Driver Attention Monitoring using Facial Features

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in Computer Science and Engineering
By Research

by

Isha Dua
20162081
isha.dua@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
May 2020

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled " **Driver Attention Monitoring using Facial Features**" by **Isha Dua**, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Adviser: Prof. C.V.Jawahar

To

Family and Friends

Acknowledgments

This thesis is the outcome of support, guidance, and encouragement I received from several people.

Firstly, I would like to thank my advisor Prof. C.V. Jawahar for all the guidance and support. He is a constant source of inspiration, and his guidance has been critical to my development as a researcher in computer vision and machine learning. Thank you for pushing me beyond what I thought were my limits. I am able to graduate with a framework that will carry me throughout my career, which is something that I am eternally grateful for. It was both an honor and a great privilege to work with him. He has helped me grow as a researcher and person. I have always defined him as "a perfect doctor" for me because he guides me to learn and solve the problem rather than just saying, "this is the problem." His passion for teaching motivated me to learn more and more. I am incredibly pleased to get his guidance and education in my life.

I would like to extend my thanks to Principal Researcher Venkat Padmanabhan and Researcher Akshay Uttama Nambhi from Microsoft Research, India, for collaborating with me on the work I have done in this thesis. Venkat sir is an inspiration, and his perspective and suggestions always give new dimensions to the discussion. I am very thankful to Akshay for being an excellent mentor during my time at MSR. He has a unique sense of resolving difficult problems and providing clarity. I learned a lot from him, and his guidance has helped me to complete the first part of this thesis successfully.

Working at CVIT was fun. I am very fortunate to learn from Thrupthi Ann John during my master's time. She is a wonderful friend and motivator. I am very thankful to her for all the guidance, discussions, and motivation. I am grateful to my seniors Praveen, Pritish, Rajvi, Aniket, Aditya, and Gaurav for their help and support. I would like to extend my special thanks to my friends Manisha and Dhaivat for believing in my potential and always pushing me to high in my work. I am thankful to my friends Tejaswi, Shweta, Bhavani, Ruchi, Vandana, and Riya. Special thanks to Anirudha and Aaron for providing me the tool for data annotation.

I am also grateful to Siva, Silar, Mahendar, and all the annotators for helping me to create the two datasets during my thesis. These are the people without whom this journey could not have a beautiful closure. I am thankful to all my dear friends and colleagues who played a significant role in my MS. Most importantly, I would like to thank my family for their continual support, especially my sister Radhika for being my constant support. I am profoundly grateful to Mati and bade papa for incorporating the learning spirit in me.

Abstract

How can we assess the quality of human driving using AI? Driver inattention is one of the leading causes of vehicle crashes and incidents worldwide. Driver inattention includes driver fatigue leading to drowsiness and driver distraction, say due to the use of cellphone or rubbernecking, all of which leads to a lack of situational awareness. Hitherto, techniques presented to monitor driver attention evaluated factors such as fatigue and distraction independently. However, to develop a robust driver attention monitoring system, all the factors affecting a driver's attention needs to be analyzed holistically. In this thesis, we present two novel approaches for driver attention analysis on the road using driver video and fusion of driver and road video.

In the first approach, we propose the driver attention rating system that leverages the front camera of a windshield-mounted smartphone to monitor the driver attention by combining several features. We derive a driver attention rating by fusing spatio-temporal features based on the driver state and behavior such as head pose, eye gaze, eye closure, yawns, use of cellphones, etc. We present a few architectures for feature aggregation like AutoRate and Attention-based AutoRate. We perform an extensive evaluation of feature aggregation networks on real-world driving data and also data from controlled, static vehicle settings with 30 drivers in a large city. We compare the proposed method's automatically-generated rating with the scores given by 5 human annotators. We introduce the kappa coefficient, an evaluation metric to compute the inter-rater agreement between the generated rating and the rating provided by human annotators. We observe that Attention-based AutoRate outperforms other proposed designs for feature aggregation by 10%. Further, we use the learned temporal and spatial attention to visualize the key frame and the key action, which justifies the model's predicted rating. Finally, to provide driver-specific results, we fine-tune the Attention-based AutoRate model using the specific driver data to give personalized driver experience.

In the second approach, we propose driver gaze mapping on the road using the fusion of driver and road videos as input. The proposed approach is used to estimate driver attention and determine which objects the driver is focusing on while driving. To solve such a task, we introduce a new dataset called DGAZE, which is an image dataset that contains the driver view and road view annotated with the driver gaze point on the road. The data is collected in a lab setting, mimicking road conditions using low-cost mobile phone cameras. It has a total of 100,000 images collected with 20 drivers and 103 unique objects on the road belonging to 7 classes, including cars, pedestrians, traffic signals, auto-rickshaw, etc. We also present I-DGAZE, a fused convolutional neural network trained on the DGAZE dataset for

predicting driver gaze on the road. Our architecture combines facial features such as facial key-point location and head pose along with the image of the left eye to get optimum results. Our model achieves an error of 94.5 pixels without calibration and 45 pixels with calibration. We compare our model with state-of-the-art eye gaze works and present extensive ablation results.

Overall, in this thesis, we propose two methods for driver attention analysis on the road. These approaches provide feedback about the quality of driver attention using driver video and fusion of driver and road video. We introduced dataset for driver attention rating and driver gaze mapping on the road. We also introduced two novel architectures, Attention-based AutoRate and I-DGAZE, corresponding to each proposed task. The evaluation metric and experimental results prove the efficacy of the same. Our two significant contributions include the proposal of a rating system for measuring driver inattention on the road and the dataset consisting of both driver and road view along with driver gaze location on the road.

Contents

Chapter Page

1 Introduction . . . 1
1.1 Related Work . . . 2
1.2 Scope . . . 5
1.2.1 Problem Definition . . . 5
1.2.2 Contributions . . . 6
1.3 Thesis Outline . . . 8

2 Facial Features for Driver Attention Analysis . . . 9
2.1 Feature identification and extraction . . . 10
2.1.1 Generic features . . . 11
2.1.2 Specific facial features . . . 11
2.2 Feature Aggregation . . . 15
2.2.1 Handpicked + SVM . . . 16
2.2.2 Handpicked + LSTM . . . 16
2.2.3 CNN (generic features) and GRU or (CNN + GRU) . . . 16
2.2.4 AutoRate Architecture . . . 17
2.2.5 Attention Based AutoRate Architecture . . . 18
2.3 Summary . . . 20

3 Evaluation and Visualization of Driver Inattention Rating . . . 21
3.1 Dataset . . . 22
3.1.1 Data Collection Setup . . . 23
3.1.2 Data Annotation Tool . . . 24
3.1.3 Driver Attention Rating . . . 25
3.1.4 Data Annotation . . . 26
3.1.5 Dataset type and its distribution . . . 26
3.1.5.1 Driving dataset . . . 27
3.1.5.2 Static dataset . . . 28
3.1.5.3 Merged dataset . . . 28
3.1.6 Dataset Summary . . . 28
3.2 Evaluation Metrics . . . 29
3.2.1 F1 Score . . . 29
3.2.2 Kappa coefficient (κ) . . . 29
3.3 Experiments and Results . . . 30
3.3.1 Implementation Details . . . 30

3.3.2 Qualitative Results . . . 32
3.3.3 Quantitative Results . . . 32
3.3.3.1 Ground Truth Rating . . . 33
3.3.3.2 Mode Based Evaluation . . . 35
3.3.3.3 Agreement Based Evaluation . . . 35
3.3.4 Ablation Study . . . 36
3.3.5 Turing Test . . . 38
3.4 Visualization . . . 39
3.4.1 Visualization Mechanism . . . 39
3.4.2 Visualization Results . . . 40
3.4.3 More Visualization Results . . . 41
3.5 Personalization . . . 42
3.6 Summary . . . 43

4 Driver Gaze Mapping on Road . . . 47
4.1 DGAZE: Driver Gaze Mapping Dataset . . . 49
4.1.1 Dataset Collection . . . 49
4.1.1.1 Dataset collection setup . . . 49
4.1.1.2 Object Annotation . . . 50
4.1.1.3 Dataset collection . . . 50
4.1.2 Dataset Statistics . . . 51
4.2 I-DGAZE: Gaze Prediction Architecture . . . 53
4.2.1 Facial features . . . 53
4.2.1.1 Face Area . . . 54
4.2.1.2 Face Location . . . 54
4.2.1.3 Head Pose . . . 54
4.2.2 I-DGAZE Architectire . . . 55
4.3 Experimental Evaluation and Results . . . 56
4.3.1 Evaluation Metrics . . . 56
4.3.2 Qualitative and Quantitative Results . . . 56
4.4 Summary . . . 58

5 Conclusions and Future Directions . . . 60
5.0.1 Conclusions . . . 60
5.0.2 Future Directions . . . 61

Bibliography . . . 64

List of Figures

Figure		Page
1.1	We focus on driver attention analysis on road using videos captured from the smart phone camera. We use driver video for driver attention rating and we use fusion of driver and road video for driver gaze mapping on road.	1
1.2	Eye gaze tracker to collect gaze data in real and simulation setting.	4
2.1	Rating system to predict driver attention based on specific and generic facial features. The figure on top shows the videos captured and annotated ratings. The figure at the bottom shows the use of AutoRate to predict driver inattention over a long video.	9
2.2	Pretrained VGG16 architecture for generic facial feature extraction.	11
2.3	Specific facial features extracted using corresponding pretrained state of the art networks	12
2.4	Facial Key points: (a) Facial Landmark[49](b) EAR: Eye Aspect Ratio (c) MAR: Mouth Aspect Ratio	13
2.5	Head pose variation for a video sample. The green region shows the range of angle for which the driver is looking on road and red region is for range of angle for which driver is looking off road.	14
2.6	Phone Detection using YOLO model finetuned for the corresponding task. The green color box is for ground truth and red box is for prediction.	15
2.7	Design choices.(a) Handpicked + SVM (b) Handpicked + LSTM (c) CNN + LSTM	17
2.8	AutoRate Architecture	17
2.9	Attention Based AutoRate Model	19
3.1	Data Collection Setup with a mobile phone camera mounted on the windshield of the car with a focus on the driver inside the vehicle. The mobile phone camera is enabled to collect both driver and road view simultaneously.	23
3.2	Data annotation tool to accelerate the annotation of ground truth rating for the collected video samples. The key features in the data annotation tool include the display of several videos per page, the number of pages can grow dynamically, the user can upload videos manually, watch the video in full screen and annotate videos by answering few related questions.	24
3.3	Description of driver attention rating from rating-1(very careless) to rating-5(very attentive). These descriptions give a soft understanding of the rating concept and do not bind them to any hard defined rules.	25

3.4 Data annotation: Five human annotators annotate each video sample. The annotations are highly subjective and annotator specific. Inter-rater agreement, like the kappa coefficient of threshold 0.80, is used to select the samples with good annotation. Note that the table below shows the impact of kappa value from None to almost perfect(k¿0.90). 27

3.5 Qualitative Results from the Attention-based AutoRate architecture 33

3.6 F1 score for four methods for all three dataset(driving, static, merged) and static2driving 34

3.7 Confusion matrix obtained for Attention Based AutoRate for (a) driving, (b) static and (c) merged datasets. For each ground truth rating, the row of numbers represents the percentage of predicted ratings from 1 to 5. The higher the percentage the darker the shade of the cell. 35

3.8 Agreement between AutoRate ratings and human annotators across datasets. 36

3.9 Turing test on misclassified samples. 38

3.10 Visualization Mechanism: Interpretability of Attention based AutoRate model for temporal and spatial visualization of videos. 40

3.11 Visualization results from Attention Based AutoRate. Darker the shade in the cell of attention map, higher is the impact of the feature in predicting driver attention rating. . 44

3.12 Driver with attention rating-4. In above figure, we observe that attention probability corresponding to generic features is high for 20% of the total video length. This means driver is attentive in 80% of the video and hence the predicted rating-4 which is equivalent to ground truth rating. Specific facial feature block corresponding to this inattentive region of video shows high attention probability for Eye Gaze X, Eye Gaze Y and Head Pitch which further concludes that driver is not looking straight on road for this section of the video. This further justifies driver attention rating as 4. We also see high probability value for seat-belt, which defines him a good driver. 45

3.13 Driver with attention rating-3 and predicted rating-3. In above figure, we observe that attention probability corresponding to generic features is high for more than 40% of the total video length. This means driver is attentive for less than 60% of the video and hence the predicted rating-3 which is equivalent to ground truth rating. Specific facial feature block corresponding to this inattentive region of video show high attention probability for Eye Gaze X, Eye Gaze Y, Face Area and Head Roll which concludes that driver is not looking straight on road for this section of the video. This further justifies driver attention rating as 3. 45

3.14 Driver with attention rating-4 and predicted rating-3. In above figure, we observe that attention probability corresponding to generic features is very high for 10% of the total video and decently high for 60% of the total video length. This means driver is properly attentive for only 30% of the video. Specific facial feature block corresponding to this inattentive region of video show high attention probability for Face Area, Eye Gaze X, and Eye Gaze Y but fails to predict Head Roll, Head Pitch and Head Yaw value which is important as video input shows a lot of variation in head pose. This means failure in specific facial feature effects final attention rating. 46

3.15 Driver with attention rating-3 and predicted rating-2. In above figure, we observe that attention probability corresponding to generic features is high for 80% of the total video length. This means driver is attentive for 20% of the video and hence the predicted rating-2 which is not equivalent to ground truth rating-3. Specific facial feature block corresponding to this inattentive region of video show high attention probability for Head Yaw, Eye Gaze X, and Eye Gaze Y. Visually the generic and specific features looks aligned with the video input but the network still fails to predict correct rating. This happens because the input video sample is ambiguous and it is difficult for even human annotators to rate it as 2 or 3. 46

4.1 In this work, we develop DGAZE the driver gaze mapping dataset that includes both driver and road view to capture driver gaze on road using low cost mobile phone cameras. Using DGAZE, we train I-DGAZE for prediction of gaze point on road. 47
4.2 DGAZE Collection Setup: Dataset is collected in lab setting which has close proximity with actual driving setting. Mobile phone camera attached to tripod stand similar to mobile phone camera mounted on wind shield of the car collects both driver view and projected road view at same frame per seconds. 50
4.3 Samples from DGAZE dataset corresponding to seven unique objects annotated on road. Note the significant variation in the size of the object, distance of the object from the driver and illuminance variation on road. These variations help the I-DGAZE model to train well on the dataset. 51
4.4 Number of annotated objects in views corresponding to each unique object on road . . 52
4.5 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. 53
4.6 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 other branch as facial features like head pose, face location and distance of driver face from mobile phone camera. 55
4.7 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. 57

