Speak2Label: Using Domain Knowledge for Creating a Large Scale Driver Gaze Zone Estimation Dataset

Shreya Ghosh[1] Abhinav Dhall[1,2] Garima Sharma[1] Sarthak Gupta[3] Nicu Sebe[4]
[1]Monash University [2]Indian Institute of Technology Ropar [3]Kroop AI [4]University of Trento
{shreya.ghosh,abhinav.dhall,garima.sharma1}@monash.edu sarthak@kroop.ai
niculae.sebe@unitn.it

Abstract

Labelling of human behavior analysis data is a complex and time consuming task. In this paper, a fully automatic technique for labelling an image based gaze behavior dataset for driver gaze zone estimation is proposed. Domain knowledge is added to the data recording paradigm and later labels are generated in an automatic manner using Speech To Text conversion (STT). In order to remove the noise in the STT process due to different illumination and ethnicity of subjects in our data, the speech frequency and energy are analysed. The resultant Driver Gaze in the Wild (DGW) dataset contains 586 recordings, captured during different times of the day including evenings. The large scale dataset contains 338 subjects with an age range of 18-63 years. As the data is recorded in different lighting conditions, an illumination robust layer is proposed in the Convolutional Neural Network (CNN). The extensive experiments show the variance in the dataset resembling real-world conditions and the effectiveness of the proposed CNN pipeline. The proposed network is also fine-tuned for the eye gaze prediction task, which shows the discriminativeness of the representation learnt by our network on the proposed DGW dataset. Project Page: `https://sites.google.com/view/drivergazeprediction/home`

1. Introduction

One of the primary drivers of progress in deep learning based human behavior analysis is availability of large labelled datasets [31, 15]. It is observed that the process of labelling becomes non-trivial for complicated tasks. In this paper, we argue that by adding domain knowledge about the task during the data recording paradigm, one can automatically label the dataset quickly. The behavior task chosen in this paper is estimation of driver gaze in car. Distracted driving is one of the main causes of traffic accidents [6]. It is important to understand the far-reaching negative impacts of this killer, which is particularly common among younger drivers [6, 16]. According to a World Health Organization report [25], there were 1.35 million road traffic deaths globally in 2016 and it is increasing day by day. In order to prevent this, efforts are being made to develop *Advanced Driver Assistance Systems (ADAS)*, which will ensure smooth and safe driving by alerting the driver or taking control of the car (handover), when a driver is distracted or fatigued. One important information, which some ADAS needs is a driver's gaze behaviour, in particular, where is the driver looking? Over past few years, monitoring driver's behaviour as well as visual attention have become interesting topics of research [44, 20]. Analysis of driver's sparse gaze zone provides an important cue for understanding a driver's mental state. In vision-based driver behaviour monitoring systems, coarse gaze direction prediction instead of exact gaze location is usually acceptable [13, 42, 40, 36, 39, 41]. The coarse gaze regions are defined as the in-vehicle areas, where drivers usually look at while driving, for e.g. windshield, rear-view mirror, side mirrors, speedometer etc. As per recent studies [41, 13], head pose information is also relevant in predicting the gaze direction. This hypothesis fits well with real and natural driving behaviour. In many cases, a driver may move both head and eyes, while looking at a target zone. Accurate driver's gaze detection requires a very specific set of sensors [28, 44], which capture detailed information about the eyes and pupil movements but these can cause an unpleasant user experience. Additionally, manual data labelling is a tedious task, which requires time as well as domain knowledge. In this work, fully automatic labelling can be fairly quickly done by introducing speech during the dataset recording. Our Speech to Text (STT) based labelling technique reduces the above-mentioned limitations of the earlier works. Moreover, this paper can provide help in collaborative driving scenarios, while the vehicle operates in semi-autonomous mode. The **main contributions** of this paper are as follows:

- Traditionally, computer vision based datasets are either

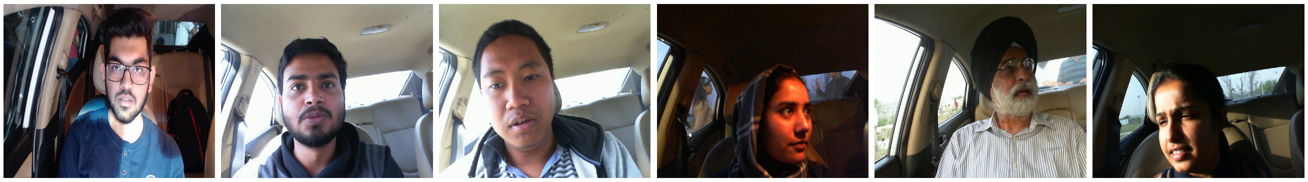

Figure 1. The proposed Driver Gaze in the Wild (DGW) dataset. Please note the different recording environments and age range.

Table 1. Comparison of in-car gaze estimation datasets.

Ref.	# Sub	# Zones	Illumination	Labelling
[2]	4	8	Bright & Dim	3D Gyro.
[18]	12	18	Day	Manual
[4]	50	6	Day	Manual
[35]	6	8	Day	Manual
[40]	10	7	Diff. day times	Manual
[13]	16	18	Day	Head-band
[41]	3	9	Day	Motion Sensor
Ours	338	9	Diff. day times	Automatic

labelled manually or using sensors, which makes the labelling process complicated. To the best of our knowledge, this is the first method for speech based automatic labelling of human behavior analysis dataset. The process of labelling in our paper is automatic and does not require any invasive sensors or manual labelling. This makes the task of collecting and labelling data with fairly large number of participants faster. Proposed method requires lesser time (∼30 sec) generating the labels as compared to manual labelling (∼10 min). Additionally, the voice frequency-based detection is used for extracting data samples missed by automatic speech to text method.

