the prediction and ground truth is measured in pixels. Note that the overall size of the road image is 1920×1080 pixels and the error should be interpreted relative to it. The mean absolute error is first observed by training I-DGAZE model using 95,000 images. This error is further reduced by fine-tuning the model using images corresponding to calibration points and creating driver-specific models.

B. Qualitative and Quantitative Results

Table III shows the quantitative results obtained by training I-DGAZE model on the DGAZE dataset and its comparison with state-of-the-art eye gaze models. We compare our results with methodologies proposed in TurkerGaze [38], where they use pixel-level face features as input and use Ridge Regression to estimate gaze point on the screen. Specifically, they rescaled each eye image to size 6×10 and then performed histogram normalization, resulting in a 120-D appearance feature vector for both eyes. The vector is then regressed to the gaze co-ordinates. We also compare our results with MPIIGaze [7], which has state-of-the-art results for eye gaze estimation in wild and Eye-tracking for Everyone [8] which predicts user gaze on phone and tablet. The implementation is adapted from the reference code provided by the authors. We observe that the performance of I-DGAZE is comparable to all the above approaches as it is a fusion of high-resolution pixel-level eye image in one branch and specific facial features relevant to the gaze prediction in the other branch.

Figure 7 presents the qualitative results of the predicted driver gaze fixation. From left to right: column 1 shows the driver image, column 2 shows road image, column 3 shows the annotated object as a green dot and the IDGAZE gaze prediction as a red dot. In column 4, we show the result of the I-DGAZE model on video sample and we observe that there may be some error between the ground truth gaze point and the prediction and results in avg error of 186.29 pixels on the road view of resolution 1920×1080 pixels. Column 5 shows the result of I-DGAZE after fine-tuning the model using images corresponding to calibration points for one driver. The error got reduced after calibration and hence the eye gaze model gives fine gaze prediction on road. We have provided additional gaze mapping results on short video clips with and without calibration in the video summary of this paper.

C. Ablation Study

We conduct ablation studies to understand the importance of the various inputs to our model. Our results are shown in Table IV. The first two rows use only the left and right eye images without the facial features. The gaze prediction error is higher for these models, indicating that the facial features provide crucial information regarding the global position of the face in the frame, which is required for gaze estimation. Rows 3 and 4 takes the left eye along with facial features as input. Our I-DGAZE model in the 4th row uses facial area in addition to facial landmarks and head pose, which serve to capture the distance between the driver's head and the camera. This gives a marginal improvement in results. Thus, we can see that the combination of inputs for the I-DGAZE model gives the best gaze prediction results.

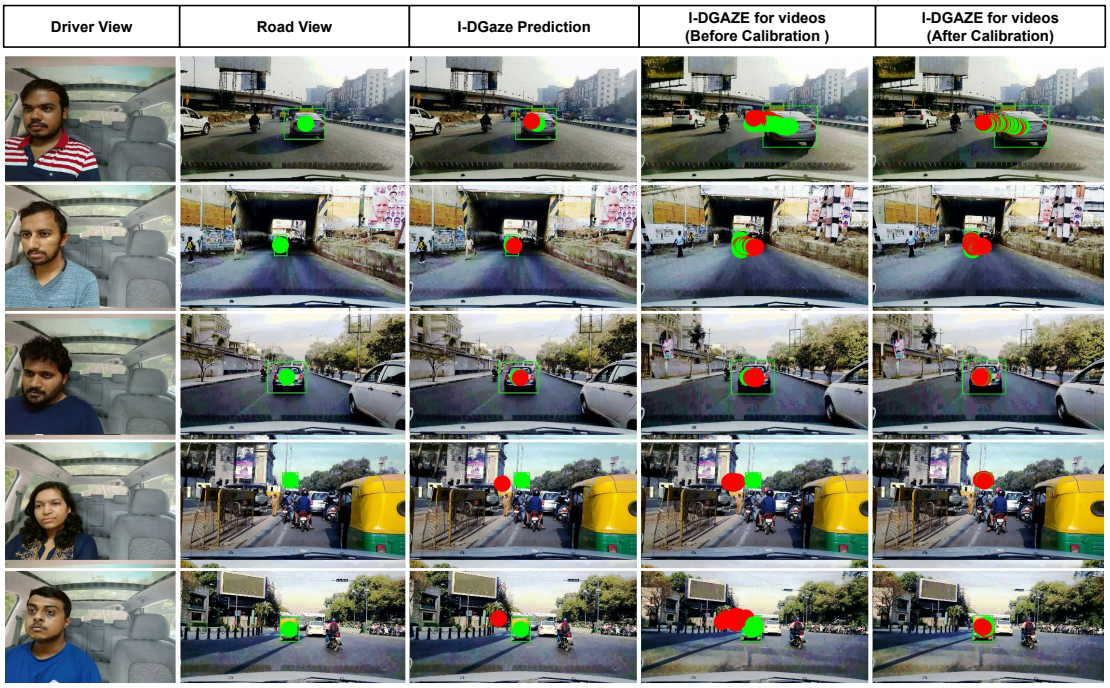

Fig. 7: Qualitative assessment of the predicted driver gaze fixation. From left to right: driver image, road image, I-DGAZE for gaze prediction in images, I-DGAZE for gaze prediction in multiple frames without calibration and I-DGAZE for gaze prediction in multiple frames with 9-point calibration.

