On Generalizing Driver Gaze Zone Estimation using Convolutional Neural Networks

Sourabh Vora, Akshay Rangesh and Mohan M. Trivedi

Abstract—The knowledge of driver distraction will be important for self driving cars in the near future to determine the handoff time to the driver. Driver’s gaze direction has been previously shown as an important cue in understanding distraction. While there has been a significant improvement in personalized driver gaze zone estimation systems, a generalized gaze zone estimation system which is invariant to different subjects, perspective and scale is still lagging behind. We take a step towards the generalized system using a Convolutional Neural Network (CNN). For evaluating our system, we collect large naturalistic driving data of 11 drives, driven by 10 subjects in two different cars and label gaze zones for 47515 frames. We train our CNN on 7 subjects and test on the other 3 subjects. Our best performing model achieves an accuracy of 93.36% showing good generalization capability.

I. INTRODUCTION

According to a recent study [1] on ‘Takeover time’ in driverless cars, drivers engaged in secondary tasks exhibit larger variance and slower responses to requests to resume control. It is also well known that driver inattention is the leading cause of vehicular accidents. According to another study [2], 80% of crashes and 65% of near crashes involve driver distraction.

Surveys on automotive collisions [3], [4] demonstrated that drivers were less likely (30%-43%) to cause an injury related collision when they had one or more passengers who could alert them to unseen hazards. It is therefore essential for Advanced Driver Assistance Systems (ADAS) to capture these distractions so that the humans inside the car [5] can be alerted or guided in case of dangerous situations. This will ensure that the handover process between the driver and the self driving car is smooth and safe.

Driver gaze activity is an important cue to recognize driver distraction. In a study on the effects of performing secondary tasks in a highly automated driving simulator [6], it was found that the frequency and duration of mirror-checking reduced during secondary task performance versus normal, baseline driving. Alternatively, Ahlstrom et al. [7] developed a rule based 2-second ‘attention buffer’ framework which depleted when the driver looked away from the field relevant to driving (FRD); and it starts filling up when the gaze direction is redirected towards FRD. Driver gaze activity can also be used to predict driver behavior [8]. Martin et al. [9] developed a framework for modeling driver behavior and maneuver prediction from gaze fixations and transitions.

Thus, there exists a need for a continuous driver gaze zone estimation system.

While there has been a lot of research in improving personalized driver gaze zone estimation systems, there hasn’t been a lot of progress in generalizing this task across different drivers, cars, perspectives and scale. We make an attempt in that direction using Convolutional Neural Networks (CNNs). CNNs have shown tremendous promise in the fields of image classification, object detection and recognition. We study their effectiveness in generalizing driver gaze estimation systems through a large naturalistic driving dataset of 10 drivers consisting of 47515 frames. Data are captured in two different cars with different camera settings of field of view (Fig 1).

The main contributions of this work are: a) A systematic analysis of CNNs on generalizing driver gaze zone estimation systems b) Comparison of the CNN based model with some other state of the art approaches and, c) A large naturalistic driving dataset of 11 drivers with extensive variability to evaluate the two methods.
II. RELATED RESEARCH

Driver monitoring has been a long standing research problem in Computer Vision. For an overview on driver inattention monitoring systems, readers are encouraged to refer to a review by Dong et al. [10].

A prominent approach for driver gaze zone estimation is remote eye tracking. However, remote eye tracking is still a very challenging task in the outdoor environment. These systems [11], [12], [13], [14] rely on near-infrared (IR) illuminators to generate the bright pupil effect. This makes them susceptible to outdoor lighting conditions. Additionally, the hardware necessary to generate the bright eye effect hinders the system integration into the car dashboard. These specialized hardware also require a lengthy calibration procedure which is expensive to maintain due to the constant vibrations and jolts during driving.