- An ‘in the wild’ dataset - Driver Gaze in the Wild (DGW) containing 586 videos of 338 different subjects is collected. To the best of our knowledge, this is the largest publicly available driver gaze estimation dataset (Fig. 1).
- As the dataset has been recorded with different illumination conditions, a convolutional layer robust to illumination is proposed.
- We have performed eye gaze representation learning to judge the generalization performance of our network. (See Supplementary Material)

2. Prior Work

With the progress in autonomous and smart cars, the requirement for automatic driver monitoring has been observed and researchers have been working on this problem for a few years now. For driver's attention estimation, eye tracking is the most evident solution. It is done via sensors or by using computer vision based techniques. Sensor based tracking mainly utilize dedicated sensor integrated hardware devices for monitoring driver's gaze in real-time. These devices require accurate pre-calibration and additionally these devices are expensive. Few examples of these sensors are Infrared (IR) camera [14], contact lenses [28], head-mounted devices [13, 12] and other systems [3, 49]. All of these above-mentioned systems have sensitivity towards outdoor lighting, difficulty in hardware calibration and system integration. Additionally, constant vibrations and jolts during driving can effect system's performance. Thus, it is worthwhile to investigate image processing based zone estimation techniques.

Prior studies for vision based gaze tracking are mainly focused on two types of zone estimation methods: head-pose based only [24, 41] and both head-pose and eye-gaze based [38, 35]. In an interesting work, Lee et al. [18] introduced a vision-based real-time gaze zone estimator based on a driver's head moment mainly composed of yaw and pitch. Further, Tawari et al. [36] presented a distributed camera based framework for gaze zone estimation using head pose only. Additionally, [36] collected a dataset from naturalistic on-road driving in streets, though containing six subjects only. For the gaze zone ground truth determination, human experts manually labelled the data. Driver's head pose provides partial information regarding the his/her gaze direction as there may be an interplay between eye ball moment and head pose [5]. Hence, methods totally relying on head pose information may fail to disambiguate between the eye movement with fixed head-pose. Later, Tawari et al. [35] combined head pose with horizontal and vertical eye gaze for robust estimation of driver's gaze zone. Experimental protocols are evaluated on the dataset collected by [36] and it shows improved performance overhead moment [36]. In another interesting work, Fridman et al. [4, 5] proposed a generalized gaze zone estimation using the random forest classifier. They validated the methods on a dataset containing 40 drivers and with cross driver testing (test on the unseen drivers). When the ratio of the classifier prediction having the highest probability to the second highest probability is greater than a particular threshold, the decision tree branch is pruned. Similarly, [38] combined 3D head pose with both 3D and 2D gaze information to predict gaze zone via a support vector machine classifier. Choi et al. [2] proposed the use of deep learning based techniques to predict categorized driver's gaze zone. Recently, Wang et al. [41] proposed an Iterative Closet Points (ICP) based head pose tracking method for appearance-based gaze estimation. The labelling is performed initially using a head motion sensor and later clustering is used on this head pose technique. The labels in this case, do not consider the scenario, where there is a difference between the eye gaze and the head pose of a subject. In another interesting work, Jha et al. [13] map 6D head pose (three head position and three head rotation angles) to an output map of quantized gaze zone. The users in their study wear headbands, which are used to label the data using the head pose information only. Few of the selected methods are also described in Table 1. Please note that most of the datasets are not available publicly, with the exception of Lee et al. [18], though it contains 12 subjects only. It is easily observable that our proposed dataset DGW has a large number of subjects and more diverse illumination settings. Further, the methods discussed above require either manual labelling of the driver dataset or it is based on a wearable sensor. We argue that the labelling of gaze can be noisy and erroneous task for labellers due to the task being monotonous. Further, with wearable sensors such a headband, it may be uncomfortable for some subjects. Therefore, in this work, we propose an alternate method of using speech as part of the data recording. This removes the need for manual labelling and the user having to wear any headgear as well. Similarly, we are interested in predicting the zone, where the driver is looking at? This considers the both eye gaze, head pose (Fig. 2) and the interplay of gaze and head pose. Nowadays, self-supervised and unsupervised learning is getting attention as it has the potential to overcome the limitation of supervised learning based algorithms, which requires large amount of labelled data. A few recent works [19, 22, 21, 10, 47, 46] explored this domain. The labelling technique used by us in this work also exploits the domain knowledge (speech) and helps in labelling a fairly large dataset quickly.

3. DGW Dataset

We curate a new driver gaze zone estimation dataset as the datasets in this domain are small in size and are mostly not available for academic purpose. Fig. 1 shows the frames from the proposed DGW dataset. Please note that the dataset and the baseline scripts will be made publicly available.

Data Recording Paradigm. Before data collection, consent was taken from participants regarding the scope of data usage. This included agreement to share data with university and industrial labs and if a face of a participant could be used in any publication in the future. We pasted number stickers on different gaze zones of the car (Fig. 3). The nine car zones are chosen from back mirror, side mirrors, radio, speedometer and windshield. The recording sensor used is a Microsoft Lifecam RGB. For recording the following protocol is followed: We asked the subjects to look at the zones marked with numbers in different orders. For each zone, the subject has to fixate on a particular zone number and speak the zone's number and then move to the next zone. For recording realistic behaviour, no constraint is mentioned to the subjects about looking by eye movements and/or head movements. The subjects choose the way in which they are comfortable. This leads to more naturalistic data (see Fig. 2). For the subjects who wear spectacles, if it is comfortable for the participant, they are requested to record twice i.e. with and without the spectacles. The RA was also present in the car and observed the subject and checked the recorded video. If there was a mismatch between the zone and the gaze, the subject repeated the recording. This insures correct driver gaze to car zone mapping.