List of Tables

Table		Page
3.1	Dataset description with train and test split.	28
3.2	Dataset Summary	29
3.3	Detailed description of the feature extraction.	32
3.4	Agreement between Attention Based AutoRate and Majority/Average ratings using kappa coefficient.	34
3.5	Comparison of attention based AutoRate model with AutoRate[20] model, CNN + GRU and other feature combinations. Note: κ denotes kappa coefficient used for inter rater agreement. The abbreviations used in the table stand for Head Pose (HP), Eye Gaze (EG) and Eye Blink (EB).	37
3.6	Performance of Attention Based AutoRate model as a result of stripping input features one by one.	37
3.7	Evaluation of personalization on 6 random drivers from dataset.	42
4.1	Comparison of DGAZE with other eye gaze mapping dataset	54
4.2	Comparison of I-DGAZE with existing gaze prediction methods	58
4.3	Ablation study of I-DGAZE model. Here, LEye = Left Eye, HP = Headpose, Y = Yaw, P = Pitch, R = Roll, FL = Face Landmark Location and FA = Face Area	59

Chapter 1

Introduction

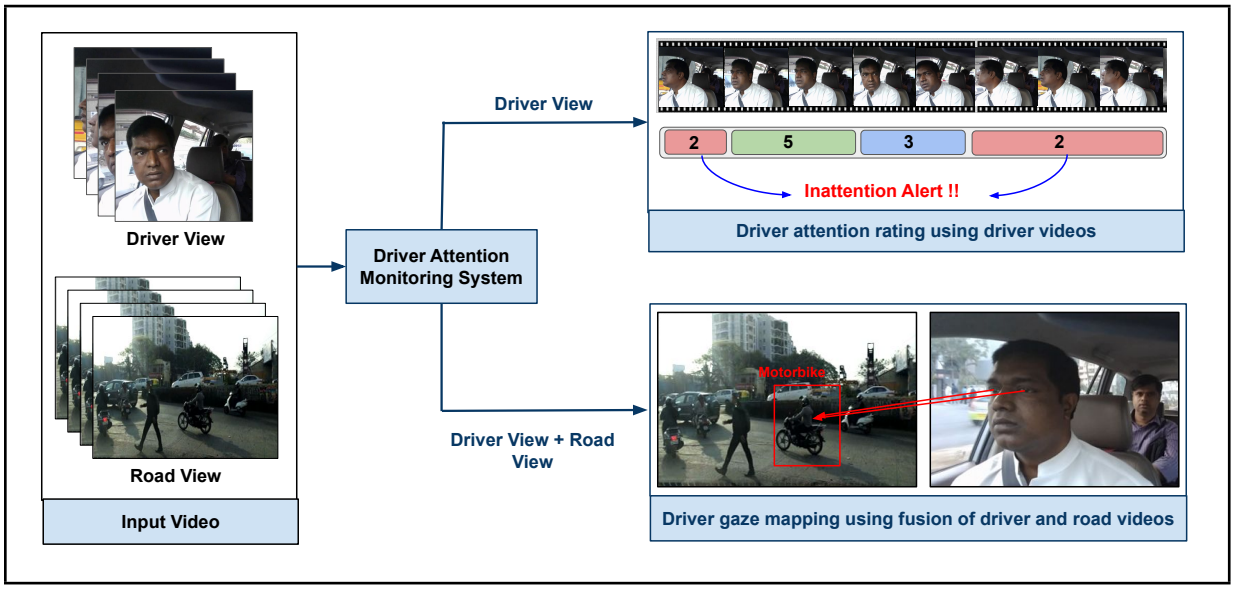


Figure 1.1: We focus on driver attention analysis on road using videos captured from the smart phone camera. We use driver video for driver attention rating and we use fusion of driver and road video for driver gaze mapping on road.

Driver inattention is one of the leading causes of road accidents in the world. According to the 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[18]. The number of annual deaths due to road accidents has reached 1.35 million[1]. Advanced driver assistance systems (ADAS) can contribute to a possible solution. ADAS are systems developed to automate, adapt, and enhance vehicle systems for safety and better driving. Since 2000, the automotive industry has gradually introduced new ADAS features in vehicles. The progress in this field has been enormous, and it has been demonstrated that the most advanced ADAS systems can contribute to reducing fatal road accidents by a factor of more than 30%. However, the major problem preventing ADAS massive

implementation is that these systems are found only in the high price vehicle segment and are not flexible to adapt new contributions made to the ADAS system.

Driver inattention occurs when the drivers divert their attention from the driving task to focus on other activities. The various factors contributing to driver inattention are fatigue, drowsiness, distraction, including talking on the phone or with other passengers, looking off the road, etc. Driver attention monitoring aims to analyze the driver's state and behavior to determine whether the driver is attentive(looking on-road). In general, a driver is considered to be attentive when (s)he concentrates on the road ahead for the majority of the time during the drive but also scans the mirrors regularly to maintain adequate situational awareness.

Traditionally, the factors affecting driver inattention have been evaluated independently. For instance, some high-end cars like Honda CR-V and Accord [2, 4] regularly monitor steering wheel input and raise alerts when the driver is frequently veering out of the lane. However, these solutions are expensive and are not present in all the vehicles. Hence, several camera-based ADAS systems have been designed. For instance, [72, 7] proposes smartphone-based drowsiness detection by analyzing features such as eye closure and yawn frequency. In [73, 66] various algorithms have been proposed to detect driver's gaze information to assess driver distraction like eyes off the road.

Thus far, most of the techniques proposed [72] have focused on monitoring the factors that affect driver's attention in individual silos. However, when humans (e.g., a supervisor or a passenger) assess a driver, they consider all of these factors in combination. Therefore, to make an effective assessment and to promote safe driving, we need to develop a comprehensive driver attention monitoring system that monitors and analyzes all the factors affecting the driver's attentiveness. Such a system could be used to provide a quantitative rating of driver attention. In this thesis, we focus on driver attention analysis on road using videos captured from the smart phone camera. We use driver video for driver attention rating and we use fusion of driver and road video for driver gaze mapping on road.

1.1 Related Work

We give an overview of few previous techniques to estimate driver attention on the road.

Sensor-based techniques: Lee et al. [35] propose a driver safety monitoring system that gathers data from different sensors such as cameras, electrocardiography, blood volume change sensor, temperature sensor, and a three-axis accelerometer, and identifies if the driver is driving safely or not. A kinect based system was developed in [14], where the driver attention was monitored using color and depth maps obtained from the kinect. The system analyzed eye gaze, arm position, head orientation and facial expressions to detect if the driver is making a phone call, drinking, sending an SMS, looking at an object inside the vehicle (either a map or adjusting the radio), or driving normally. In [75] and [9], head tracking sensors and 3D range cameras were used to monitor driver's head pose and driver distraction. The above techniques require installation of additional physiological sensors into the vehicle, which is intrusive and

cumbersome to maintain. In contrast, driver attention rating uses just a windshield-mounted smartphone to monitor driver's attention.

Camera-based techniques: Several camera-based ADAS systems have been proposed to determine driver distraction and fatigue [72, 7]. Dong et al. [19] present a review of various state-of-the-art techniques proposed to detect driver drowsiness, fatigue and distraction. Rezaei et al. [48] present an ADAS system that correlates the driver's head pose information to road hazards by analyzing two camera views simultaneously. The system combines the head pose information with distance to the vehicle in front to reason rear-end collisions. A technique to detect driver drowsiness based on eye blinking pattern was proposed in [29]. These approaches can only monitor specific aspects of driver's attention, however, to have a robust driver attention monitoring system all the factors affecting the driver's attention needs to be monitored holistically. Vicente et al. [64] propose a system to detect eyes off the road. The system uses head pose information to detect where the driver is looking. In real driving scenarios, head pose information alone may not be sufficient to accurately determine where the driver is looking as the driver can perform a quick scan by rolling the eyes. Song et al. [56] describe a system to detect talking over the phone using the microphone's audio data and driver's voice features. Sheshadri et al. [51] detect driver cell phone usage by analyzing the face view videos. The authors develop a custom classifier to detect if the phone is present or not in an image. In contrast, our driver attention rating system takes a holistic approach to identify and monitor all the factors that affect driver attention monitoring such as fatigue, drowsiness and distraction using a windshield-mounted smartphone.

Our work on driver attention rating goes beyond existing works [46] to derive a driver attention on road, which can be used by insurance companies to determine the premium, or to provide effective feedback to the drivers. We show that deriving a robust driver attention rating is non-trivial due to the ambiguity in rating driver's attention. To this end, we propose a deep learning system that combines generic and specific facial features towards deriving a driver attention rating. We show the efficacy of AutoRate on a real-world dataset comprising of 30 drivers in a large city.

Driver Gaze Data: A.Palazzi et al. [42] propose the driver gaze data collection in a real driving setup using the eye gaze tracker. The SMI Eye Tracking Glass is used to collect data for this task. The tracker, along with the rooftop camera, is aligned to obtain the road video and driver gaze on the road. The eye tracker used for this data collection is very costly and not affordable for all vehicle users. The dataset consist of 500,000 registered frames. Eye Tracking for Everyone [32], collects the data using mobile phone applications via crowdsourcing, enabling the collection of large scale datasets. The paper "Eye Tracking for Everyone" focuses on predicting the user's gaze on mobile phones and tablets. They introduce GazeCapture, the first large-scale dataset for eye tracking, containing data from over 1450 people consisting of almost 2.5M frames. Other accessible eye gaze dataset are [37, 67, 55, 38, 58, 74, 27, 32]. The downside of these datasets is that they do not contain significant variation in the head pose or have a coarse gaze point sampling density. The above techniques require costly eye trackers and roof top cameras to capture driver gaze on road in real driving setting. Figure 1.2 presents few hardware devices for driver gaze collection in real and simulation setting. As data collection in real

Figure 1.2: Eye gaze tracker to collect gaze data in real and simulation setting.

setting environment is risky, many researchers collects the data in simulation. We also collect the data in simulation for both driver and road video using mobile phone camera mounted on tripod stand.

Driver Gaze on Road: Driver Gaze Region Estimation Without Using Eye Movement[21] predicts the gaze in separate regions using the driver face videos. The region are possibly left view, right view or centre which is coarse data annotation. A.Palazzi et al. [42] uses a computer vision model based on a multi-branch deep architecture that integrates three sources of information: raw video, motion and scene semantics to predict driver attention on road. Eye Tracking for Everyone [32]proposes itracker, network for predicting the driver gaze on the road. The network uses images captured from the front camera of the mobile phone. The model achieves a prediction error of 1.7cm and 2.5cm without calibration on mobile phones and tablets respectively. With calibration, this is reduced to 1.3cm and 2.1cm. On the contrary, our input image consists of a driver view from the mobile phone camera. We use a late fusion convolution neural network with a specific facial feature as input to one branch of the network. Our output is projected on wall in front and we get error of 94.5 pixels without calibration and 45 pixels with calibration. There exist other appearance-based models that use a geometric model of an eye and can be subdivided into corneal-reflection-based[71, 76, 77]and shape-based methods[10, 61, 24]. There exist another popular technique[74] that directly uses eyes as input and can potentially work on low-

resolution images, but this requires a large amount of user-specific training. Our model generalizes well to the training data collected over 20 users. The model generates better results after fine-tuning with a small amount of user-specific data.

Other Specific Indian Efforts for Driver Safety: Harnessing AutoMobiles for Safety, or HAMS, project [40] by Microsoft Research is an initiative to monitor the state of the driver and how the vehicle is being driven in the context of a road environment that the vehicle is in. They use low-cost sensing devices to construct a virtual harness for vehicles. According to the ACM article on "Technology Interventions for Road Safety and Beyond[28]", the general goal is to have affordable technologies that work with humans through effective monitoring and feedback, rather than replacing humans through full autonomy. Driver monitoring will reduce the number of accidents and provide safe driving environment. Other efforts that contribute driver safety includes IDD dataset[63] by IIIT Hyderabad and Intel. The dataset contributes to benchmark computer vision techniques on the unstructured Indian road conditions. It is also helping spur the development of new techniques for such data collection, such as low-cost inspection of road infrastructure (potholes, signage, and street lights) using computer vision and inertial sensing. There are many such efforts from the leading companies like BOSCH, NISSAN, INTEL towards creating safe driving environment.

1.2 Scope

1.2.1 Problem Definition

In this thesis, we focus on driver attention monitoring using the driver video and fusion of driver and road videos, as shown in Figure 1.1. Driver inattention occurs when the driver has diverted away from the task of safe driving. We use the mobile phone camera mounted on the windshield of the car to collect both driver and road videos simultaneously. Traditionally, researchers used sensors like Kinect devices, head tracking sensors, 3D range cameras, etc. to identify if the driver is driving safely or not. All these sensors are specific for the proposed task, which means 'n' task requires 'n' sensors. Deploying a solution for a new job related to driver monitoring becomes costly and non-adaptive. Instead of many such sensors, we use only mobile phone cameras to collect both driver and road view simultaneously. We extract and aggregate various facial features to estimate the quality of driver attention on the road. This provides flexibility to add features corresponding to the new solved task with minimum training. Note that driver attention prediction is based on driver state and behavior and not driving. The collected driver video can predict driver gaze on the road as left view, center view, and right view, which is coarse prediction. To this end, we combine both driver view and road view to predict fine-grained driver gaze location on the road. The eye gaze prediction on the road can be used to analyze the change in driver gaze with a change in distance of object on the road, it can aid in avoiding collision with pedestrian crossing road, object driver attends on the road, etc. We describe in detail about the driver attention monitoring using only driver videos and fusion of driver and road videos.

Driver attention monitoring using driver videos: This work aims to monitor driver attention while driving using the driver videos collected using mobile phone cameras mounted on the windshield of the car. Thus far, most of the techniques proposed [72] have focused on monitoring the factors that affect driver's attention in individual silos. However, when humans (e.g., a supervisor or a passenger) assess a driver, they consider all of these factors in combination. Therefore, to make an effective assessment and to promote safe driving, we need to develop a comprehensive driver attention monitoring system that monitors and analyzes all the factors affecting the driver's attentiveness. Such a system could be used to provide a quantitative rating of driver attention.

Driver attention monitoring using a fusion of driver and road videos: In this work, we aim to combine driver and road videos to predict fine-grained driver gaze on the road. The driver gaze can be used for the analysis of driver attention on the road, length of the driver gaze on a specific object, the order in which visual elements are fixated upon. These eye-gaze trackers can be used for avoiding various accidents due to driver inattention. But the major disadvantage of using these wearable eye-gaze trackers is that they range from a few thousand to few lakh rupees, which is not affordable. The other downside is that as these trackers are mounted on lightweight eyeglasses, we cannot obtain an unobstructed driver image. We can only get the road view. Thus, we aim to utilize the mobile phone videos of both driver and road view to obtaining driver gaze on the road.