Due to the above mentioned problems, vision based systems appear to be an attractive solution for gaze zone estimation. These systems can be grouped into two categories: Techniques that only use the head pose [15], [16] and those that use the driver’s head pose as well as gaze [17], [18], [19], [20]. Driver head pose provides a decent estimate of the coarse gaze direction. For a good overview of vision based head pose estimation systems, readers are encouraged to refer to a survey by Murphy-Chutorian and Trivedi [21]. However, methods which rely on head pose alone fail to discriminate between adjacent zones separated by subtle eye movement, like front windshield and speedometer. Using a combination of gaze and head pose was shown to provide a more robust estimate of gaze zones by Tawari et al. [17] for personalized gaze zone estimation systems. Fridman et al. [22], [23] take a step towards generalized gaze zone estimation by performing the analysis on a huge dataset of 40 drivers and doing cross driver testing. However, they employ a high confidence decision pruning of 10 i.e. they only make a decision when the ratio of the highest probability predicted by the classifier to the second highest probability is greater than 10. Because of the pruning step as well as missed frames due to inaccurate detection of facial landmarks and pupil, the decision making ability of their model is limited to 1.3 frames per second (fps) in a 30 fps video. A system with a low decision rate would miss several glances for mirror checks making it unusable for driver attention monitoring. Thus, there exists a need for a better system which generalizes well to different drivers for the gaze zone estimation task.

We take a step towards that direction using CNNs. There hasn’t been many research studies which use CNNs for predicting driver’s gaze. Choi et al’ [24] use a five layered CNN to classify driver’s gaze in 9 zones. However, to the best of our knowledge, they don’t do cross driver testing. In this study, we further systematize this approach by having separate subjects in train and test sets. Cross driver testing is particularly important as it better resembles the real world conditions where the system will need to run on subjects which it has not seen during training. We also evaluate our model across variations in camera position and field of view.

III. DATASET

Extensive naturalistic driving data was collected to enable us to train and evaluate our convolutional neural network model. Ten subjects drove two different cars instrumented with two inside looking cameras as well as one outside looking camera. The inside looking cameras capture the driver’s face from different perspectives: one is mounted near the rear view mirror while the other is mounted near the A-pillar on the side window. All cameras capture color video stream at a frame rate of 30 frames per second and a resolution of 2704 x 1524 pixels. The camera suite is time synchronized.

While only images from the camera mounted near the rear-view mirror were used for our experiments, the other views were given to to a human expert for labeling the ground truth gaze zone. Seven different gaze zones (Fig 2) are considered in our study, namely, front windshield, right, left, center console (infotainment panel), center rear-view mirror, speedometer as well as the state of eyes closed which usually occurs when the driver blinks.

The frames for each zone were collected from a large number of ‘events’ separated well across time. An event is defined as a period of time in which the driver only looks at a particular zone. In a naturalistic drive, the front facing events last for a longer time and also occur with maximum frequency. Events corresponding to zones like Speedometer or Rearview Mirror usually last for a very small time and are much sparse as compared to front facing events. The objective of collecting the frames from a large number of events is to ensure sufficient variability in the head pose and

---

**TABLE I: Dataset: Number of annotated frames, frames used for training and frames used for testing per gaze zone**

<table>
<thead>
<tr>
<th>Gaze Zones</th>
<th>Annotated frames</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forward</td>
<td>21522</td>
<td>3505</td>
<td>1023</td>
</tr>
<tr>
<td>Right</td>
<td>4216</td>
<td>3195</td>
<td>1021</td>
</tr>
<tr>
<td>Left</td>
<td>4751</td>
<td>3725</td>
<td>1022</td>
</tr>
<tr>
<td>Center Stack</td>
<td>4143</td>
<td>2831</td>
<td>1159</td>
</tr>
<tr>
<td>Rearview Mirror</td>
<td>4489</td>
<td>3533</td>
<td>956</td>
</tr>
<tr>
<td>Speedometer</td>
<td>4721</td>
<td>3580</td>
<td>1140</td>
</tr>
<tr>
<td>Eyes Closed</td>
<td>3673</td>
<td>2565</td>
<td>1093</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>47515</strong></td>
<td><strong>22934</strong></td>
<td><strong>7414</strong></td>
</tr>
</tbody>
</table>

---

![Fig. 2: Gaze zones considered in this study](image-url)
Fig. 3: An overview of the proposed pipeline. It consists of two major blocks namely the Input Pre-processing Block and the Network Finetuning Block. One of the three region crops and one of the two networks are chosen for training and testing.

pupil location in the frames as well as to obtain highly varied illumination conditions.