The data is collected during different times of the day for recording different illumination settings (as evident in Fig. 1). Recording sessions are also conducted during the evening after sunset at different locations in the university. This enables different sources of illumination from street lights (light emitting diodes, compact fluorescent lamp and sodium vapour lamps) and also from inside the car. There are a few sessions during which the weather was cloudy. This brings healthy amount of variation in the data.

4. Automatic Data Annotation

As manual data labelling can be an erroneous and monotonous task, our method is based on automatic data labelling. Following are the details of the labelling process.

Speech To Text. Post extraction of the audio from the recorded samples, the IBM Watson's STT API [43] is used to convert the audio signal into text. We searched for the keywords 'one', 'two', 'three', 'four', 'five', 'six', 'seven', 'eight' and 'nine' in the extracted speech in ascending order. As we recorded the data in 'one' to 'nine' in sequence, therefore, sequentially ordered texts having high probability is considered. Further, we extracted the frames corresponding to the detected time stamps by adding an offset (10 frames chosen empirically) before and after the detection of the zone number. We used the US English model (16 kHz and 8 kHz). In a few cases, this model was unable to detect correctly, this may be due to different pronunciation of English words across different cultures. In order to overcome this limitation, we applied STT rectification.

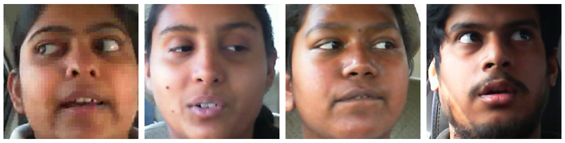

Figure 2. Challenging samples from our DGW dataset in which the head pose and eye gaze differ for subjects. The labels should not just be based on the head pose as in prior works [13, 41].

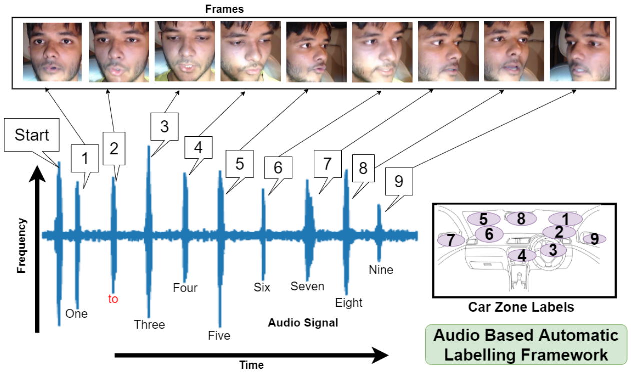

Figure 3. Overview of the automatic data annotation technique. On the top are the representative frames from each zone. Please note the numbers written in alphabets below the curve. The red colored 'to' shows the incorrect detection by the STT library. This is correct by frequency analysis approach in Sec. 4. On the bottom right are reference car zones.

STT Rectification. Generally, human voice frequency lies in the range of 300-3000 Hz [37]. We use the frequency and energy domain analysis of the audio signal to detect time duration of the audio signal from the data. The assumption based on the recording paradigm of DGW dataset is that the numbers are spoken in a sequence. If during scan of the numbers generated from the STT process, there is a mismatch for a particular number, the following steps are executed to find the particular zone's data: *Step 1:* Convert stereo input signal to mono audio signal and start scanning with a fixed window size T. Calculate frequency over the time domain of the audio for a window size T. *Step 2:* If the frequency lies in the human voice range 300-3000 Hz, then this window is a probable candidate. *Step 3:* Compute ratio between energy of speech band and total energy for this window. If ratio is above a threshold, then this window is a probable candidate. *Step 4:* If there is an overlap between the timestamps generated from steps 3 and 4 above, the zone label is assigned to the frames between the timestamps. This process extracted an extra 4000 frames, which were earlier missed due to the noise generated from STT. We checked manually for some recordings randomly, most of the useful data has been extracted following the steps above. *Refer supplementary document for dataset statistics and validation of automatic data annotation process.* Please note that the label generated after STT is treated as ground truth label.

4.1. Label Refining

Our proposed automatic data annotation may generate noisy labels during the gaze transition between two zones in the car. For example, a subject utters the word 'one' and looks at region 'one'. After that the subject shifts gaze from region 'one' to region 'two' and utters 'two'. During the transition between these two utterances, some frames may have been incorrectly annotated. Similarly, in a few cases, the zone utterance and the shifting of gaze may not have occurred simultaneously. To handle such situations, we perform label rectification based on an auto-encoder network followed by latent features based clustering.

Encoder-Decoder. The encoder part of the network is based on the backbone network (Inception-V1, refer Fig. 4). The decoder network consists of series of alternate convolution and up-sampling layers. The details of decoder network is as follows: the convolution layers have 1024, 128, 128, 64 and 3 kernels having 3 × 3 dimension. The first up-sampling layer has 2 × 2 kernel. The second and third up-sampling layers have 4 × 4 kernel. It is to be noted that the facial embedding representation learnt in this network encodes the eye gaze with the head pose information.

Clustering. After learning the auto-encoder, we perform k-means clustering on the facial embeddings. Here the value of k=9 is same as the number of zones. After k-mean clustering of all the samples, previous labels of the transition frames are updated on the basis of its Euclidean distance from the cluster center. More specifically, we measure the distance between a transition frame's facial embedding with the 9 cluster centers and assign the frame the label of the nearest neighbor cluster. Please note that it is a static process. The refined labels are then considered as the ground truth label for the dataset.