Method	Without Calibration			Description
	Train Error	Val Error	Test Error	
Turker Gaze [38]	171.300	176.37	190.717	Pixel features + Ridge Regression
MPII Gaze [7]	144.32	229.0	189.63	CNN + head pose
iTracker [8]	140.1	205.65	190.5	fc1 of iTracker [8] + SVR
I-DGAZE	133.34	204.77	186.89	CNN + Facial Features

TABLE III: Comparison of I-DGAZE with existing gaze prediction methods. The facial features include location of the face, nose and pupils, head pose, face area and image of the left eye.

Method	Without Calibration			With Calibration	
	Train Error	Val Error	Test Error	Val Error	Test Error
LEye	151.07	200.07	202.6	176.70	206.28
LEye + REye	166.36	162.77	226.48	160.33	196.88
LEye + HP+ FL	136.58	199.73	187.77	185.48	186.50
LEye + HP + FL + FA (I-DGAZE)	133.35	204.77	**186.89**	187.18	**182.67**
LEye + REye + HP + FL + FA	117.05	198.08	201.26	186	200.5

TABLE IV: Ablation study of I-DGAZE model. Here, LEye = Left Eye, HP = Headpose, FL = Face Landmark Location and FA = Face Area

We also experiment with using both left and right eye images along with facial features. For this model, we take three branches for the left eye, right eye and features which are merged into one branch similar to the I-DGAZE model. Our experiments show that adding both eyes do not improve the gaze prediction. This is because the movement of both eyes are correlated and the facial features include the position of both pupils.

VI. Conclusion

In this work, we presented DGAZE, a large-scale dataset for driver eye gaze prediction on the road. Our dataset provides 227,178 image pairs of driver view and road view. The road view is collected in a range of lighting and driving conditions with the ground-truth gaze points covering a large area of the road view. DGAZE may be used for point-level or object-level gaze prediction. We also presented I-DGAZE, a deep model trained on DGAZE for point prediction of driver eye gaze. The architecture is a two-branch late fusion neural network that takes as input the image of the left eye of the driver as well as other facial features to give an accuracy of 186.89 pixels on the road view of resolution 1920×1080 pixels. We introduce a calibration procedure that allows us to create driver-specific models.

Unlike previous driver gaze systems, our dataset and model can be used to predict driver gaze without the use of any specialized hardware, thus reducing the cost of deployment and making the technology widely accessible. The dataset also enables the study of driver behavior such as road conditions that increase driver distraction, the way eye gaze changes for far or near objects and salient areas of the road which the driver pays attention to. We believe that our dataset and models will encourage the creation of better ADAS systems, thereby reducing the number of road accidents and improving driver safety. The code, data, and models are available at https://github.com/duaisha/DGAZE.git.

ACKNOWLEDGEMENT

We would like to thanks Akshay Uttama Nambi and Venkat Padmanabhan from Microsoft Research for providing us with resources to collect DGAZE dataset. This work is partly supported by DST through the IMPRINT program. Thrupthi Ann John is supported by Visvesvaraya Ph.D. fellowship.

REFERENCES

[1] M. C. Catalbas, T. Cegovnik, J. Sodnik, and A. Gulten, "Driver fatigue detection based on saccadic eye movements," in 2017 10th International Conference on Electrical and Electronics Engineering (ELECO). IEEE, 2017, pp. 913–917.

[2] T. Richard and D. Knowles, "Driver inattention: Drivers' perception of risks and compensating behaviours," IATSS Research, vol. 28, no. 1, pp. 89 – 94, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0386111214600959

[3] "NHTSA Distracted Driving," https://www.nhtsa.gov/risky-driving/distracted-driving.

[4] T. Dingus, S. Klauer, V. Lewis, A. Petersen, S. Lee, J. Sudweeks, M. Perez, J. Hankey, D. Ramsey, S. Gupta, C. Bucher, Z. Doerzaph, J. Jermeland, and R. Knipling, "The 100-car naturalistic driving study: Phase ii results of the 100-car field experiment," 01 2006.

[5] B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar, "Gaze locking: passive eye contact detection for human-object interaction," in UIST, 2013.

[6] K. A. Funes Mora, F. Monay, and J.-M. Odobez, "Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras," in Proceedings of the Symposium on Eye Tracking Research and Applications, ser. ETRA '14. New York, NY, USA: ACM, 2014, pp. 255–258. [Online]. Available: http://doi.acm.org/10.1145/2578153.2578190

[7] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun 2015. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2015.7299081

[8] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016, pp. 2176–2184.

[9] T. Fischer, H. Jin Chang, and Y. Demiris, "Rt-gene: Real-time eye gaze estimation in natural environments," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 334–352.