Fig 1 shows some sample instances of drivers looking at different gaze zones. The videos were deliberately captured for different drives in different settings of fields of view (wide angle vs normal). The subjects also adjusted the seat position according to their comfort. We believe that all such variations in the dataset are necessary to build a robust model that generalizes well.

Since the forward facing frames dominate the dataset, they are sub sampled so as to create a balanced dataset. Further, the dataset is divided such that the drives from 7 subjects are used for training while the drives from 3 subjects are used for testing our model. Table I shows the number of frames per zone finally used in our train and test datasets.

IV. METHODOLOGY

CNNs are good at transfer learning. Oquab et al. [25] showed that image representations learned with CNNs on large-scale annotated datasets can be efficiently transferred to other visual recognition tasks with limited amount of training data. We attempt to finetune two CNNs originally trained on the ImageNet dataset [26]. We consider the following options: a) AlexNet [27] and b) VGG with 16 layers [28].

Fig. 3 describes the block diagram of our complete system. It consists of two major blocks namely: a) Input pre-processing block and b) Network finetuning block. The input pre-processing block extracts the portions of the raw input image that are most relevant to gaze zone estimation. The network fine tuning block then finetunes the ImageNet trained CNNs using the sub images output by the input pre-processing block. Both blocks are described in greater detail in sections IV-A and IV-B.

A. Training

We remove the last layer of the network (which has 1000 neurons) from both architectures and add a new fully connected layer with 7 neurons and a softmax layer on top of it. We initialize the newly added layer using the method proposed by He et al. [29]. We fine tune the entire network using our training data. Since the networks are pre-trained on a very large dataset, we use a low learning rate. For both networks, we start with a hundredth of the learning rate used to train the respective networks and observe the training and validation loss and accuracy. If the loss function oscillates, we further decrease the learning rate. It was found that a learning rate of $10^{-4}$ works well with both the networks. The networks were fine tuned for a duration of 5 epochs with mini batch gradient descent using adaptive learning rates. Based on GPU memory constraints, batch sizes of 64 and 32 were used for training AlexNet and VGG16 respectively. The Adam optimization algorithm, introduced by Kingma and Ba [30], was used.

B. Input to the CNNs

We choose three different approaches for preprocessing the inputs to the CNNs. In the first case (Fig 4b), driver’s face was detected and used as an input. The face detector presented by Yuen et al. [31] was used. In the second case, some context was added to driver’s face by extending the face bounding box in all directions (Fig 4c). Context has given a boost in performance in several computer vision problems and this input strategy will help us determine whether adding context to the face bounding box will improve the performance of the CNN. In the third case, only the top half of the face was used as an input (Fig 4a). The cropped images were all resized to 224x224 or 227x227 according to the network requirements and finally, the mean was subtracted.

V. EXPERIMENTAL ANALYSIS & DISCUSSION

The evaluation of the experiments performed in IV are presented using three metrics. The first two forms of evaluation metrics are the weighted and unweighted accuracy.
TABLE II: Weighted accuracy for both networks when presented with different image region crops. Cross driver testing was performed for each experiment; drives by 7 subjects were used for training while drives by 3 different subjects were used for testing.

<table>
<thead>
<tr>
<th></th>
<th>Half Face</th>
<th>Face</th>
<th>Face+Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlexNet</td>
<td>88.91</td>
<td>82.08</td>
<td>75.56</td>
</tr>
<tr>
<td>VGG16</td>
<td>93.36</td>
<td>92.74</td>
<td>91.21</td>
</tr>
</tbody>
</table>

They are calculated as:

\[
\text{Weighted Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{True positive}_i}{\text{Total Population}_i} \quad (1)
\]

\[
\text{Unweighted Accuracy} = \frac{\sum_{i=1}^{N} \text{True positive}_i}{\sum_{i=1}^{N} \text{Total Population}_i} \quad (2)
\]

where, \( N \) = Number of gaze zones.

The third evaluation metric is the N class confusion matrix where each row represents true gaze zone and each column represents estimated gaze zone.