1.2.2 Contributions

In first part of the thesis, we propose driver attention rating system that leverages the front camera of a windshield-mounted smartphone to monitor driver attention on road by combining several features. We derive a driver attention rating by fusing spatio-temporal features based on the driver state and behavior such as head pose, eye gaze, eye closure, yawns, use of cellphones, etc. We perform extensive evaluation of AutoRate(driver attention rating system) on real-world driving data and also data from controlled, static vehicle settings with 30 drivers in a large city. We compare AutoRate's automatically-generated rating with the scores given by 5 human annotators. We compute the agreement between AutoRate's rating and human annotator rating using the kappa coefficient.

Second, we use fusion of driver and road videos to estimate driver attention and determine which objects the driver is focusing on road while driving. Currently, there are no large scale datasets for driver gaze prediction on the road. Collection of such gaze data requires wearable eye gaze trackers, which are costly and do not provide an unobstructed view of the driver. We introduce a new dataset called DGAZE to tackle this problem. DGAZE is an image dataset for driver gaze mapping which contains the driver view and road view annotated with the driver gaze point. It is collected in a lab setting mimicking road conditions using low cost mobile phone cameras. The dataset has a total of 100,000 images collected with 20 drivers and 103 unique objects on road belonging to 7 classes including cars, pedestrian, traffic signal, auto rickshaw etc. 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 location and pose along with the image of the left eye to get optimum results.

This work proposes a framework for driver attention monitoring using driver videos and fusion of driver and road videos with following specific contributions:

Specific contributions in this thesis:

1. We propose a system to monitor driver inattention by capturing driver video through a windshield-mounted mobile phone camera and predicting a rating of 1 to 5 for 10 second segments of the video, where 1 implies least attentive and 5 implies most attentive. Our system employs spatial and temporal facial features extracted from the video using state-of-the art pre-trained models which are then combined to automatically rate driver attention.

2. We create a driver video dataset consisting of 3200 videos in static and driving setting that can be used for building a comprehensive driver inattention prediction system.

3. We introduce the kappa coefficient, an evaluation metric to compute the inter-rater agreement between the ratings provided by the proposed model and human annotators. This helps to take into consideration the subjectivity related to rating, which can not be captured using other evaluation metrics like accuracy, precision, recall, and F1-score.

4. We explain the ratings predicted by our model by visualizing the key frames and key actions that influence the rating. We do this using learned temporal and spatial attention.

5. We introduce the DGAZE dataset for predicting driver gaze on road. It consists of both driver videos and corresponding road videos collected using the mobile phone camera mounted on wind shield of the car.

6. We propose I-DGAZE, a model for predicting driver gaze on road using a mutli-branch fused convolutional neural network.The model is able to predict the gaze with an error of 94.5 pixels without calibration and 45 pixels with calibration.

7. We have made the dataset, network model and the source code publicly available.

Specific contributions to the driver attention monitoring community:

1. We propose a rating system for driver attention analysis in the range of 1 to 5, where rating-1 means very careless and rating-5 means very attentive. The rating system combines all the features responsible for driver inattention on-road rather than analyzing them independently.

2. We introduce the kappa coefficient, an evaluation metric to obtain an inter-rater agreement between the subjective ratings by different annotators.

3. We introduce the DGAZE dataset for driver gaze mapping on the road. The dataset is collected in a lab setting using a mobile phone camera. It consists of both driver and road view along with gaze point on the road. Eye gaze trackers and eyeglass trackers are very expensive, and hence, collection of such dataset is challenging.

1.3 Thesis Outline

The next three chapters (chapters 2, 3, and 4) provide a self-contained description of our contributions to the two proposed novel techniques for driver attention analysis on the road. A brief outline of the text in this thesis is as follows. In chapter 2, we explain feature extraction for determining driver attention on the road. We then propose different feature aggregation architectures like AutoRate, Attention-based AutoRate, etc to predict driver attention on the road.

In chapter 3, we explain dataset collection for driver attention rating task; we further introduce evaluation metrics like kappa coefficient and Turing test to evaluate subjective problems like driver attention rating on the road. We then present an extensive set of experimental results to compare different feature aggregation techniques. We use the learned temporal and spatial attention to visualize the key frame and the key action, which justifies the model's predicted rating. Further, we observe that personalization in driver attention rating can improve driver-specific results by a significant amount.

In chapter 4, we predict driver attention on-road using the fusion of both driver and road videos. We introduce DGAZE, an image dataset for driver gaze mapping, which contains the driver view and road view annotated with the driver gaze point. We also present I-DGAZE, a fused convolutional neural network for predicting driver gaze on the road, which was trained on the DGAZE dataset. We also compare the result from I-DGAZE with state of the art eye gaze prediction models.

Chapter 2

Facial Features for Driver Attention Analysis

Figure 2.1: Rating system to predict driver attention based on specific and generic facial features. The figure on top shows the videos captured and annotated ratings. The figure at the bottom shows the use of AutoRate to predict driver inattention over a long video.

In this chapter, we employ a camera-based system to determine the driver's attention rating automatically. We use the front camera of a windshield-mounted smartphone, which gives a 60-degree view of the scene centered on the driver. The driver attention rating system derives a rating (in the range of 1 to 5) using the visual features from the camera feed, which is equivalent to a rating provided by a human annotator looking at the driver's video. Note that our rating of driver attention is based on driver behavior, and not on their driving. An assessment of driving would likely need additional sensing

streams to detect sharp braking, jerks, honking, etc.. It would be quite challenging to do (and even more subjective) if attempted based just on the driver-facing video. We use human annotation instead of physiological sensors [6] to detect inattention as sensors are intrusive. The proposed system derives a rating by identifying and fusing Spatio-temporal features that affect the driver's attention. The rating system is trained and tested using an extensive real-world dataset comprising over 3200 unique video snippets, each of length 10 seconds, across 30 drivers in a large city (i.e., 160,000 total images when sampled at 5 fps). We used 5 human annotators to rate each 10-second video snippet on a 5-point scale to get ground truth driver attention rating.

Designing such a system to derive an accurate driver attention rating is challenging because: (i) Unlike typical image classification tasks, classifying a video snippet is more challenging as the system needs to identify and extract Spatio-temporal information across the sequence of frames to capture the dynamics of driver attention. (ii) Ratings provided by human annotators (even highly reputed ones) are subjective and therefore differ from person to person, as the task of rating is inherently ambiguous (e.g., the difference between adjacent levels of attention rating is not clear-cut). This results in ground truth not being precise, making it hard for the prediction task. (iii) To our knowledge, there exists no dataset with driver attention information in real-world driving scenarios that could be used to train the system comprehensively.

To address these challenges, in this work, we propose a camera-based system to determine the driver's attention rating automatically. We use the front camera of a windshield-mounted smartphone, which gives a 60° view of the scene centered on the driver. The objective of the rating system is to derive a driver's attention using the visual features from the camera feed, such that it is equivalent to a rating provided by a human annotator looking at the driver's video. We provide a detail description of the state of the art networks for identification and extraction of features like facial landmarks, head pose, and eye gaze. We have finetuned YOLO for phone and seatbelt detection for predicting the illegal driving. We present several feature combination techniques to predict driver attention on the road. More specifically, we explain the AutoRate architecture, which is a late fusion of CNN features from one branch and specific features like head pose, eye gaze, etc. from the other branch. The system worked okay for the samples, but it failed for videos in which specific features get undetected. To address this problem, we proposed Attention-based AutoRate, which learns to use the selective input features and hence works right irrespective of presence or absence of any feature. We have provided a detailed explanation of all this architecture in this chapter.

2.1 Feature identification and extraction

AutoRate extracts two types of features, (i) generic features and (ii) specific facial features.

2.1.1 Generic features

The idea of extracting generic features is to ensure high-level object patterns in the image is captured. To this end, we use the transfer learning approach outlined in Section 2.2.3, with a pre-trained VGG16 [54] convolutional network being used to extract a low-dimensional feature representation (or bottleneck features) of the frames as shown in Figure 2.2.

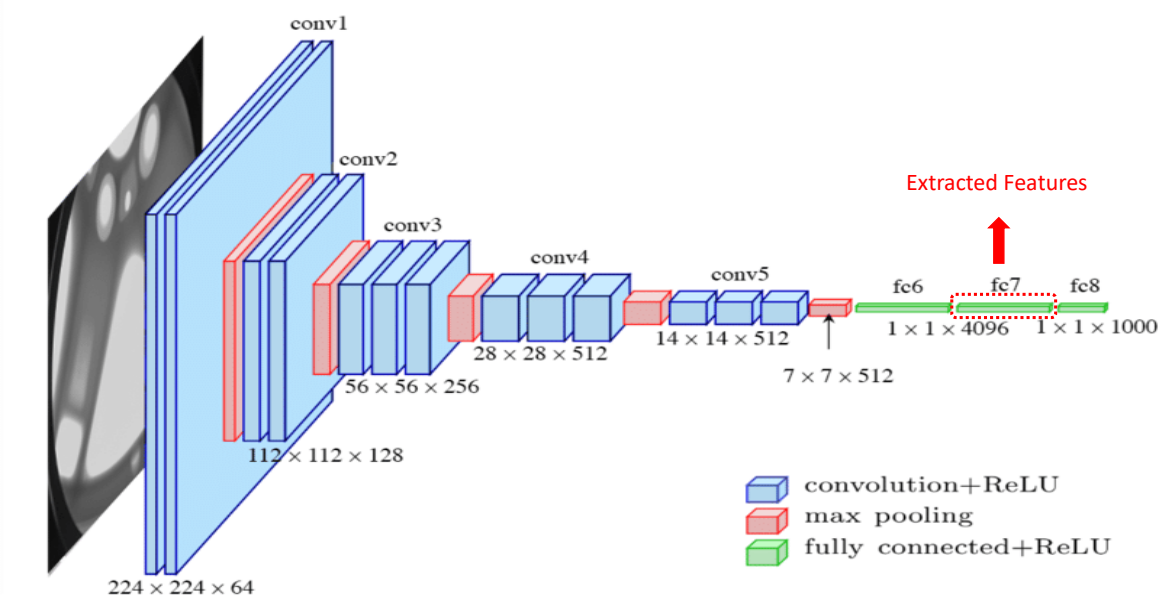


Figure 2.2: Pretrained VGG16 architecture for generic facial feature extraction.

2.1.2 Specific facial features

Generic features alone are not sufficient to adequately capture the dynamics entailed in driver attention monitoring. Therefore, AutoRate identifies a comprehensive set of features that are relevant to the rating task, *viz.,* facial landmarks, eye closure, yawns, head pose, eye gaze, talking over the phone, and face area. These features were identified after an extensive analysis of real-world driving videos and understanding driver behavior [19]. We use state-of-the-art pre-trained models to extract these specific facial features from a sequence of frames. Figure 2.3 shows the facial feature extraction block for each frame. We now discuss the key facial features and describe how these are extracted from an input image:

1. Facial landmarks:

Facial landmark detection is a fundamental component in AutoRate to extract features. It aims to localize facial feature points such as eye corners, mouth corners, nose tip, etc. AutoRate uses facial landmarks to detect eye closure, yawns, and eye gaze, which form the features of interest. Real-world

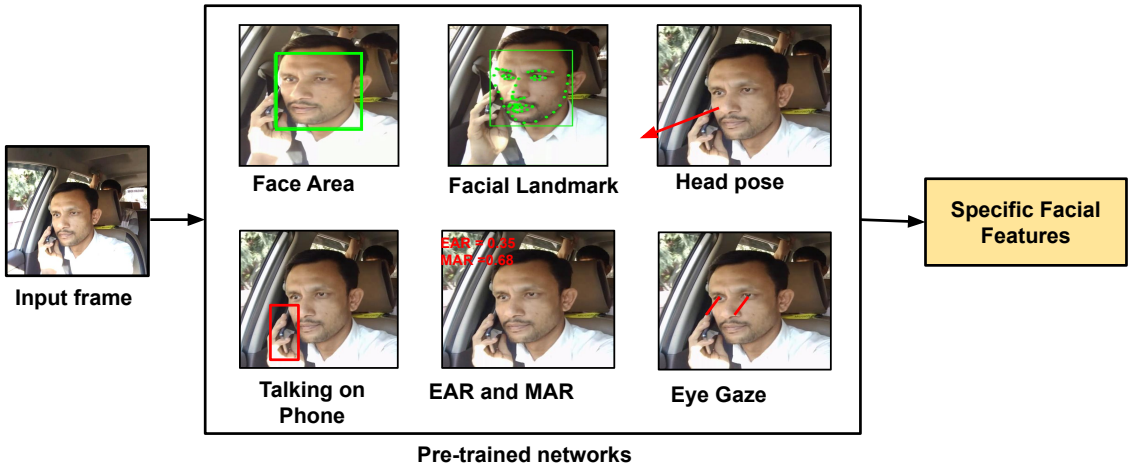


Figure 2.3: Specific facial features extracted using corresponding pretrained state of the art networks

conditions call for the facial landmark detection to handle (i) large head pose variation due to frequent mirror scanning or looking off the road, and (ii) diverse lighting conditions like sunny, shadows, etc. There exists several techniques from active appearance model to Convolutional Neural Networks (CNNs) to extract facial landmarks from an image [30, 5, 25]. Here, we employ a pre-trained Face Alignment Network(FAN)[8] to extract facial landmarks. The FAN network employed is based on the Hour-Glass (HG) network [69] that aims to learn relevant features at different scales in the image and outputs pixel-wise predictions. The FAN network is trained on multiple in-the-wild datasets to obtain robust landmarks in the face of large pose variations (e.g., yaw in the range $[-90^o, 90^o]$) and diverse lightning conditions. Further, the FAN network achieves state-of-the-art result across multiple datasets such as 300-W [49], 300-VW [12, 53, 60] and LS3D-W [8].

2. Eye closure & yawns:

Several studies have identified behavioral measures such as eye closure and yawn frequency to detect drowsiness [50]. AutoRate leverages facial landmarks to detect eye closure and yawns [40]. Specifically, to detect eye closure we use the eye aspect ratio (EAR) [57] metric, which is the ratio of the height of the eye to its width. EAR(Figure 2.4(b)) is defined as,

$$EAR = \frac{||p_{38} - p_{42}|| + ||p_{39} - p_{41}||}{2||p_{37} - p_{40}||}, \tag{2.1}$$

where p_{38}, ..., p_{42} are the eye landmarks. EAR is close to zero when the eye is closed and non-zero when it is open.