[10] Y. Sugano, Y. Matsushita, and Y. Sato, "Learning-by-synthesis for appearance-based 3d gaze estimation," in Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, ser. CVPR '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 1821–1828. [Online]. Available: https://doi.org/10.1109/CVPR.2014.235

[11] E. Wood, T. Baltrušaitis, L.-P. Morency, P. Robinson, and A. Bulling, "Learning an appearance-based gaze estimator from one million synthesised images," in Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications, 2016, pp. 131–138.

[12] T. Schneider, B. Schauerte, and R. Stiefelhagen, "Manifold alignment for person independent appearance-based gaze estimation," in 2014 22nd international conference on pattern recognition. IEEE, 2014, pp. 1167–1172.

[13] V. Parekh, R. Subramanian, and C. Jawahar, "Eye contact detection via deep neural networks," in International Conference on Human-Computer Interaction. Springer, 2017, pp. 366–374.

[14] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "It's written all over your face: Full-face appearance-based gaze estimation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 51–60.

[15] Y. Cheng, F. Lu, and X. Zhang, "Appearance-based gaze estimation via evaluation-guided asymmetric regression," in ECCV, 2018.

[16] Y. Yu, G. Liu, and J.-M. Odobez, "Deep multitask gaze estimation with a constrained landmark-gaze model," in ECCV Workshops, 2018.

[17] S. Park, A. Spurr, and O. Hilliges, "Deep pictorial gaze estimation," in ECCV, 2018.

[18] D. Lian, L. Hu, W. Luo, Y. Xu, L. Duan, J. Yu, and S. Gao, "Multiview multitask gaze estimation with deep convolutional neural networks," IEEE Transactions on Neural Networks and Learning Systems, vol. 30, pp. 3010–3023, 2018.

[19] Y. Xiong, H. woo Kim, and V. Singh, "Mixed effects neural networks (menets) with applications to gaze estimation," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7735–7744, 2019.

[20] A. Palazzi, D. Abati, S. Calderara, F. Solera, and R. Cucchiara, "Predicting the driver's focus of attention: the dr(eye)ve project," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018.

[21] A. Tawari and M. M. Trivedi, "Robust and continuous estimation of driver gaze zone by dynamic analysis of multiple face videos," 2014 IEEE Intelligent Vehicles Symposium Proceedings, pp. 344–349, 2014.

[22] 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," 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 988–994, 2014.

[23] B. Vasli, S. Martin, and M. M. Trivedi, "On driver gaze estimation: Explorations and fusion of geometric and data driven approaches," 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), pp. 655–660, 2016.

[24] A. Fridman, P. Langhans, J. Lee, and B. Reimer, "Driver gaze region estimation without use of eye movement," IEEE Intelligent Systems, vol. 31, pp. 49–56, 2015.

[25] A. Fridman, J. Lee, B. Reimer, and T. Victor, "'owl' and 'lizard': patterns of head pose and eye pose in driver gaze classification," ArXiv, vol. abs/1508.04028, 2016.

[26] F. Vicente, Z. Huang, X. Xiong, F. D. la Torre, W. Zhang, and D. Levi, "Driver gaze tracking and eyes off the road detection system," IEEE Transactions on Intelligent Transportation Systems, vol. 16, pp. 2014–2027, 2015.

[27] R. A. Naqvi, M. Arsalan, G. Batchuluun, H. S. Yoon, and K. R. Park, "Deep learning-based gaze detection system for automobile drivers using a nir camera sensor," in Sensors, 2018.

[28] I. Dua, A. Nambi, C. V. Jawahar, and V. Padmanabhan, "Autorate: How attentive is the driver?" in The IEEE Conference on Automatic Face and Gesture Recognition(FG2019), May 2019.

[29] I. Dua, A. U. Nambi, C. V. Jawahar, and V. N. Padmanabhan, "Evaluation and visualization of driver inattention rating from facial features," IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 2, no. 2, pp. 98–108, 2020.

[30] M. Danelljan, G. Hger, F. Shahbaz Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in Proceedings of the British Machine Vision Conference, 2014.

[31] C. D. McMurrough, V. Metsis, J. Rich, and F. Makedon, "An eye tracking dataset for point of gaze detection," in Proceedings of the Symposium on Eye Tracking Research and Applications, ser. ETRA '12. New York, NY, USA: ACM, 2012, pp. 305–308. [Online]. Available: http://doi.acm.org/10.1145/2168556.2168622

[32] U. Weidenbacher, G. Layher, P.-M. Strauss, and H. Neumann, "A comprehensive head pose and gaze database," pp. 455 – 458, 10 2007.

[33] Q. Huang, A. Veeraraghavan, and A. Sabharwal, "Tabletgaze: A dataset and baseline algorithms for unconstrained appearance-based gaze estimation in mobile tablets," ArXiv, vol. abs/1508.01244, 2015.

[34] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1867–1874.

[35] M. Patacchiola and A. Cangelosi, "Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods," Pattern Recognition, vol. 71, pp. 132 – 143, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320317302327

[36] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in Proceedings of the IEEE, 1998, pp. 2278–2324.

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[38] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.