5. Method

Baseline. For the baseline methods, we experiment with several standard networks like Alexnet [17], Resnet [8] and Inception Network [34]. The input to the network is the cropped face computed using the Dlib face detection library. The proposed network is shown in Fig. 4. The baseline network takes 224 × 224 × 3 facial image as input. From the results using the standard networks mentioned above, one of the limitations observed is that as the DGW dataset has been recorded in diverse illumination conditions, some samples, which contained illumination change across the face are mis-classified. Sample images can be seen in Fig. 6 top. To the backbone network, we add the illumination layer presented below.

Illumination Robust Layer. For illumination robust facial image generation, we follow a common assumption proposed by Lambert and Phong [26]. The authors adopted the concept of ideal matte surface, which obey Lambert's cosine law. The law states that the incoming incident light at any point of an object surface is diffused uniformly in all possible directions. Later, Phong has added a specular highlight modelization term with Lambertian model. This term is independent of the object's geometric shape. Moreover, it is also independent of the lighting direction of each surface point. For illumination robust learning, we follow the computationally efficient Chromaticity property (Zhang et al. [50]). c = {r, g, b} is from the following skin color formation equation:

$$c_i = \frac{f_i {\lambda_i}^{-5} S(\lambda_i)}{(\prod_{j=1}^{3} f_j {\lambda_j}^{-5} S(\lambda_j))^{\frac{1}{3}}} \times \frac{e^{-\frac{k_2}{\lambda_i T}}}{e^{\frac{1}{3} \Sigma_{j=1}^{3} - \frac{k_2}{\lambda_j T}}} \qquad (1)$$

Here, i = {1, 2, 3}, correspond to the R, G & B channels, respectively. f_i is from Dirac delta function; λ_i's are tri-chromatic wavelengths (the wavelengths of R, G, B lights wherein $\{\lambda_1 \in [620, 750], \lambda_2 \in [495, 570], \lambda_3 \in [450, 495]$, unit : nm}); S(λ) is spectral reflectance function of skin surface; $k2 = \frac{hc}{k_B}$ (h: Plank's constant $h = 6.626 \times 10^{-34} J.s$, k_B: Boltzmann's constant $k_B = 1.381 \times 10^{-23} J.k^{-1}$ and $c = 3 \times 10^8 ms^{-1}$) refer to first and second radiation constants in Wien's approximation and T represents the lighting color temperature. If we write the Equation (1) in $c_i = A \times B$ format, then the left part of Equation 3 is the illumination robust (A) and right part (B) is illumination dependent due to the colour temperature factor T, which varies throughout the dataset. Thus, for illumination robust feature extraction, we initialize a constant kernel having the T independent value of part (B). T is initialized with a Gaussian distribution. Further, the product of constant and Gaussian kernel is considered for learning.

Attention based Gaze Prediction. The eye region of a person's face is important in estimating driver's gaze zone as it gives vital information about the eye gaze. We already show that there are images in DGW dataset (Fig. 2), where the head pose is frontal even though the driver may be looking at a particular zone, which is not in the front using the change in the eye gaze. Motivated by this hypothesis, we add attention augmented convolution module [1] to the network. Let's consider a convolution layer having F_{in} input filters, F_{out} output filters and k kernels. H and W represent the height and width of an activation map. d_v and d_k denote the depth of values and the depth of queries/keys in MultiHead-Attention (MHA). $v = \frac{d_v}{F_{out}}$ is the ratio of attention channels to number of output filters and $k_a = \frac{d_k}{F_{out}}$ is the ratio of key depth to number of output filters.

The **A**ttention **A**ugmented **Conv**olution (**AAConv**) [1] can be written as follows: $AAConv(X) = Concat[Conv(X), MHA(X)]$ Where, X is the input. MHA consists of a 1×1 convolution with F_{in} input filters and $(2d_k + d_v) = F_{out}(2k_a + v)$ output filters to compute queries/keys and values. An additional 1×1 convolution with $d_v = \frac{F_{out}}{v}$ input and output filters is also added to mix the contribution of different key heads. AAConv is robust to translation and different input resolution dimensions.

Network Architecture. The proposed network architecture is shown in the left box of Fig. 4. In this part of the network, we basically perform the gaze zone classification task with Inception-V1 as backbone network. The input of this network is facial image. The illumination robust layer and attention layer are introduced in the beginning and end of the backbone network to enhance the performance. After attention layer, the resultant embedding is passed through two dense layers (1024, 512) before predicting the gaze zone.

Table 2. Comparison of backbone networks on the proposed DGW dataset (validation set) with original labels (9 classes).

Networks	**Accuracy (%)**	
	Network	**Network + Illumination Robust Layer**
Alexnet	56.25	57.98
Resnet-18	59.14	60.87
Resnet-50	58.52	60.05
Inception-V1	60.10	61.46

6. Experiments

Data partition. The dataset is divided into train, validation and test sets. The partition is performed randomly. 203 subjects are used in training partition, 83 subjects are used in validation partition and rest of the 52 subjects are used in test partition. Having unique identities in the data partitions helps in learning more generic representations.

Experimental Setup. The following experiments were evaluated and compared to understand the complexities of the data and create baselines: 1) Baseline: based on Inception-V1 as the backbone network; 2) Baseline + Illumination Layer: On top of Inception-V1, an illumination robust layer is added; 3) Baseline + Attention: Attention augmented convolution layer is introduced in Inception-V1; 4) Baseline + Illumination Layer + Attention: This is the combination of illumination robust layer and attention; 5) Performance with standard backbone networks: Comparison of several state-of-the-art networks is performed; 6) Ablation study for illumination robust layer's configuration; 7) Eye-gaze representation learning: Transfer learning experiments to check the effectiveness of the representation learnt from DGW; 8) Evaluation on Nvidia Jetson Nano platform.