Similarly, to detect yawns we use the mouth aspect ratio (MAR) metric, which is the ratio of the height of the mouth to its width. MAR (Figure 2.4(b)) is defined as:

$$MAR = \frac{||p_{62} - p_{68}|| + ||p_{63} - p_{67}|| + ||p_{64} - p_{66}||}{3||p_{61} - p_{65}||}, \tag{2.2}$$

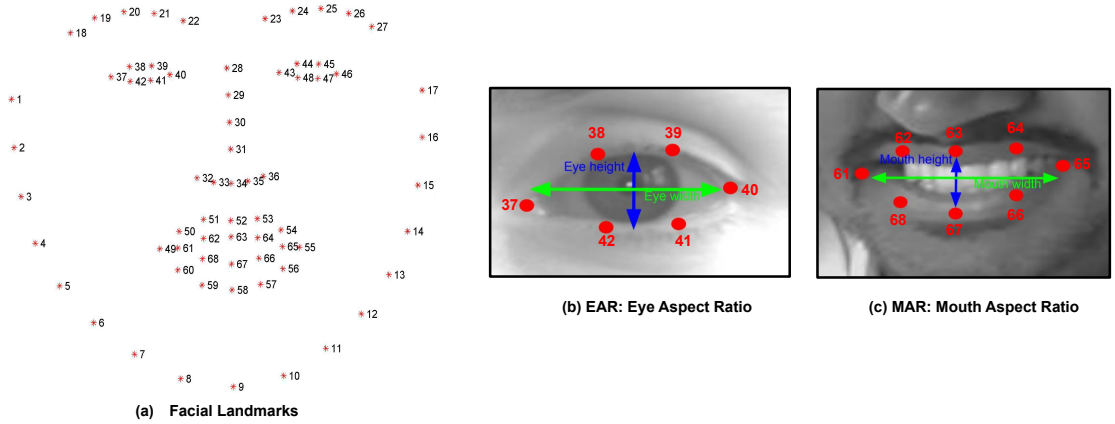


Figure 2.4: Facial Key points: (a) Facial Landmark[49](b) EAR: Eye Aspect Ratio (c) MAR: Mouth Aspect Ratio

where p_{61},...p_{68} are the landmarks corresponding to upper and lower lip. MAR is close to zero when the mouth is closed and a non-zero value when it is open. A yawn can be detected when the mouth is opened (i.e., when MAR crosses a threshold) continuously for a prolonged period. Unlike past work that has used EAR and MAR to detect eye closure and yawns as signs of drowsiness, AutoRate uses the raw EAR and MAR values as features towards driver attention rating.

3. Head pose:

Head pose information is a key feature for determining where the driver is looking and monitoring the driver's alertness. In a real driving scenario, the driver tends to scan her/his environment to maintain situational awareness, hence head pose detection should be robust to such variation. While head pose can be derived using traditional techniques such as PnP (Perspective-n-Point) algorithms [41], we employ a pre-trained CNN [33] due to its robustness. The pre-trained network *viz.,* Deepgaze [45] is trained using datasets such as Prima [22], AFLW [36], and AFW [31] to handle large pose variations. The input to the network is the driver's face region, which is obtained by cropping it along the landmarks in order to *de-noise the background.* The network then outputs the corresponding head pose information *viz.,* yaw, pitch and roll angles. Figure 2.5 shows head pose variation obtained using the above method for a video sample.

4. Eye gaze:

In a driving scenario, eye gaze is also an important cue to determine where the driver is looking in addition to head pose. Hence, eye gaze information is important to determine where the driver is looking [40]. We employ a standard LeNet-5 [34] network that takes an eye patch as the input and outputs

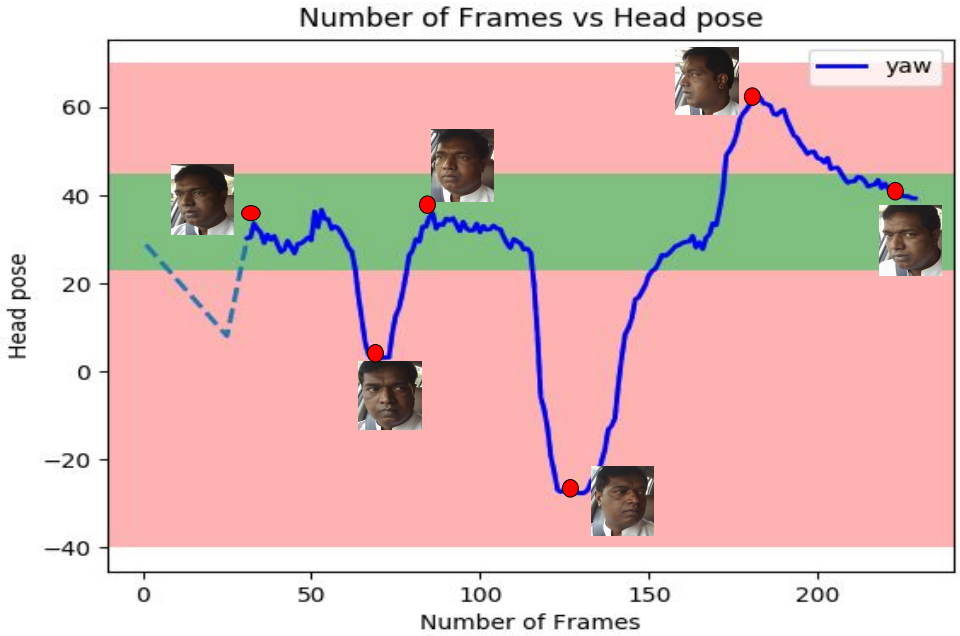


Figure 2.5: Head pose variation for a video sample. The green region shows the range of angle for which the driver is looking on road and red region is for range of angle for which driver is looking off road.

the gaze information, *viz.,* yaw and pitch values. The input eye patch is obtained by considering the landmarks associated with the eye region. We train the LeNet model using in-the-wild MPIIGaze [74] dataset, which contains 213,659 images from 15 participants.

5. Talking over the phone:

Talking over the phone while driving is a form of distracted driving. Identifying talking over the phone is a challenging task, as the phone object varies in type and size. We are not aware of any pre-trained network for phone detection, so we collected around 1200 sample images when the driver is talking on the phone (by holding it up to their face) and manually marked the bounding box around the phone. The labeled images with bounding box of the phone was used to train a custom object detector using CNNs. We use a pre-trained YOLOv2 [47] network trained on COCO dataset [70], where we freeze all but last few layers and fine tune the network with our dataset. The final predictions are then restricted to only detection of a phone and the corresponding bounding box in an image as shown in Figure 2.6

6. Face area: AutoRate uses face area as a feature to determine the change in driver's seating position, e.g., leaning forward or leaning back. To detect face area, we use a robust face detection algorithm *viz.,*

Figure 2.6: Phone Detection using YOLO model finetuned for the corresponding task. The green color box is for ground truth and red box is for prediction.

Tiny Faces [26] that can deal with extreme illumination, blurring, pose variation, and occlusion. The intuition is that if the face area is large, then the driver is sitting closer to the camera and vice versa.

To summarize, AutoRate identifies and extracts both generic and specific facial features to form a comprehensive feature vector (V_i) for each frame 'i', *viz.,*

$$V_i = [Bottleneck_f, EAR, MAR, yaw_h, pitch_h, roll_h, yaw_e, pitch_e, talk, area_f] \tag{2.3}$$

where, $Bottleneck_f$ represents the generic features from VGG16, *EAR* & *MAR* represents the raw eye and mouth aspect ratios. $yaw_h, pitch_h, roll_h$ represents the head pose information and $yaw_e, pitch_e$ represents the eye gaze information. $talk$ is a boolean value indicating if the driver is talking over a phone or not and $area_f$ indicates the driver's face area in an image.

2.2 Feature Aggregation

We now describe how to aggregate feature vectors (V_i) obtained for a sequence of frames in a video snippet. The objective of the aggregation function is to combine the feature vectors across frames (in our setup it is 50 frames) to capture both spatial and temporal information.

To this end, we present different approaches for determining the rating from the given video. IN one of the approach, we use pre-trained CNN (Convolutional Neural Network) to extract the generic features and then apply a GRU (Gated Recurrent Unit) across the frames to get a final representation of the entire video snippet. In other approach which we refer to as the AutoRate architecture, besides

having the features from a CNN, we also have other specific features which are then combined using GRU to get overall feature vector for the video. In the third approach which we refer to as Attention-based AutoRate, we use attention layer after applying LSTM (Long short-term memory) to both specific and facial features. The attention layer learns the attention probabilities corresponding to every frame of video.

2.2.1 Handpicked + SVM

Figure 2.7(a) shows the Handpicked + SVM architecture for determining driver attention rating. The key idea is that for each input frame we extract the *specific facial features*. The intuition here is that specific facial features guides the network to learn key actions performed by the driver. It takes a sequence of frames as input; we used a 10-second video snippet sampled at 5 frames per second (fps), resulting in 50 frames. The input frames, are fed to a series of state of the art pre-trained networks corresponding to each task to extract relevant features. The facial features obtained are

2.2.2 Handpicked + LSTM

Figure 2.7(b) shows the Handpicked + LSTM architecture for determining driver attention rating. The key idea is that for each input frame we extract the *specific facial features* like face area, head pose, eye gaze, MAR, EAR and phone detection. The intuition here is that specific facial features guides the network to learn key actions performed by the driver. It takes a sequence of frames as input; we used a 10-second video snippet sampled at 5 frames per second (fps), resulting in 50 frames. The input frames, are fed to a series of state of the art pre-trained networks corresponding to each task to extract relevant features. The facial features obtained are fed into the sequential model, i.e., a series of GRU (gated recurrent unit) [11] blocks to extract spatiotemporal information. The features from the final layers of the GRU models are then fed to classifier to obtain the rating in the range of 1 to 5.

2.2.3 CNN (generic features) and GRU or (CNN + GRU)

As recent works [43] have shown that deep neural networks (DNNs) trained for one task capture relationships in the data that can be reused for different problems in the same domain. The pre-trained models have a strong ability to generalize to images outside the training dataset. This has led to transfer learning, where the idea is to use pre-trained models such as VGG16 [54] trained on the ImageNet [16]

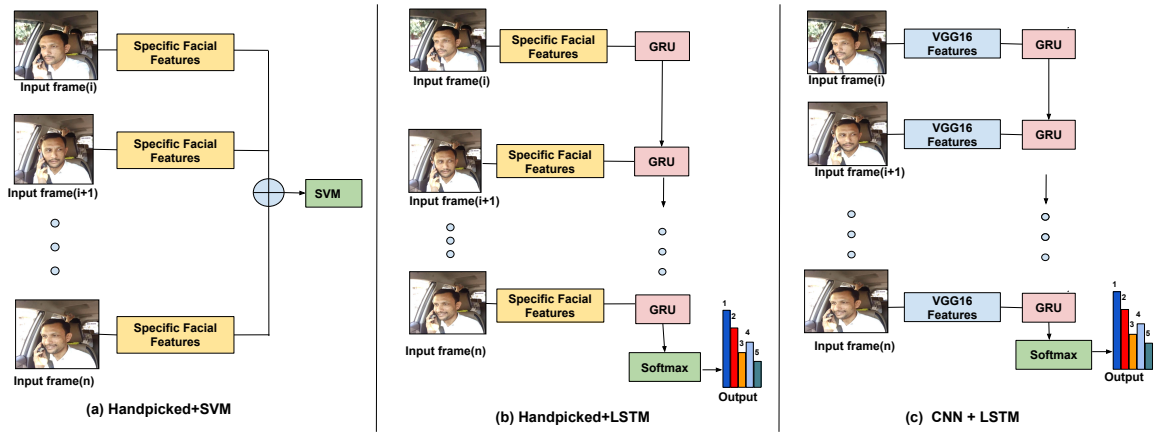


Figure 2.7: Design choices.(a) Handpicked + SVM (b) Handpicked + LSTM (c) CNN + LSTM

dataset, to extract bottleneck features. Figure 2.7(c) shows the architecture of such an approach. These features are then used to extract the temporal information. In detail, the input to the network is a sequence of frames from a 10-second video snippet. Each image is fed to a pre-trained VGG16 network that extracts bottleneck features at the first fully connected layer. These features are then aggregated using GRU to predict driver attention rating.

2.2.4 AutoRate Architecture

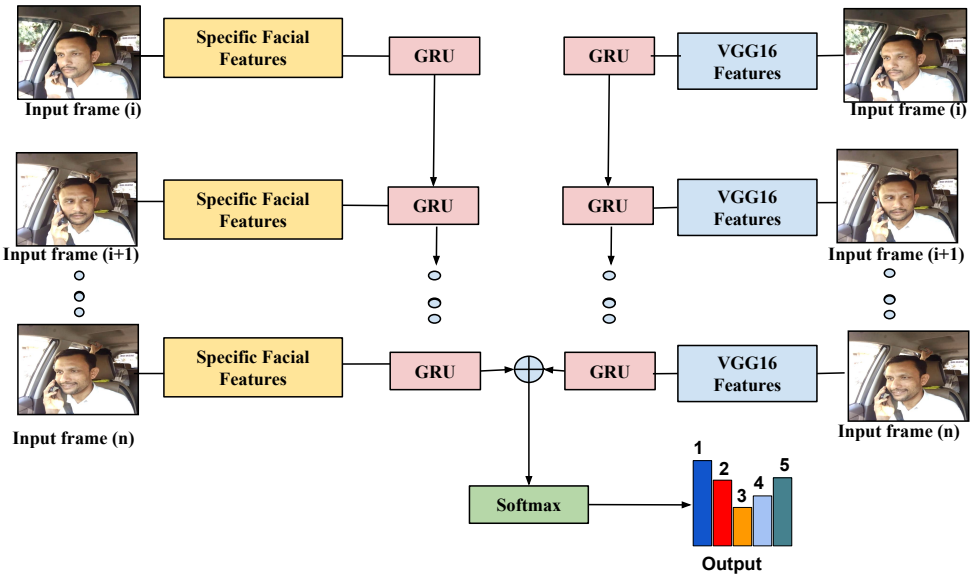


Figure 2.8: AutoRate Architecture

Figure 2.8 shows the proposed architecture of AutoRate for determining driver attention rating. The key idea is that for each input frame we extract both the *generic features* and *specific facial features*. The intuition here is that generic features capture high-level patterns in a frame and specific facial features guides the network to learn key actions performed by the driver, which may not be captured by the previous approach that uses only generic features. AutoRate takes a sequence of frames as input; we used a 10-second video snippet sampled at 5 frames per second (fps), resulting in 50 frames. The input frames, along with ground truth ratings, are fed to a series of pre-trained networks to extract relevant features. The facial and generic features obtained are separately fed into two different sequential models, i.e., a series of GRU (gated recurrent unit) [11] blocks to extract spatiotemporal information. The features from the final layers of both the GRU models are then concatenated to obtain the overall representation of the video.