A. Analysis of the networks and face bounding box size

Table II presents the weighted accuracy obtained on the test set for different combinations of networks and input pre-processing. Our best performing model achieves an accuracy of 93.36% clearly demonstrating the generalization capabilities of the features learned through CNN. Two trends are clearly observable based on the results. First, the performance of both the networks increases as the face bounding box size is reduced. Second, finetuned VGG16 outperforms finetuned AlexNet for all input pre-processing forms.

The low performance of AlexNet can be attributed to large kernel size (\( 11 \times 11 \)) and a stride of 4 in the first convolution layer. The gaze zones change with very slight movement of the pupil or eyelid. This fine discriminating information of the eye is missed out in the first layer due to convolution with large kernel size and a stride of 4. In our experiments, we found that the network easily classifies zones with large head movement (left and right) whereas it struggles to classify zones with slight eye movement (Eg. Front, Speedometer and Eyes Closed). The large increase in accuracy when only the top half of the face is provided as an input as compared to when the large Face+Context sub image is provided further confirms the fact.

VGG16 is composed of convolution layers that perform \( 3 \times 3 \) convolutions with a stride of 1. These small convolution kernels coupled with the larger depth of the network allows for discriminating gaze zones with even slight movements of the pupil or eyelid. The advantage of small \( 3 \times 3 \) kernel size is clearly visible when we evaluate the performance of both the networks fine tuned on Face+Context images. While the performance of AlexNet decreases significantly to 75.56% from 88.9% when trained on Face+Context images as compared to when trained on HalfFace images, this is not the case with VGG16. For VGG16, there is only a slight drop from 93.36% to 91.21%. This shows that small \( 3 \times 3 \) kernels help preserve the fine discriminating features of the eye, even when the eyes are such a small part of the image.

B. Comparison of our CNN based model with some current state of the art models

In this section, we compare our best performing model (VGG16 trained on upper half of face images) with some other recent gaze zone estimation studies. The technique presented by Tawari et al. [17] was implemented on our dataset so as to enable a fair comparison. They use a Random Forest classifier with hand crafted features of head pose and gaze surrogates which are calculated using facial landmarks.

Table III presents the confusion matrix for 7 gaze zones using the Random Forest model. Drives by 7 subjects were used for training while drives by 3 different subjects were used for testing.

<table>
<thead>
<tr>
<th>True Zone</th>
<th>Recognized Gaze Zone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forward</td>
<td>95.31 0.2 1.56 0.97 1.00 0.99 0.98</td>
</tr>
<tr>
<td>Right</td>
<td>0 99.51 0 0 0.1 0 0.39</td>
</tr>
<tr>
<td>Left</td>
<td>1.96 0 85.71 0 0 0.1 12.23</td>
</tr>
<tr>
<td>Center Stack</td>
<td>0.17 6.56 0.35 87.58 0.26 4.92 0.17</td>
</tr>
<tr>
<td>Rearview Mirror</td>
<td>0.01 2.2 0 0.21 0.09 0.61 0.09</td>
</tr>
<tr>
<td>Speedometer</td>
<td>1.49 0.61 0 5.27 0 99.87 1.76</td>
</tr>
<tr>
<td>Eyes Closed</td>
<td>0.91 0.82 0.18 0.73 0.09 2.2 95.06</td>
</tr>
</tbody>
</table>

Weighted Accuracy = 93.36%
Unweighted Accuracy = 93.17%

Table IV: Confusion matrix for 7 gaze zones using the Random Forest model. Drives by 7 subjects were used for training while drives by 3 different subjects were used for testing.

<table>
<thead>
<tr>
<th>True Zone</th>
<th>Recognized Gaze Zone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forward</td>
<td>84.16 0.7 0.68 1.47 5.38 0.59</td>
</tr>
<tr>
<td>Right</td>
<td>0 99.51 0 0 0.39 0 0.49</td>
</tr>
<tr>
<td>Left</td>
<td>6.17 0 71.17 0.33 0.67 1.83 19.83</td>
</tr>
<tr>
<td>Center Stack</td>
<td>0.78 8.57 0 32.55 15.41 0 42.68</td>
</tr>
<tr>
<td>Rearview Mirror</td>
<td>0 0.84 0 0.21 98.74 0 0.21</td>
</tr>
<tr>
<td>Speedometer</td>
<td>27.81 0 6.84 2.89 0 40.96 21.49</td>
</tr>
<tr>
<td>Eyes Closed</td>
<td>6.99 4.56 11.36 10.87 7.18 4.37 54.66</td>
</tr>
</tbody>
</table>

Weighted Accuracy = 67.86%
Unweighted Accuracy = 67.15%
matrix obtained by the Random Forest Model.