Evaluation Matrix and Training Details. Overall accuracy in % is used as evaluation matrix for gaze zone prediction. For gaze representation learning, the angular error (in °) is used as evaluation matrix for the CAVE dataset [30] and mean error (in cm) is used for the TabletGaze dataset [9]. For CAVE the angular error is calculated as $mean\ error \pm std.\ deviation$ (in °).

For the backbone network, Inception-V1 network architecture is used. For training the following parameters are used: 1) SGD optimizer with 0.01 learning rate with $1 \times e^6$ decay per epoch. 2) Kernels are initialized with Gaussian distribution with initial bias value 0.2. 3) In each case, the models are trained for 200 epochs with batch size 32.

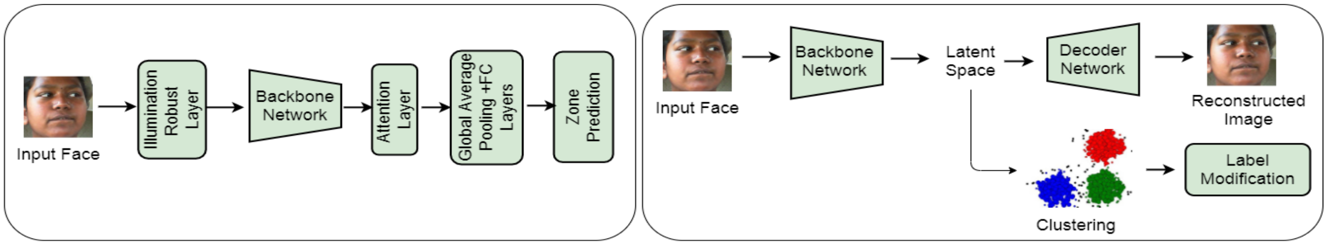

Figure 4. LEFT: Overview of the proposed network. RIGHT: Label refinement architecture.

For gaze representation learning, we fine tuned the proposed network with the following changes. Two FC layers (256, 256 node dense layers for the TabletGaze and 1024, 512 node dense layers for the CAVE dataset) are added with ReLU activation to fine tune the network. The learning rate was set to 0.001 with SGD optimizer. For both these datasets, we froze the first 50 layers of the network and fine tuned the rest part.

7. Results

7.1. Network Performance

Experiment with state-of-the-art backbone networks. We experimented with several network architectures to get an overview of the trade-off between the number of parameters and accuracy. Specifically, we choose lightweight networks like AlexNet [17], ResNet-18 [8], ResNet-50 and Inception [34]. Among these networks due to robust handling of different scales, the inception network performs better. The further results are based on the Inception-V1 as the backbone network. Based on this empirical analysis, the baseline network is the Inception-V1 network plus global average pooling and Fully Connected (FC) layers. The quantitative analysis of these networks are shown in Table 2. It also reflects that the addition of illumination robust layer increases the performance. It is effective in real-world scenarios in which different sources of illumination play vital role and many existing techniques may not perform properly. The classification performance increases as illumination robust layer is added to the baseline network as compared to the baseline network only.

9 zones vs 7 zones. Additionally, we experimented with a simpler task i.e. seven gaze zone classification. Although the data is collected with nine zones, zones 1 and 2 can be merged to represent the right half of the windscreen and zones 5 and zone 6 can be merged to represent the left half of the windscreen. We call this experiment setting as '7-zone'. Please refer to Fig. 3 (on Page 4) for car zone label reference. From Table 3, it is observed that the classification accuracy is higher in the case of 7 classes. The validation and test accuracies increase by 6.46% and 6.41%, respectively. This also means that for small sized cars fine-grained zone classification is non-trivial. Please note that all the following experiments in the paper are performed for 9 car zones only.

Improvement over baseline.

Quantitative Analysis. Table 3 shows the gradual improvement over the baseline due to the addition of attention and illumination layers. By adding the attention layer, we introduce guided learning. In the next step, we added an illumination robust layer to encode illumination robust features, which also increase the performance of the model. Our final model has both illumination and attention layer followed by FC layers.

Significance Test. One way ANOVA test is performed on the models is to calculate the statistical significance of the models. The p-values of the 'Baseline + Illumination', 'Baseline + Attention' and 'Baseline + Illumination + Attention' models are 0.03, 0.04 and 0.01, respectively. The p-values of the models are < 0.05, which indicates that the results are statistically significant.

Qualitative Analysis. Fig. 6a shows few examples, where previously mis-classified images are classified correctly after the addition of the illumination layer. We noted that for

Table 3. Comparison of the proposed network architecture with and without illumination robust layer and attention. Here, Base: Baseline, Illu: Illumination Layer and Attn:Attention.

Inception-V1 (Trained with original labels)		Accuracy (%)	
		Validation	**Test**
Baseline (7 classes)		66.56	67.39
(9 classes)	Base	60.10	60.98
	Base + Attn	60.75	60.08
	Base + Illu	61.46	60.42
	Base + Attn + Illu	64.46	62.90

Table 4. Variation in network performance (in %) w.r.t the illumination layer size and position.

Illumination Layer	**Layer Details**	**Accuracy (%)**	
		Validation	**Test**
Dense Layer	1024	56.16	57.93
	4096	58.51	61.18
Convolution Layer	32	53.48	52.16
	64	60.47	58.38
	128	**61.46**	**60.42**
	256	57.71	57.35

few subjects with spectacle glare, performance increased.

Label Rectification Performance. To avoid error in automatic labelling process, clustering based label modification (Sec. 7.1) was also performed.

Qualitative Analysis. After clustering, the labels of approximately 400 frames changed. Zone 9 class set changed most with frames in this zone increasing by 5.2%. Frames in Zone 8 and 5 increased by 3.5% and 2.3%. Other zones have less than 1% increment. This suggests that whenever there is a significant distance change in among consecutive zones, the error increases.