2.2.5 Attention Based AutoRate Architecture

For challenging real world videos, we observe that AutoRate sometimes fails to predict the correct rating. We define videos as challenging where state-of-the-art methods for identifying facial features like face area, head pose, eye gaze, etc. perform poorly. To address this, we propose an approach called attention based AutoRate which uses a technique similar to the technique used by humans. Humans rate driver videos by using selective attention to tune out irrelevant information and concentrate on what really matters. In attention based AutoRate architecture, we achieve selective attention by introducing an attention module in both the generic feature branch and specific facial feature branch of AutoRate architecture. As shown in Figure 2.9, we introduce the attention module after applying LSTM to generic features (4096 dimensional vector per frame) and specific features. This attention module is the weighted combination of attention probabilities (α_1 ... α_n) as shown in Equation (2.4).

$$X = \sum_{t=1}^{n} \alpha_{k,t} h_t \tag{2.4}$$

where, k is S or V. S stands for specific features in right branch of Figure 2.9, V stands for VGGFace features in left branch of Figure 2.9, t specifies the frame number that varies from 1 to 50 and h_t specifies the hidden unit from LSTM block at time t. Note that we have used VGGFace features as input instead of VGG features and Bi-directional GRU instead of GRU for better results.

Few other additions made to attention based AutoRate model are: (1) We use VGGFace [44] features as generic feature instead of VGG16 features used in AutoRate [20] architecture because VGGFace

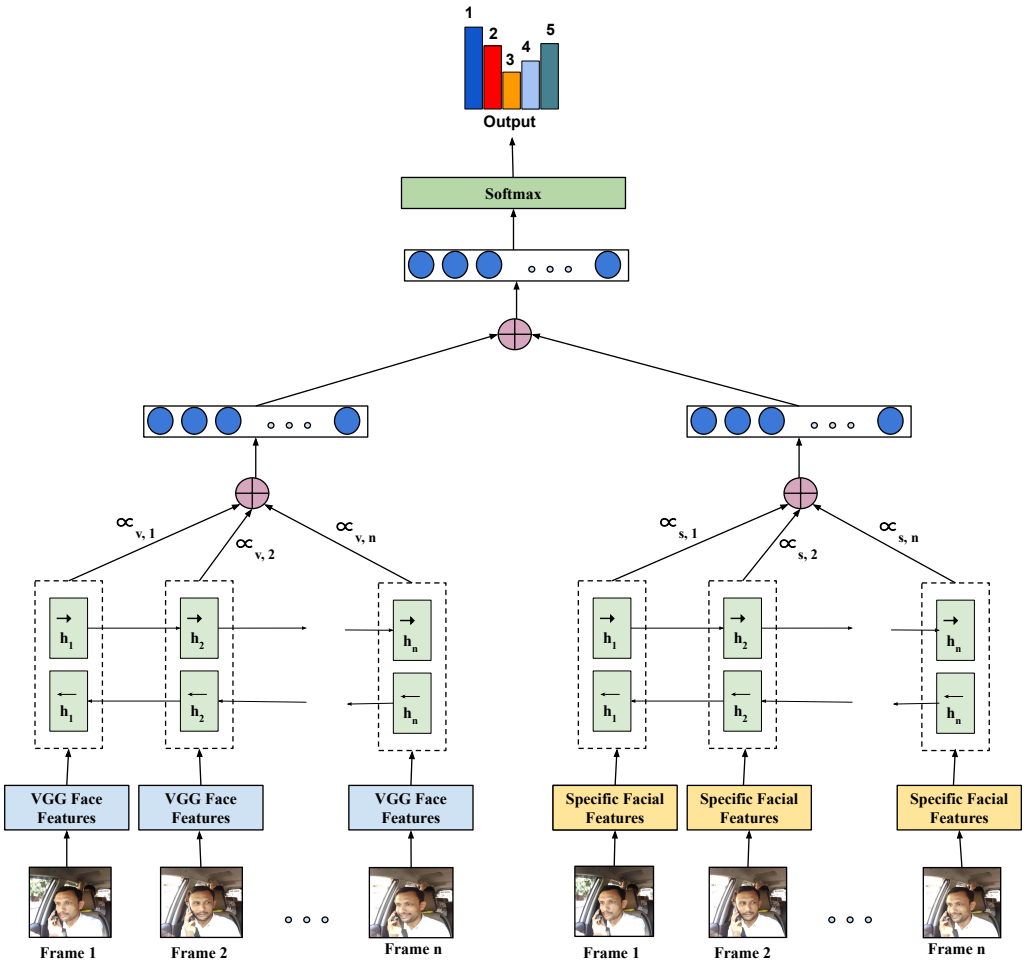


Figure 2.9: Attention Based AutoRate Model

features provide a better representation of the face. This is because VGG16 is trained on imagenet dataset which has 10,000 general objects as classes while VGGFace uses VGG16 as the base model and is trained on 2.6M face images and 2.6K people. (2) We use Bidirectional LSTM instead of GRU to capture driver behavioral information from both previous and future frames.

In addition, the attention probabilities corresponding to each frame are learned when training Attention based AutoRate model. In Attention-Based AutoRate model, instead of using a single vector from the GRU's last hidden state, we add an attention layer to create a weighted connection between the entire source input and the fully connected layer. Length of the entire source input is equivalent to the number of frames in the video. It means that instead of connecting 256 dimensional hidden state

from last block of GRU, we use a weighted linear combination of hidden state from each block of the GRU. Equation 2.4 shows that X is a weighted combination of the hidden state of each block in the GRU that are weighted by alpha score (i.e. attention probabilities). This weighted linear combination of each hidden state weighted by alpha score to the fully connected layer is called attention layer. The same attention layer is used for both left branch with VGG Face features as input and right branch with specific facial features as input. The attention probabilities in both branches is learned by training this Attention-Based AutoRate model end to end.

2.3 Summary

In this chapter, we discussed the facial feature extraction algorithms and presented several approaches for feature combination to predict the final driver attention rating. We showed the proposed architecture AutoRate, a smartphone-based system for driver attention rating. AutoRate employs deep learning techniques that combine generic and specific facial features towards deriving the driver's attention rating. We also present the Attention-based AutoRate model, which learns to attend the key frames and key features required for driver attention rating.

Chapter 3

Evaluation and Visualization of Driver Inattention Rating

In previous chapter, we discussed several techniques like AutoRate and Attention-based AutoRate for feature combination to predict the quality of driver attention on road. We now collect the dataset for training model to predict driver attention rating. In this chapter, we discuss in detail about the process for dataset collection and annotation. Since the objective of driver attention rating system is to derive the rating using the visual features from the camera feed, such that it is equivalent to a rating provided by a human annotator looking at the driver's video. We use human annotation instead of physiological sensors [6] to detect inattention as sensors are intrusive. Due to the inherent subjectiveness of the ratings provided by human annotators, "equivalent to" in this context means making AutoRate "indistinguishable from" human annotators rather than exactly matching a particular human annotator. AutoRate derives a rating by identifying and fusing Spatio-temporal features that affect the driver's attention. AutoRate is trained and tested using an extensive real-world dataset comprising over 3200 unique video snippets, each of length 10 seconds, across 30 drivers in a large city (i.e., 145,000 total images when sampled at 5 fps). We used 5 human annotators to rate each 10-second video snippet on a 5-point scale to get ground truth driver attention rating.

The task is challenging because unlike typical image labeling tasks, the task of annotating video snippets is inherently subjective because there is no clear-cut definition of what constitutes (in)attentiveness. Therefore, we need to rely on multiple human annotators for each video clip. However, that brings up the question of how to reconcile the disagreements in the ratings. One way to overcome this is to eliminate the instances in which human annotators ratings do not match, resulting in a reduced dataset. Another approach is to learn using privileged information (LUPI) [62, 52], where confidence associated with a snippet is used to distinguish between easy and difficult snippets. While LUPI based techniques can be used, our objective is not to distinguish between snippets (easy vs. hard) but rather it is to make

AutoRate's rating of the driver's attention indistinguishable from human rating. To this end, we introduce the concept of kappa coefficient to evaluate AutoRate. We also evaluate the AutoRate model with three approaches: (1) **Mode-based:** In this approach, the mode of the ratings for a video snippet among all the annotators, i.e, the rating with the highest number of votes, is considered as the ground truth rating. We then show AutoRate's efficacy using the F_1 score metric in deriving driver's attention rating that closely matches the majority rating. (2) **Agreement-based:** In this approach, we compute the kappa coefficient (κ) that measures inter-rater agreement between raters [65, 13]. This is considered as a more robust measure than majority-based agreement. We compute the kappa coefficient (κ) between AutoRate's rating and human annotators to show an agreement between the two. (3) **Turing test based [59]:** In this approach, a new human evaluator is presented with the ratings from another human annotator and from AutoRate and is asked to tell which rating came from a human vs. from AutoRate. If the evaluator cannot distinguish between the ratings provided by humans and AutoRate, then AutoRate has done a good job in providing a rating that resembles a human annotator.

Further, we use temporal and spatial attention to visualize the key frames and the key actions which justify the models predicted rating. The temporal visualization identifies the key regions in videos and spatial visualization identifies key features to predict driver attention rating. For example, if attention based AutoRate correctly predicts a driver attention rating as 1 (least attentive and doing illegal activities like phone usage, talking to passengers, etc.) on a 5-point scale, we would want to confirm that the model bases its decision on the features related to using phone or frequent talking with passengers. We observe that driver attention rating is a very subjective problem and features responsible for predicting this rating may vary from one driver to another. To this end, we finetune attention based AutoRate model for a specific user to provide a personalized driver attention rating.

3.1 Dataset

In this section, we describe how we achieve our goal of data collection for the AutoRate task. Each video sample in the dataset consists of 50 frames and a total of 10 sec of length. We collected videos in real driving settings and stationary settings simulating closely to the actual setup. The original frame rate at which videos are sampled is 25fps on an average. The 25fps comes to a total of 250 frames for 10 seconds. But the time taken for an eye blink is 15 blinks per minute [39], which means one blink takes 4 seconds. This means we need to process (25 x 4) 100 frames to get 1 blink. Since the processing

of each frame is costly, we wanted to reduce the frame rate while not missing any crucial details. We decided to use 4 frames to detect eye blink as a compromise between reducing the processing time and not missing any detail. Thus, we chose a frame rate of 5fps for our experiments of video samples of 10 sec each and is collected at the rate of 5 frames per second. We collected 3200 video samples from 30 drivers in different cars in a large city. We then annotated the videos in the range of 1 to 5. We then used the kappa coefficient to eliminate the videos with the less inter-rater agreement as those videos samples are highly ambiguous and challenging to use. We discuss in detail each step below:

3.1.1 Data Collection Setup

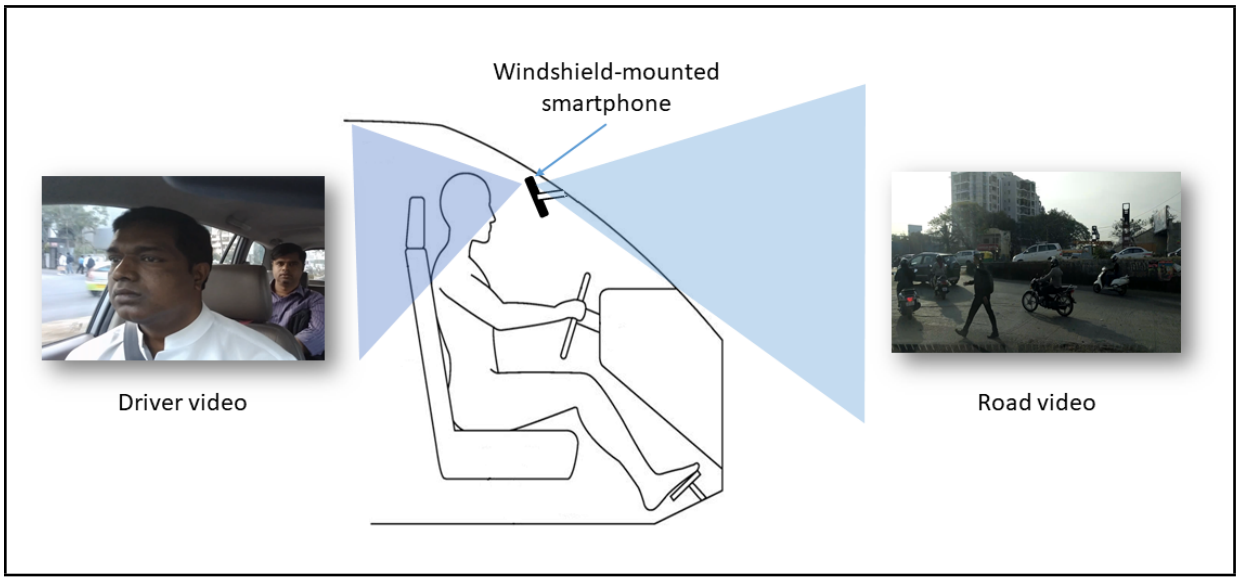


Figure 3.1: Data Collection Setup with a mobile phone camera mounted on the windshield of the car with a focus on the driver inside the vehicle. The mobile phone camera is enabled to collect both driver and road view simultaneously.

We collected the dataset for AutoRate task in both driving and stationary setting. Figure 3.1 shows the data collection setup for AutoRate task. We collected data using the low cost and low power mobile phone camera mounted with the support of the mobile stand on the windshield of the car. We have used the HAMS data collection application, which uses both front and rear cameras of mobile phones to collect both driver and road view simultaneously. The mobile phone is positioned at an angle which gives a 60° view of the scene centered on the driver. Figure 3.1 shows the position of the mobile phone camera collecting both driver and road view for achieving the goal of AutoRate task. The original frame rate at which videos are captured is 25 frames per second. It is collected for 10 drivers in driving settings

in different cars spanning different areas of a large city. The driver has no constraint in this setting and is free to perform any maneuver. As dataset collected in a real driving setting does not provide a balanced dataset with a uniform distribution, we collected data in a stationary setting with 20 drivers by taking the car to different areas of the city. Here the driver was asked to do random maneuvers corresponding to the defined class. We divide this dataset into videos of 10 seconds; each sampled at 5 fps. This improved the distribution of the dataset, and we captured the rare events related to critical but rare classes. We then sent the dataset for the annotation to create clean and unbiased ground truth for each sample.