We see that our CNN based model clearly outperforms the Random Forest model by a substantial margin of 24.6%. There are several factors responsible for the low performance of the Random Forest model. The biggest ones are the position and orientation of the driver with respect to the camera. The Random Forest model relies on the head pose and gaze angles to discriminate between different gaze zones and these angles are not robust to the position and orientation of the driver with respect to the camera. Further, for determining the eye openness, the area of the upper eyelid was used as a feature which also changes with different subjects, different seat position and camera settings. All these factors combined limit the Random Forest model with hand crafted features to generalize as shown by the results on our dataset. Further, the accuracy of 68.76% was calculated when only the frames which pass the landmarks and pupil detection steps were considered. Because the classifier cannot make a prediction for frames which do not pass the above mentioned steps, these frames should ideally be considered as misclassifications. The weighted accuracy of the Random Forest model when calculated using this scheme further drops down to 64.1%. Inaccurate estimation of these intermediate tasks was also seen to seriously limit the performance in [22], [23]. In our CNN based approach there is no dependency on accurate facial landmark estimation and pupil detection which is another huge advantage over the Random Forest approaches.

We also compare our work with Choi et al. [24]. They trained a truncated version of AlexNet and achieved a high accuracy of 95% on their dataset. However, to the best of our knowledge, they don’t do cross driver testing and divide each drive temporally. The first 70% frames for each drive were used for training, next 15% frames were used for validation and the last 15% were used for testing. In our experiments, we show that AlexNet does not perform very well as compared to VGG16. We replicated their experimental setup by dividing our drives temporally and achieve a very high accuracy of 98.5% using AlexNet trained on face images. This clearly shows that the network is learning driver specific features and therefore overfits to the subjects.

Finally, to further evaluate the generalization ability of our CNNs, tests were also performed for subjects wearing glasses in a leave-one-subject-out fashion. The accuracy obtained by both the networks was only slightly less (less than 3%) than the accuracy seen when the networks were tested on subjects not wearing glasses (Table II). These results are very promising as the traditional approach of first detecting landmarks and pupil seriously suffer when the subjects are wearing glasses. Extensive analysis still needs to be performed as there were only two subjects in our dataset who wore glasses. We plan to do that in the future.

VI. CONCLUDING REMARKS

Correct classification of driver’s gaze is important as alerting the driver at the ‘correct time’ can prevent several road accidents. It will also help autonomous vehicles to determine driver distraction so as to calculate the appropriate takeover time. In literature, a large progress has been made towards personalized gaze zone estimation systems but not towards systems which can generalize to different drivers, cars, perspective and scale. Towards this end, this research study uses CNNs to classify driver’s gaze into seven zones. The model was evaluated on a large naturalistic driving dataset (NDS) of 11 drives, driven by 10 subjects in 2 separate cars. Two separate CNNs (AlexNet and VGG16) were fine tuned on the collected NDS using three different input pre-processing techniques. VGG16 was seen to outperform AlexNet because of the small kernel size ($3 \times 3$) in the convolution layer. Further, it was seen that the input strategy of using only the upper half of the face works better as compared to when the entire face or face+context images were used. Our best performing model (VGG16 finetuned on Half Face images) achieves an accuracy of 93.36% which shows tremendous improvement when compared to some recent state of the art techniques. Future work in this direction will be towards adding more zones and utilizing temporal context.

VII. ACKNOWLEDGMENTS

The authors would like to specially thank Dr. Sujitha Martin, Kevan Yuen and Nachiket Deo for their suggestions to improve this work. The authors would also like to thank our sponsors and our colleagues at Laboratory of Intelligent and Safe Automobiles (LISA) for their massive help in data collection.

REFERENCES