Quantitative Analysis. After label modification, the validation accuracy changes from 64.46% to 66.44% as shown in Table 5. This supports the hypothesis that the classification accuracy increases for frames in between transition from one zone to another and specially for the zones, with large physical distance (eg: zone 7 and 8).

Test Set Results. For all of the methods, test set performances was calculated. Both the 'Baseline + Illumination' and 'baseline+attention' perform slightly lower than the baseline (Table 3). The combined effect of illumination and attention improves the test set performance from 60.98% to 62.90%. Error in automatic labelling could be the cause for this performance. Further, training is performed on 'train+validation' set and performance is evaluated on test set. The test performance increased to 64.31% over the baseline (60.98%) and 'Baseline + Illumination + Attention' network (62.90%).

Results with State-of-the-art Methods. We evaluate our method (Sec. 5) on the LISA Gaze Dataset v1 [40], which contains 7 zones. Our method achieves 93.45% classification accuracy. [40]'s method gives 91.66% on their own data. This validates the discriminative ability of our proposed network. Further, we evaluate the method proposed by Vora et al. [40] on our data as well. The method achieves 67.31% and 68.12% classification accuracy on the validation and test sets, respectively. We evaluate standard networks [7, 33, 29, 11] and other state-of-the-art methods [39, 40, 38, 35, 4, 45] on DGW dataset in Table 6. It is observed that Resnet 152 [7] and Inception V3 [33] perform better than the others, however, the performance is not high as observed for other dataset such as [40]. This can be attributed to the large number of subjects (338) and different illumination conditions under which DGW was recorded.

Table 5. Results of the proposed methods. The 'Inception-V1 + Illumination + Attention' model is used for the experiments.

Methods	Accuracy (in %)		F1 Score	
	Validation	Test	Validation	Test
Proposed Network	64.46	62.90	0.52	0.52
Train on (Train + Val)	-	64.31	-	0.59
Label Modified	65.97	61.98	0.63	0.59

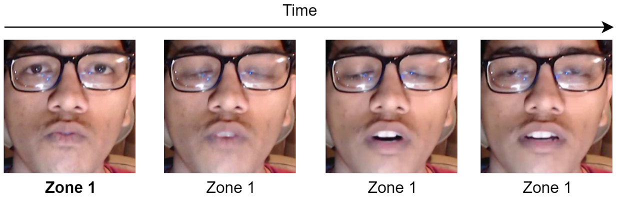

Figure 5. Gaze label assignment during an eye blink. The assigned labels (last three frames) are mentioned below each frame.

7.2. Ablation Study

Effect of Illumination layer. We also conducted experiments (Table 4)for observing the variation in network performance w.r.t the illumination layer variation.

1) Position Vs Performance. First, we experiment to check the ideal position of the illumination robust layer. For this analysis, the layer is implemented in the beginning and end conv-layers. For the beginning conv-layer, the performance increased. On the other hand, for the end conv-layer (before the flatten layer) performance did not increase. The reason could be that after several convolutions and max-pooling operation, the information is changed enough to be useful with the illumination robust layer.

2) Filter size Vs Performance. This analysis is conducted in two settings: 1) The robustness is implemented in convolutional layer and 2) The layer is implemented in dense layer (fully connected layer). From the Table 4, we can observe that as the illumination layer filter size increases, the performance also increases. We used the baseline + illumination + attention framework to compute these results.

Table 6. Performance comparison with existing CNN-based driver gaze estimation models.

Method	Val Acc (%)	Test Acc (%)
VGG 16 [29]	58.67	58.90
Inception V3 [33]	67.93	68.04
Squeezenet [11]	59.53	59.18
Resnet 152 [7]	68.94	69.01
Vora et al. [40]	67.31	68.12
Vora et al. (Alexnet face) [39]	56.25	57.98
Vora et al. (VGG face) [39]	58.67	58.90
Vasli et al. [38]	52.60	50.41
Tawari et al. [35]	51.30	50.90
Fridman et al. [4]	53.10	52.87
Yoon et al. (Face + Eyes) [45]	70.94	71.20
Lyu et al. [23]	85.40	81.51
Stappen et al. [32]	71.03	71.28
Yu et al. [48]	80.29	82.52

Qualitative Analysis of Failure Cases. Fig. 6 shows few mis-classified examples. The reason for this could be incomplete information of the eyes in these samples. All the cases (except bottom left in Fig. 6) eyes are not visible properly and illumination layer even unable to recover information. In these cases head pose and neighbour frame's gaze information could be vital clue to predict gaze zone. Additionally, there could be an interplay between pose and gaze.

Discussion on Effect of Eye Blink. Eye blinks are involuntary and a periodic event. During an eye blink event, a major cue for gaze estimation (i.e. pupil region) is missing. In the absence of the eye information (partially or fully closed eyes), the head pose may still provide useful cues required for gaze estimation. The same is also observed in some gaze datasets (example: Gaze360) i.e. the head pose information is considered as the gaze information in case of partial or complete occlusion scenarios. In our work, we too assume that head pose and eye information provide complementary information. In case of an eye blink, head pose information can be useful. To analyse this in the driver gaze context, we conduct following experiments. If we remove eye blink frames (detected using eye aspect ratio [27]) the validation accuracy improved by 6.27%, which means that partial eye close or fully closed samples are challenging. However, if we consider practical deployment scenario during which the driver gaze detection system will be used in a car, temporal information in the form of previous frames will also be available. So, there is a possibility of borrowing information from earlier frames, when the current frame has incomplete information due to an eye blink. A simple method is using labels of the neighbour previous frames (where eyes are open) for assigning them to frames containing eye blink. With this assignment, we note that the validation accuracy improves by 4.7%. This small experiment is an indication that in the presence of an eye blink, we can still consider the information from the previous frames, which leads to correct prediction of current gaze zone. An example is shown in the Fig. 5, here, the neighbour frame is the first frame in which the eyes are open. We assign the same labels for the subsequent frames (i.e. eye blink frames).