3.1.2 Data Annotation Tool

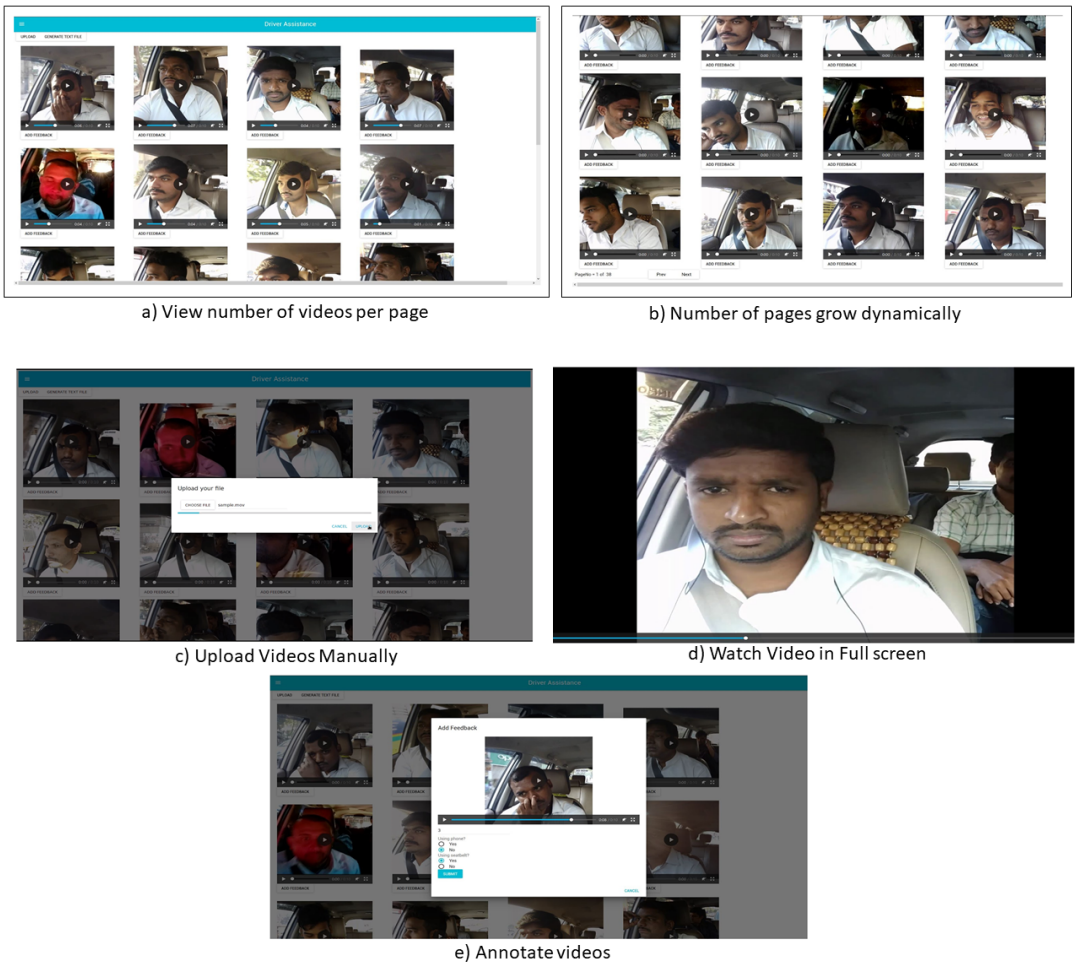


Figure 3.2: Data annotation tool to accelerate the annotation of ground truth rating for the collected video samples. The key features in the data annotation tool include the display of several videos per page, the number of pages can grow dynamically, the user can upload videos manually, watch the video in full screen and annotate videos by answering few related questions.

We used the data annotation tool created internally in React and MongoDB. React is used to develop the front end of the annotation tool because it is used as a base in the development of single-page or mobile applications, as it is optimal for fetching rapidly changing data that needs to be recorded. MongoDB is used for backend operations as it is a cross-platform document-oriented database program. Classified as a NoSQL database program, MongoDB uses JSON-like documents with the schema. Figure 3.2 shows the set of features available in the tool for facilitating the data annotation process. The features include a) Displaying number of videos per page b) Number of pages in tool can grow dynamically concerning the size of the dataset uploaded c) User can upload videos manually d) User can view the video in full-screen e) Annotate videos – a window dialog that provide the user with facility to watch video, enter the rating, check the boxes corresponding to use of phone and seatbelt.

3.1.3 Driver Attention Rating

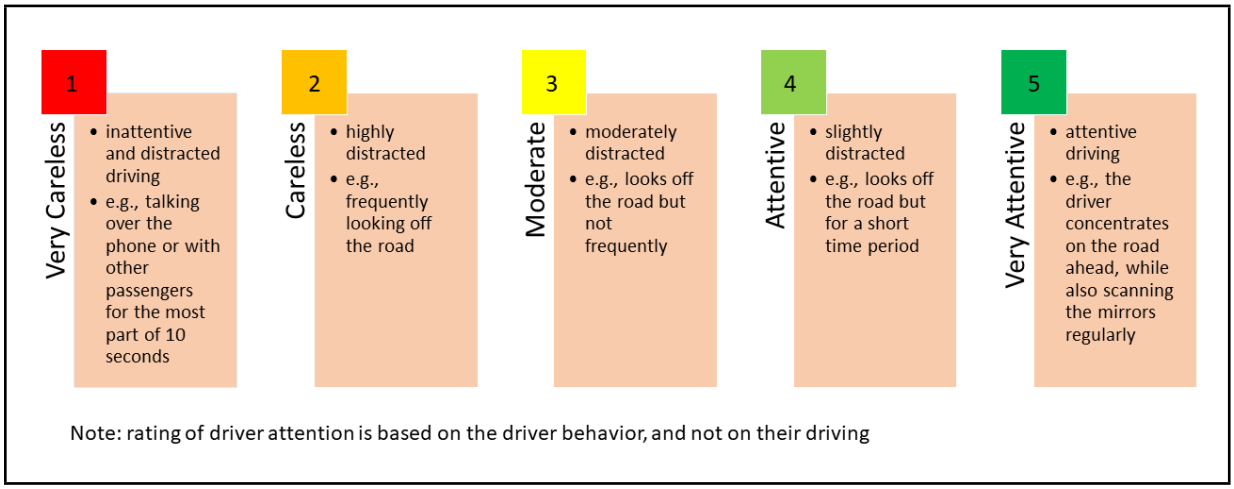


Figure 3.3: Description of driver attention rating from rating-1(very careless) to rating-5(very attentive). These descriptions give a soft understanding of the rating concept and do not bind them to any hard defined rules.

Rating defines the quality of products usually on a scale of 1 to 5 or a scale of 1 to 10. Here 1 means the poor quality and 5 means the superior quality. Driver attention is measured in terms of a combination of various factors like driver drowsiness, driver distraction, eye gaze off-road, etc. These factors may vary in intensity and can have different outcomes. To measure driver attention with robustness, we introduce driver attention rating in the range of 1 to 5. Rating is the overall value given based on the driver state and behavior; it takes into consideration the change in intensity of drowsiness, distraction,

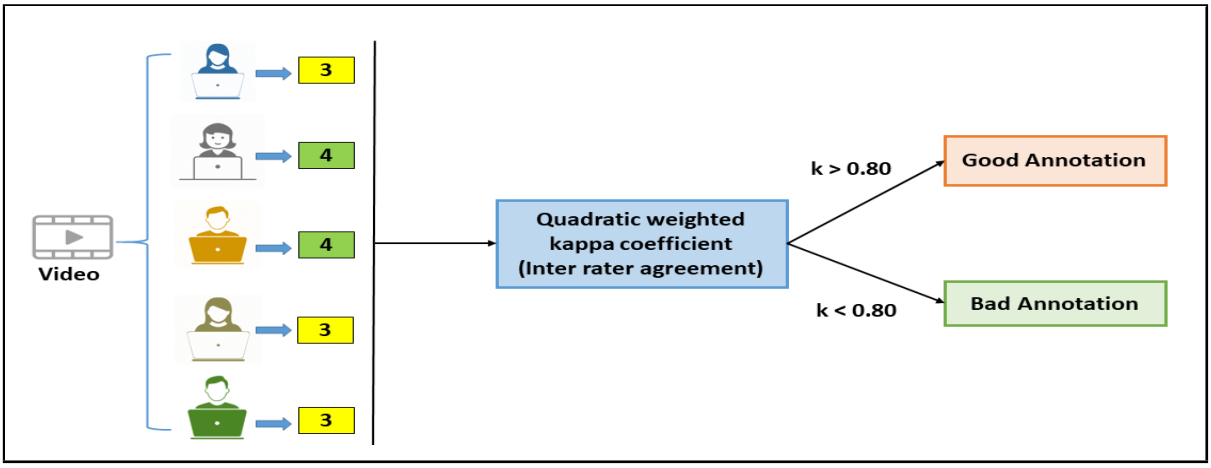
loose focus, illegal driving, etc. Here in this chapter, we use only driver state and behavior to rate the quality driver's attention. After months of the experiment, we came up with the following definitions for driver rating. These ratings are not hard rule-based and are very generic. It helps to bring all annotators on the same norms for rating. Here, `Rating-1` represents *inattentive* and distracted driving, e.g., talking over the phone or with other passengers for the most of 10 seconds. `Rating-2` represents driver being *highly distracted*, e.g., frequently looking off the road. `Rating-3` represents driver being *moderately distracted*, e.g., looks off the road but not frequently. `Rating-4` represents driver being *slightly distracted*, e.g., looks off the road but for a short time period. `Rating-5` represents *attentive driving*, e.g., the driver concentrates on the road ahead, while also scanning the mirrors regularly to maintain situational awareness.

3.1.4 Data Annotation

Figure 3.4 presents the pipeline for data annotation process. Each video sample is presented to five human annotators. Based on their understanding of driver attention rating, they are asked to rate each video sample in the range of 1 to 5. We explained each rating to annotators so that they are not too far in the opinions for the same rating. This is a subjective task, and all annotators may have a different opinion about the same video sample. We used the inter-rater agreement metric (kappa coefficient)to get the good annotation and remove the with high ambiguity. We use the kappa coefficient value of 0.80 to decide the good or bad annotation. The threshold is decided based on the kappa value(0.80) corresponding to a strong level of understanding.

3.1.5 Dataset type and its distribution

We considered two datasets [40]: (i) Driving dataset, where we collected data from real driving scenarios, and (ii) Static dataset, where we collected data in a static vehicle setting. We split the video into 10-second snippets, allowing fine-grained driver attention analysis. Each 10-second video snippets was then rated by the human annotators based on the driver's attention level, ranging from rating-1 (least attentive) to rating-5 (most attentive). Note that, such an annotator would not have access to the full range of signals (e.g., vehicle jerks, honks, etc.) that might inform the assessment of a person who was actually at the scene. So this is a limitation of our study. We now provide a detailed description of our datasets.



Interrater Reliability		
Value of κ	Level of agreement	% of data that is reliable
0 - .20	None	0 – 4%
.21 - .39	Minimal	4 – 15%
.40 - .59	Weak	15 – 35%
.60 - .79	Moderate	35 – 63%
.80 - .90	Strong	64 – 81%
Above .90	Almost perfect	82 – 100%

Rating is subjective and annotator-specific.
If κ > .80, we consider it as a good rating.

Figure 3.4: Data annotation: Five human annotators annotate each video sample. The annotations are highly subjective and annotator specific. Inter-rater agreement, like the kappa coefficient of threshold 0.80, is used to select the samples with good annotation. Note that the table below shows the impact of kappa value from None to almost perfect(k¿0.90).

3.1.5.1 Driving dataset

In this dataset, we collected real-world driving data by deploying smartphones in a fleet of 10 cabs across multiple days[1]. In total 8 hours of data was gathered across the 10 cabs. As mentioned earlier, we then split the video's into 10-second snippets. Finally, only a subset of 10-second snippets is selected to ensure that the correlation between consecutive videos is avoided. In total we retained around 1000 video snippets from 10 drivers. The training and test split for this dataset is shown in Table I. We see that rating-5 has over 800 samples (out of the 1000 snippets in all) whereas rating-1 has fewer than 100

[1]HAMS Project: `https://aka.ms/HAMS`

Dataset	Driving		Static		Merged	
	Train	Test	Train	Test	Train	Test
Rating-1	68	14	582	129	650	143
Rating-2	59	13	183	33	242	46
Rating-3	55	13	289	62	344	75
Rating-4	133	37	294	64	427	101
Rating-5	566	121	434	91	1000	212
Total	881	198	1782	379	2663	577

Table 3.1: Dataset description with train and test split.

samples. This reflects the situation that drivers are attentive most of the time. Nevertheless, the instances of inattentiveness, even if relatively few, could have serious safety consequences, so it is important to be able to rate these accurately.

3.1.5.2 Static dataset

As noted above, the data is skewed towards the driver being attentive and it is challenging and also risky to gather inattentive driving data in real-world settings. To get around this difficulty and augment the inattentive driving data, we performed targeted data collection with 20 different drivers in a static vehicle to improve the data distribution for ratings 1 to 4. We asked the driver to perform various actions (as realistically as possible) corresponding to the definitions of each rating described in Section 3.1.3. Table 3.1 shows the training and test split for the static dataset.

3.1.5.3 Merged dataset

To create this dataset, we merge both the driving and static datasets. In total this dataset includes data from 30 drivers with approximately 3200 videos each of 10 seconds. Table I shows the training and test split in the merged dataset.

3.1.6 Dataset Summary

Table 3.2 summarizes the characteristics of the dataset.

Dataset Summary	
Features	**Value**
Number of clips	3240
Length of clips	10 seconds
Number of annotations per clip	5
Number of drivers	30
Number of classes	5
Ground truth	Majority/Average Rating

Table 3.2: Dataset Summary

3.2 Evaluation Metrics

We now describe the various metrics used for evaluation of performance of our proposed architecture. We use F1 Score which is the weighted average of Precision and Recall and kappa coefficient (κ) which measures the inter rater agreement between various annotators.

3.2.1 F1 Score

It is a measure of test's accuracy and is defined as harmonic mean of the precision and recall.

$$F1 = 2 \cdot \frac{P \cdot R}{P + R}, \tag{3.1}$$

where P and R represents the precision and recall, respectively. Precision (P) is computed by first considering each predicted class (i.e., predicted driver attention rating) in turn and computing the fraction of predictions in that class that are correct, i.e., match the ground truth. Then the fractions are combined across the classes, using the weighted arithmetic mean to obtain the overall precision. The Recall (R) is computed analogously, by considering the ground truth classes instead of the predicted classes.

3.2.2 Kappa coefficient (κ)

Kappa coefficient (κ) between two annotators [65] measures agreement between two annotators and defined as,

$$\kappa = \frac{P_o - P_e}{1 - P_e},$$
$$P_o = \sum_{i}^{R}\sum_{j}^{R} w_{ij}x_{ij}, \;\; P_e = \sum_{i}^{R}\sum_{j}^{R} w_{ij}m_{ij}, \tag{3.2}$$

where P_o is the relative observed agreement between annotators and P_e is the probability of chance agreement if the annotators were totally independent. R represents the total number of ratings (in our case, 5) and w_{ij}, x_{ij} and m_{ij} corresponds to the weight, observed and expected values, respectively. If the annotators are in complete agreement then $\kappa = 1$ and if there is no agreement then $\kappa = 0$.