Other. Please refer the supplementary material for the ablation study regarding gaze representation, effect of lip movement and deployment on Jetson Nano environment.

8. Conclusion, Limitations and Future Work

In this paper, we show that automatic labelling can be performed by adding domain knowledge during the data recording process. We propose a large scale gaze zone estimation dataset, which is labelled fully automatically using the STT conversion. It is observed that the missed information from STT can be recovered by analyzing the frequency and energy of the audio signal. The dataset recordings are performed in different illumination conditions, which makes the dataset closer to the realistic scenarios. To take care of the varying illumination across the face, we propose an illumination robust layer in our network. The results show that the illumination robust layer is able to correctly classify some samples, which have different or low illumination. Further, the experiments on eye gaze prediction using the features learnt from our network on the DGW dataset show that the features learnt are effective for gaze estimation task. In order to record even more realistic data, car driving also needs to be added in the data recording paradigm. The trickier part is about how to use speech effectively in this case as the drivers will be concentrating on the driving activity. Perhaps, a smaller subset of driving dataset can be labelled using a network trained on the existing stationary recorded dataset and it can be validated by human labellers. Further, the DGW dataset will be extended with more female subjects to balance the current gender distribution. At this point, our method does not consider the temporal information. It will be interesting to understand the effect of temporal information on the gaze estimation as a continuous regression-based problem. Few prior works used IR cameras [14, 41] for their superior performance in dealing with illumination effects such as on the driver glasses. Our use of a webcam-based RGB camera validated the process of STT labelling and illumination invariance. It will be of interest to try distillation based knowledge transfer from our DGW dataset into the smaller sized network later fine-tuned on smaller gaze estimation datasets.

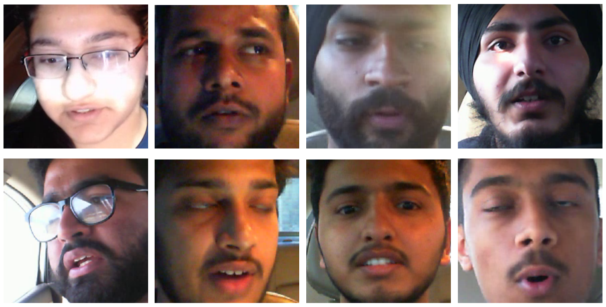

Figure 6. TOP: Correctly classified samples with illumination robust layer, which were earlier mis-classified by the baseline network. Bottom: Incorrectly classified samples by our network.

Currently, on Nvidia Jetson Nano, we achieve 10 FPS. It should further improve if network optimization techniques such as quantization and separable kernels are experimented with. Our proposed method is implicitly learning the discriminativeness, due to the head pose. In future, we plan to integrate the head pose information explicitly to evaluate its usefulness. We will also evaluate the performance of the network and the usefulness of the learnt features for the task of distracted driver detection. One future direction can be joint gaze zone and distraction detection as a multi-task learning problem.

References

[1] I Bello, B Zoph, A Vaswani, J Shlens, and Q V Le. Attention augmented convolutional networks. In *IEEE International Conference on Computer Vision*, pages 3286–3295, 2019.

[2] I H Choi, S K Hong, and Y G Kim. Real-time categorization of driver's gaze zone using the deep learning techniques. In *International Conference on Big Data and Smart Computing*, pages 143–148. IEEE, 2016.

[3] Yunlong Feng, Gene Cheung, Wai-tian Tan, Patrick Le Callet, and Yusheng Ji. Low-cost eye gaze prediction system for interactive networked video streaming. *IEEE Transactions on Multimedia*, 15(8):1865–1879, 2013.

[4] L Fridman, P Langhans, J Lee, and B Reimer. Driver gaze estimation without using eye movement. *IEEE Intelligent Systems*, pages 49–56, 2015.

[5] L Fridman, J Lee, B Reimer, and T Victor. 'owl'and 'lizard': patterns of head pose and eye pose in driver gaze classification. *IET Computer Vision*, 10(4):308–314, 2016.

[6] E Gliklich, R Guo, and R W Bergmark. Texting while driving: A study of 1211 us adults with the distracted driving survey. *Preventive Medicine Reports*, 4:486–489, 2016.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[8] K He, X Zhang, S Ren, and J Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[9] Q Huang, A Veeraraghavan, and A Sabharwal. Tabletgaze: unconstrained appearance-based gaze estimation in mobile tablets. *Machine Vision and Applications*, pages 445–461, 2015.

[10] Yifei Huang, Sheng Qiu, Changbo Wang, and Chenhui Li. Learning representations for high-dynamic-range image color transfer in a self-supervised way. *IEEE Transactions on Multimedia*, 2020.

[11] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size, 2016. cite arxiv:1602.07360Comment: In ICLR Format.

[12] S Jha and C Busso. Challenges in head pose estimation of drivers in naturalistic recordings using existing tools. In *IEEE International Conference on Intelligent Transportation Systems*, pages 1–6, 2017.

[13] S Jha and C Busso. Probabilistic estimation of the gaze region of the driver using dense classification. In *IEEE International Conference on Intelligent Transportation Systems*, pages 697–702, 2018.

[14] M W Johns, A Tucker, R Chapman, K Crowley, and N Michael. Monitoring eye and eyelid movements by infrared reflectance oculography to measure drowsiness in drivers. *Somnologie-Schlafforschung und Schlafmedizin*, 11(4):234–242, 2007.

[15] P Kellnhofer, A Recasens, S Stent, W Matusik, and A Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *IEEE International Conference on Computer Vision*, pages 6912–6921, 2019.