In this paper, we use quadratic weighted kappa [13], where we treat disagreements differently, for e.g., difference between ratings off by 2 is penalized more than ratings off by 1. The weight assigned to each rating category is given by,

$$w_d = 1 - \frac{d^2}{(R-1)^2}, \tag{3.3}$$

where d is the difference between ratings.

3.3 Experiments and Results

3.3.1 Implementation Details

We used original frame rate at which videos are sampled is 25fps on an average. The 25fps comes to a total of 250 frames for 10 seconds.The time taken for an eye blink is 15 blinks per minute [39] that means 1 blink takes 4 seconds of time. This means we need to process (25 x 4) 100 frames to get 1 blink. Since the processing of each frame is costly, we wanted to reduce the frame rate while not missing any crucial details. We decided to use 4 frames to detect eye blink as a compromise between reducing the processing time and not missing any detail. Thus, we chose a frame rate of 5fps for our experiments.

Feature Extraction: In this section, we explain the complete facial feature extraction procedure and the relation between these models. We first extract **face area** which is used to determine the distance of the driver from the camera. Face area is obtained from the bounding box of the state of the art face detection model proposed in "Finding Tiny Faces" [26]. For **facial landmark detection**, we used FAN(Face Alignment Network) [17] which is the state of the art network for face landmark detection. It is trained on LS3D-W dataset of size 230,000 images. It works for pose values ranging from -90° to +90°. Now, we determine the driver **head pose** (yaw, pitch and roll) using "Head pose estimation in

the wild using convolutional neural networks and adaptive gradient methods" [45] trained on PRIMA dataset. This paper uses tightly cropped face image as input to the network. In order to use this pre-trained network for head pose estimation, we use the landmarks detected using FAN to tightly crop the face image before using it as input to the pre-trained head pose model. As we don't have ground truth annotation for head pose on our dataset, we plot the face images corresponding to the head pose values across the video to verify the correctness of the model.

Further, we use facial landmarks detected using FAN for predicting **eye aspect ratio(EAR) and mouth aspect ratio(MAR)** which uses the detected eye landmarks and mouth landmarks respectively. The EAR is computed as ratio of height of the eye to the width of the eye and similarly, we compute the MAR. Now, to determine the **eye gaze** value we use the modified LENET-5 architecture proposed in "Appearance based gaze estimation in the wild" [74]. The modified LENET-5 takes left eye image as input to the network and predicted head pose value is concatenated in the last fully connected layer. So, we use the eye landmarks detected using FAN to crop the left eye image from the driver face for input to eye gaze network. We also use the head pose value predicted using the above head pose network for concatenation in the last fully connected layer for eye gaze prediction.

For **phone detection**, we collected around 1200 sample images when the driver is talking on the phone (by holding it up to their face) and manually marked the bounding box around the phone suing YoloMark tool. The bounding boxes are marked such that there is not too much margin around the object and the object annotation is of good quality. The labeled images with bounding box of the phone was used to fine tune the pre-trained YOLOv2 [47] network trained previously on COCO dataset [70]. The final predictions are then restricted to only detection of a phone and the corresponding bounding box in an image. The YOLO was fine tuned for two classes (Phone and No Phone). The batch size used for fine-tuning the model is 64 and total number of epochs 1000. The output is phone detection confidence and the bounding box location. If the confidence value is above the threshold value(0.7), the phone is detected else not detected. **seat belt detection** follows the same procedure as phone detection. The accuracy for phone and seatbelt detection is 97% and 80% respectively.

We have used pre-trained network for all specific features except for phone and seatbelt detection. As there was no model for phone and seatbelt detection so, we fine tune YOLOv2 for the same. Yes, the performance of individual model affects the final accuracy of Attention based AutoRate. For each individual feature, we tried multiple competitive models and we picked the one that worked best on our

dataset. Table 3.3 shows the detailed information about the feature, state of the art algorithm used to extract the features, output type, output dimension and performance of different pre-trained models.

Feature	Description	Output Type	Dimension
Face Detection [26]	Face detected or not	Binary (0 or 1)	1
Face Area [26]	X,Y location on frame of size 500 x 500	0 to 500	4
Facial Landmark [17]	X,Y location on frame of size 500 x 500	0 to 500	128
Head Pose [45]	Yaw, Pitch and Roll	-90° to +90°	3
Eye Gaze [74]	Gaze in X and Y direction	Real Values	2
EAR [57]	Ratio of height to width of eye	Real Values	1
MAR [57]	Ratio of height to width of mouth	Real Values	1
Phone Detection [47]	Using phone or not	Binary(0 or 1)	1
Seat Belt Detection [47]	Wearing seatbelt or not	Binary(0 or 1)	1

Table 3.3: Detailed description of the feature extraction.

System Performance: Each video sample has 50 frames and total time taken to extract facial features is approximately 6 minutes/video on an average. The time used to predict driver rating using Attention based AutoRate is 1.2 second approximately. Total time taken at run time is 7 minutes and 20 seconds due to which the model cannot provide real time performance.

3.3.2 Qualitative Results

Figure 3.5 shows the qualitative results obtained from the Attention-based AutoRate mode. The first line shows the actual rating obtained from the annotators and rating we get from the proposed design. The video frames is plotted along with the corresponding attention probabilities. Darker the color, more inattentive is the driver. In row-1, we observe that driver is inattentive for very less time and is rated 5 which matches the ground truth rating of 5. While in row-3, driver is inattentive for most of the time and is rated 3 by the model and ground truth rating is 2 which is close.

3.3.3 Quantitative Results

We now present our evaluation of the CNN+GRU architecture, AutoRate, and Attention-based AutoRate for driver attention rating. We also show the efficacy of our model on datasets captured under

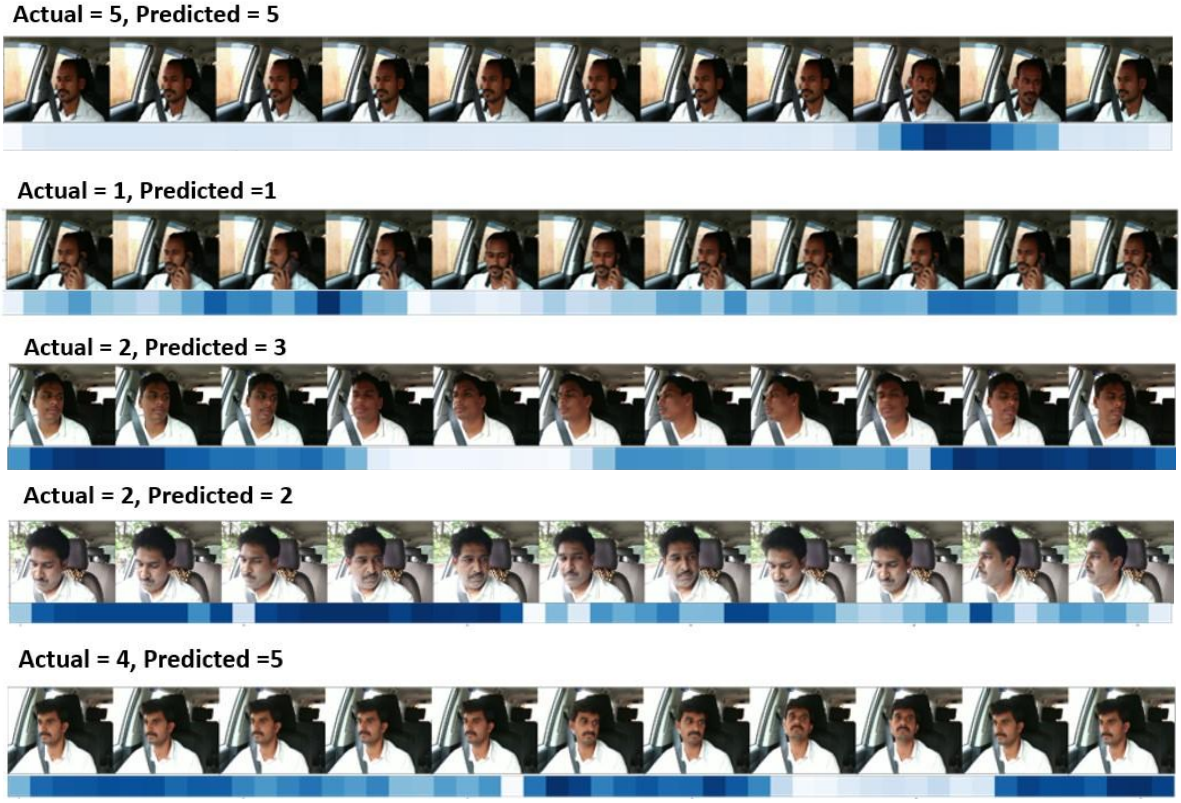


Figure 3.5: Qualitative Results from the Attention-based AutoRate architecture

various conditions. We also present a detailed ablation study to compare different models for predicting driver attention ratings. We also perform the Turing test evaluation for the misclassified videos.

3.3.3.1 Ground Truth Rating

Driver attention rating is a non-trivial task as there is no clear-cut definition of what constitutes (in)attentiveness. In some cases it may be hard for the annotators to distinguish between driver frequently looking off the road against moderately looking off the road. This results in ambiguity, where the ratings obtained differ from one annotator to another. Hence, it is important to first understand the agreement between annotators before evaluating AutoRate's efficacy. In our experiments, we used five human annotators to rate the 10 second video snippets. In the driving dataset, the average agreement between all the five annotators is 0.90 and in static dataset the agreement is 0.87. The kappa coefficient for the merged dataset is 0.89.

This exhibits that there is no perfect agreement among the five annotators and hence some of the video snippets may not have *true* ground truth ratings. In light of this, in the sections that follow, we evaluate Attention Based AutoRate using the three approaches; Mode-based, Agreement-based, and Turing test based evaluation.

We asked the five annotators to rate the video snippets based on their notion of driver attention, i.e., without providing them any guidelines or definition for each rating. However, this resulted in poor agreement, with a kappa of just 0.5 in the driving dataset. Hence, we proceeded to provide the annotators some broad guidelines and definitions for the various rating levels, to boost the degree of agreement.

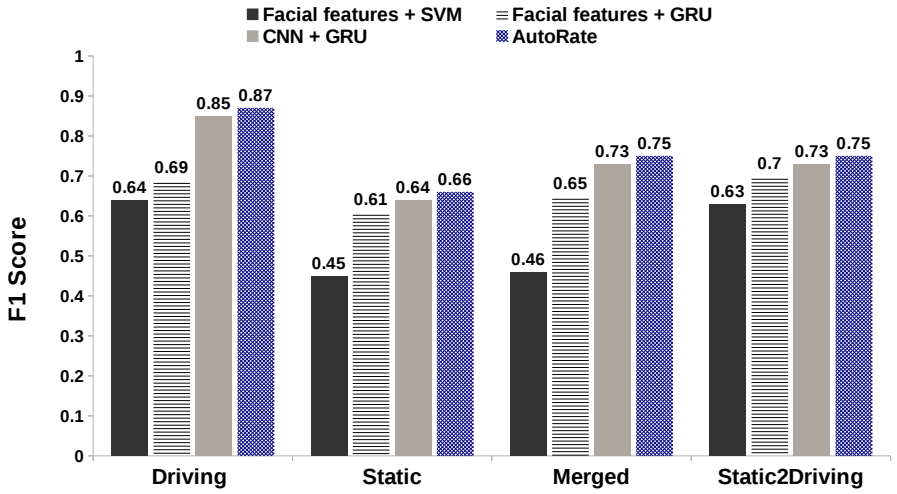


Figure 3.6: F1 score for four methods for all three dataset(driving, static, merged) and static2driving

Datasets	Attention Based AutoRate **vs Majority**	Attention Based AutoRate **vs Average**
Driving	0.90	0.83
Static	0.87	0.8
Merged	0.89	0.86
Stat2Driving	0.85	0.82

Table 3.4: Agreement between Attention Based AutoRate and Majority/Average ratings using kappa coefficient.

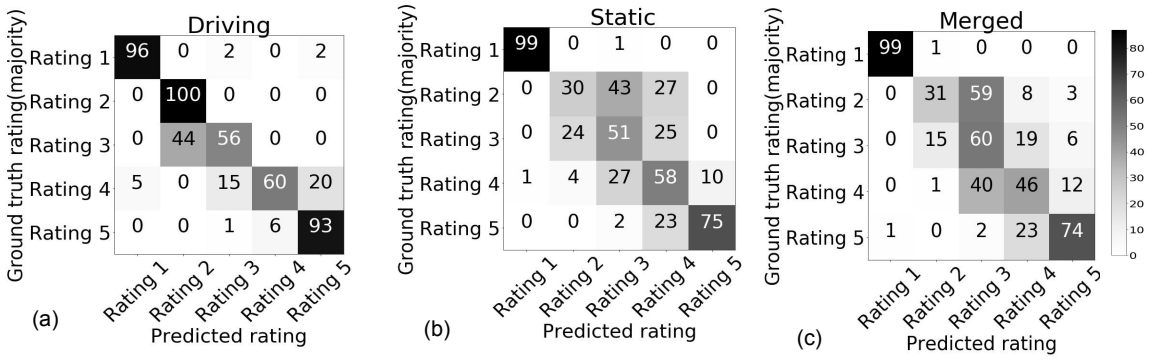


Figure 3.7: Confusion matrix obtained for Attention Based AutoRate for (a) driving, (b) static and (c) merged datasets. For each ground truth rating, the row of numbers represents the percentage of predicted ratings from 1 to 5. The higher the percentage the darker the shade of the cell.

3.3.3.2 Mode Based Evaluation

We now present results where we consider the mode of the ratings for a video snippet among all the annotators as our ground truth rating. Figure 3.6 shows the F1 score for AutoRate, CNN+GRU and other approaches across all the three datasets. The results are obtained after doing 10-fold cross-validation across all the datasets. F1 score of AutoRate and CNN+GRU is consistently higher than other approaches. We also plot the F1-score for the model trained on static data and fine-tuned on 1000 driving data. The F1 score reported 0.75 is purely on driving data, which is on par with that of a model trained entirely on driving data, 0.87. This indicates that our pre-trained model can be used for different road conditions just by fine-tuning using a minimal amount of data. Figure 3.7 shows the confusion matrix for AutoRate, where each cell of confusion matrix shows the percentage of predicted rating. The off-diagonal values are high for adjacent rating levels indicating the ambiguity in the ground truth which results in majority of misclassifications.

3.3.3.3 Agreement Based Evaluation

We now present evaluation based on the kappa coefficient to quantify the agreement between Attention Based AutoRate's predicted rating with the individual human annotator rating. Table III shows the agreement between Attention Based AutoRate, mode and average rating among the 5 human annotators using the kappa coefficient(κ). We first compute the mode and average ratings (rounded using the floor function) for each video snippet across all the human annotators. We then compute kappa coefficient between the human rating and AutoRate's rating using Equation 3.2.