[16] Leah Knapp. "https://theharrispoll.com/pop-quiz-what-percentage-of-drivers-have-brushed-or-flossed-their-teeth-behind-the-wheel-while-its-crazy-to-think-that-anyone-would-floss-their-teeth-while-cruising-down-the-highway-it/". *Erie Insurance*, 2015.

[17] A Krizhevsky, I Sutskever, and G E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[18] S Lee, J Jo, H Jung, K Park, and J Kim. Real-time gaze estimator based on driver's head orientation for forward collision warning system. *IEEE Transactions on Intelligent Transportation Systems*, 12(1):254–267, 2011.

[19] M Leo, D Cazzato, T De Marco, and C Distante. Unsupervised eye pupil localization through differential geometry and local self-similarity. *Public Library of Science*, 9(8), 2014.

[20] Nanxiang Li, Jinesh J Jain, and Carlos Busso. Modeling of driver behavior in real world scenarios using multiple noninvasive sensors. *IEEE Transactions on Multimedia*, 15(5):1213–1225, 2013.

[21] X Liu, Joost Van D W, and Andrew D B. Exploiting unlabeled data in cnns by self-supervised learning to rank. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1862–1878, 2019.

[22] Feng Lu, Yue Gao, and Xiaowu Chen. Estimating 3d gaze directions using unlabeled eye images via synthetic iris appearance fitting. *IEEE Transactions on Multimedia*, 18(9):1772–1782, 2016.

[23] Kui Lyu, Minghao Wang, and Liyu Meng. Extract the gaze multi-dimensional information analysis driver behavior. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 790–797, 2020.

[24] Sankha S Mukherjee and Neil Martin Robertson. Deep head pose: Gaze-direction estimation in multimodal video. *IEEE Transactions on Multimedia*, 17(11):2094–2107, 2015.

[25] World Health Organization. World Health Organization. Technical report, 2016.

[26] B T Phong. Illumination for computer generated pictures. *Communications of the ACM*, 18(6):311–317, 1975.

[27] Souvik Ray. Eye Blink. Technical report.

[28] DA Robinson. A method of measuring eye movemnent using a scieral search coil in a magnetic field. *IEEE Transaction on Bio-Medical Electron.*, pages 137–145, 1963.

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[30] B A Smith, Q Yin, S K Feiner, and S K Nayar. Gaze locking: passive eye contact detection for human-object interaction. In *ACM symposium on User interface software and technology*, pages 271–280, 2013.

[31] Mohammad Soleymani, Martha Larson, Thierry Pun, and Alan Hanjalic. Corpus development for affective video indexing. *IEEE Transactions on Multimedia*, 16(4):1075–1089, 2014.

[32] Lukas Stappen, Georgios Rizos, and Björn Schuller. X-aware: Context-aware human-environment attention fusion

for driver gaze prediction in the wild. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 858–867, 2020.

[33] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015.

[34] C Szegedy, V Vanhoucke, S Ioffe, J Shlens, and Z Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[35] 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 *IEEE Conference on Intelligent Transportation Systems*, pages 988–994, 2014.

[36] A Tawari and M M Trivedi. Robust and continuous estimation of driver gaze zone by dynamic analysis of multiple face videos. In *IEEE Intelligent Vehicles Symposium*, pages 254–265, 2014.

[37] I R Titze and D W Martin. Principles of voice production, 1998.

[38] B Vasli, S Martin, and M M Trivedi. On driver gaze estimation: Explorations and fusion of geometric and data driven approaches. In *IEEE Intelligent Transportation Systems*, pages 655–660, 2016.

[39] S Vora, A Rangesh, and M M Trivedi. On generalizing driver gaze zone estimation using convolutional neural networks. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 849–854. IEEE, 2017.

[40] S Vora, A Rangesh, and M M Trivedi. Driver gaze zone estimation using convolutional neural networks: A general framework and ablative analysis. *IEEE Transactions on Intelligent Vehicles*, pages 254–265, 2018.

[41] Y Wang, G Yuan, Z Mi, J Peng, X Ding, Z Liang, and X Fu. Continuous driver's gaze zone estimation using rgb-d camera. *Sensors*, page 1287, 2019.

[42] Y Wang, T Zhao, X Ding, Ji Bian, and X Fu. Head pose-free eye gaze prediction for driver attention study. In *IEEE International Conference on Big Data and Smart Computing*, pages 42–46, 2017.

[43] IBM Watson. Speech to Text. Technical report, 2016.

[44] D Xia and Z Ruan. IR image based eye gaze estimation. In *IEEE ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, volume 1, pages 220–224, 2007.

[45] Hyo Sik Yoon, Na Rae Baek, Noi Quang Truong, and Kang Ryoung Park. Driver gaze detection based on deep residual networks using the combined single image of dual near-infrared cameras. *IEEE Access*, 7:93448–93461, 2019.

[46] Y Yu, G Liu, and J Odobez. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11937–11946, 2019.

[47] Y Yu and J Odobez. Unsupervised representation learning for gaze estimation. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–13, 2020.

[48] Zehui Yu, Xiehe Huang, Xiubao Zhang, Haifeng Shen, Qun Li, Weihong Deng, Jian Tang, Yi Yang, and Jieping Ye. A multi-modal approach for driver gaze prediction to remove identity bias. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 768–776, 2020.

[49] Cong Zhang, Qiyun He, Jiangchuan Liu, and Zhi Wang. Exploring viewer gazing patterns for touch-based mobile gamecasting. *IEEE Transactions on Multimedia*, 19(10):2333–2344, 2017.

[50] W Zhang, X Zhao, J Morvan, and L Chen. Improving shadow suppression for illumination robust face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):611–624, 2019.