It can be seen that for the driving dataset, Attention Based AutoRate has an overall agreement of 0.90 and 0.83 with the mode and average ratings, respectively. Note that, for the same driving dataset, among the 5 annotators the agreement was 0.89. Further, Attention Based AutoRate has around 0.89 agreement for mode and 0.86 average ratings provided by human annotators in the merged dataset. This indicates that the driver attention rating predicted by AutoRate matches closely with the ratings provided by human annotators which is 0.88.

Figure 3.8 shows the agreement between the ratings obtained by AutoRate and each individual human annotator across all datasets. For the driving dataset, the kappa coefficient is around 0.90. Given that the agreement among the human annotators in driving dataset was itself low (i.e., 0.88), we conclude that Attention Based AutoRate is doing quite well in mimicking a human annotator.

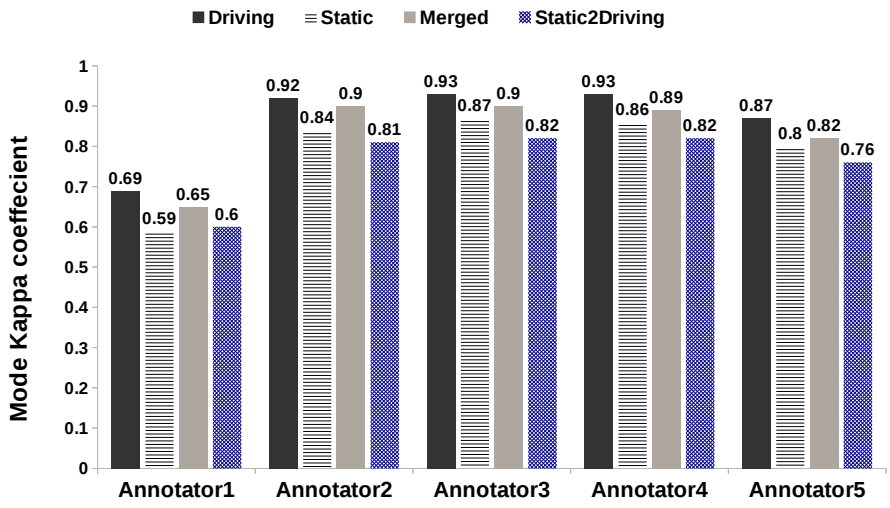


Figure 3.8: Agreement between AutoRate ratings and human annotators across datasets.

3.3.4 Ablation Study

We performed extensive ablation studies on our Attention-based AutoRate model which we present in Table 3.5. In the first 5 rows, we present the impact of various facial features and their combinations on the predicted rating. The facial features were concatenated across all 50 frames before classifying. Here, we have used an SVM to classify the driver rating instead of GRU as the feature dimensionality is too small for a complex network like GRU and can lead to overfitting. We observe that all the facial features combined together give the best results among the first 5 rows. We further noticed that on

Method	Acc	F_1	Mode κ	Avg κ
HP + SVM	0.33	0.54	0.30	0.28
HP + EG + SVM	0.3	0.45	0.27	0.19
EB + yawning + SVM	0.23	0.32	0.13	0.11
HP + EG +EB + yawning + SVM	0.35	0.46	0.42	0.37
All facial features + SVM	0.42	0.53	0.7	0.68
All facial features + GRU	0.45	0.59	0.8	0.74
CNN(VGG16) + GRU	0.58	0.60	0.84	0.82
CNN(VGGFace) + GRU	0.62	0.64	0.85	0.82
AutoRate [20]	0.64	0.66	0.88	0.86
VGGFace + AttentionLSTM	0.73	0.74	0.87	0.85
Attention based AutoRate	**0.75**	**0.75**	**0.89**	**0.86**

Table 3.5: Comparison of attention based AutoRate model with AutoRate[20] model, CNN + GRU and other feature combinations. Note: κ denotes kappa coefficient used for inter rater agreement. The abbreviations used in the table stand for Head Pose (HP), Eye Gaze (EG) and Eye Blink (EB).

Method	Acc	F_1	Mode κ	Avg κ
Attention Based AutoRate(AA)	**0.75**	**0.75**	**0.89**	**0.86**
AA without any facial features	0.73	0.74	0.87	0.85
AA without VGGFace	0.53	0.42	0.54	0.54
AA without EAR and MAR	0.70	0.69	0.85	0.84
AA without EG and HP	0.69	0.69	0.81	0.79
AA without P and S	0.69	0.68	0.87	0.84
AA without FaceArea	0.72	0.72	0.88	0.86

Table 3.6: Performance of Attention Based AutoRate model as a result of stripping input features one by one.

comparing Row 1, which uses only head pose as the facial feature and Row 4 which uses a combination of four facial features, they have almost the same accuracy but a decent improvement in the mode kappa value. This shows that head pose and eye gaze are the most important features for driver inattention prediction but using combination of all features is definitely beneficial.

Next, we explore the effect of using a GRU instead of concatenating the features for all the frames in rows 6, 7 and 8. We notice that using GRU gives a slight improvement over the former. In rows 7 and 8, we have used generic features extracted from the VGG16 network pretrained on ImageNet or the VGGFace network. We can see that deep features work much better than specific facial features. Using VGG-Face features instead of VGG16 gives a 4% increase in accuracy. AutoRate, which combines both

VGG-Face features as well as specific facial features gives an increase of 10% accuracy over the model that uses only specific facial features.

The last two rows of Table 3.5 explores the effect of adding attention to the models. We can see a significant jump in accuracy of 10% compared to AutoRate. This shows that Attention-Based AutoRate learns to attend to key frames in the video and key actions performed to predict driver attention rating. This selection of features and frames assist the model to deal with videos for which specific facial feature values is not detected. Note that Attention-Based AutoRate has kappa coefficient of 0.89 which is in close agreement with kappa coefficient computed for human annotators(0.89).

Table 3.6 shows the results of an ablation study conducted on the Attention-based AutoRate model. We strip one input feature at a time from the proposed model and observe the performance. We obtained the highest drop of 29% and 44% in accuracy and mode kappa value respectively by removing the VGGFace features. This implies that VGGFace features are the most important features to predict driver attention rating. We then removed the facial features one by one and found that eye gaze, head pose, phone, and seatbelt are the most influential of the specific features. Other facial features like face area, eye blink, and yawning have less impact as these have a high correlation with the prominent specific features determined earlier. We further found that all features combined together get the best results.

3.3.5 Turing Test

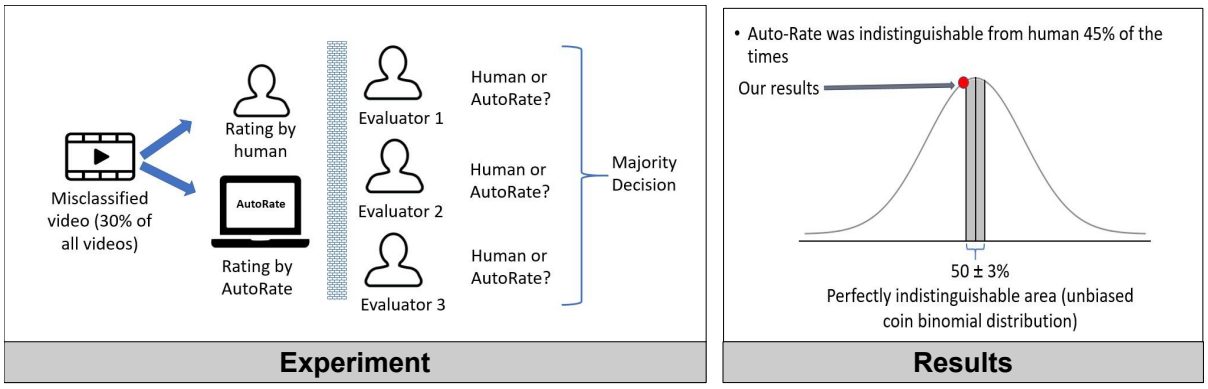


Figure 3.9: Turing test on misclassified samples.

We now report on a *Turing test* [59], where a new human *evaluator* is presented ratings from another human annotator and from Attention Based AutoRate. The job of the evaluator is to tell which rating came from the human vs. from Attention Based AutoRate. If the evaluator cannot reliably tell which

rating came from whom, then Attention Based AutoRate would have done a good job in rating driver's attention. Note that the focus here is on having Attention Based AutoRate be indistinguishable from a human annotator, not on accuracy per se, although the latter would likely have a bearing on the former.

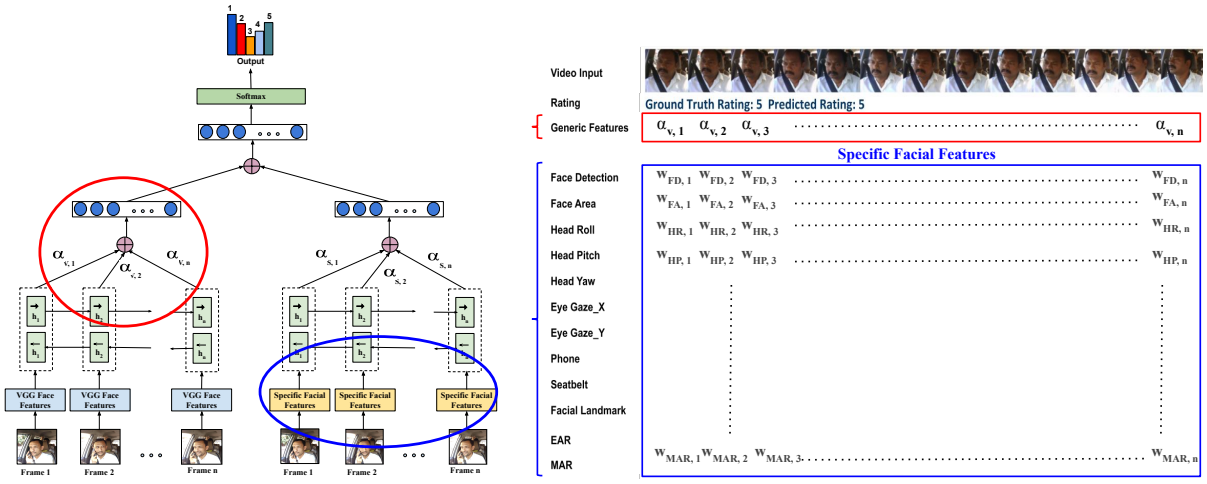
On our unseen dataset (i.e., 782 test videos), we first determined the ratings predicted by Attention Based AutoRate. Proposed network's rating match with the human rating for 70% of the videos (i.e., 549 out of 782). The samples that were misclassified (i.e., 782-549=233 video snippets) were presented to 3 evaluators along with the human rating and the Attention Based AutoRate rating. Each evaluator decided which of the ratings across the 233 snippets came from a human and which from Attention Based AutoRate. For each snippet, we picked the majority decision, i.e., where two or three of the evaluators were in agreement. We found that in 55% of cases, the majority decision was correct, i.e., it correctly called out human ratings vs Attention Based AutoRate ratings. Thus, the Attention Based AutoRate ratings in majority of the cases is perfectly indistinguishable from human ratings, i.e., based on an unbiased coin binomial model we would have expected the majority decision to have been correct 50% of the time, with a standard deviation of 3%. Hence ratings derived by AutoRate is mostly indistinguishable from a human, and can be applied to rate driver attention effectively.

3.4 Visualization

Visualization is key to confirm that the model bases its decision on the right set of features. For example, if attention based AutoRate correctly predicts a driver attention rating as 1 (least attentive and doing illegal activities like phone usage, talking to passengers, etc) on a 5-point scale, we would want to confirm that the model bases its decision on the features related to phone usage or frequent talking with passengers. We now present the method used for spatial and temporal visualization.

3.4.1 Visualization Mechanism

Figure 3.10a shows Attention-based AutoRate model which learns attention probabilities for both generic feature branch and specific feature branch of attention based AutoRate model. These attention probabilities are used for spatial and temporal visualization. The temporal visualization identifies key frames in a video and spatial visualization identifies key features in the frame to predict driver attention rating. For temporal visualization, we plot the attention probabilities ($\alpha_{V,1}...\alpha_{V,n}$) learned corresponding to each frame from generic feature (VGGFace feature) branch of attention based AutoRate model.



(a) Attention based AutoRate model

(b) Visualization of attention probabilities

Figure 3.10: Visualization Mechanism: Interpretability of Attention based AutoRate model for temporal and spatial visualization of videos.

The attention probability for temporal visualization is marked with red circle in Figure 3.10a and red box with generic feature label in Figure 3.10b. For spatial visualization, we plot the learned attention probabilities corresponding to features like head pose, eye gaze, phone, etc in frame from specific facial feature branch of attention based AutoRate model. The attention probabilities for spatial visualization is marked with blue circle in Figure 3.10a and blue box under the label specific facial feature block in Figure 3.10b. Each row under the label specific facial feature block in 3.10a represents specific feature. Darker the color on attention map, more is the relevance of frame or feature in predicting the rating. For example: row 3 of specific feature block shows attention probabilities assigned to head roll feature in each frame ($w_{HR,1}...w_{HR,50}$), where HR = Head Roll. Similarly for Head Yaw, Head Pitch, Eye Gaze X and Eye Gaze Y. Heat map corresponding to phone and seatbelt detection shows the time stamp in video during which the driver is using phone and wearing seatbelt respectively.

3.4.2 Visualization Results

Figure 3.11 presents the analysis for visualization results from attention based AutoRate model. Darker the color on attention map, more is the relevance of frame or feature in predicting the rating. The first row in figure show frames of video input (Every fourth frame is plotted for ease of visualization). The second row shows the ground truth rating given by annotators and rating predicted by attention based AutoRate model. In the third row, we plot attention probabilities learned for each frame

using generic features (VGGFace features) as input. From fourth row onwards, we show the attention probabilities corresponding to each specific facial features. Note that generic features are used as input in the left branch of the attention based AutoRate model and specific facial features are used in right branch of the attention based AutoRate model. Now, we present in detail analysis for two videos for which the visualization result is shown in figure 3.11. In Case I, we observe that predicted driver attention rating is 5 which is equivalent to ground truth driver attention rating given by majority of annotators. The third row in this example corresponding to generic feature label show that attention probabilities for generic features are high for a very small section of the video input. Corresponding to this small section of video, high probability value for 'Head Roll' and 'Eye Gaze' concludes that the driver is not looking straight on road for this section of the video. This small section is approximately 10% of the total video which shows that the driver is attentive for the majority of the video. Hence predicted rating is 5. We also observe high attention probability for seatbelt and low attention probability for phone, which is an important factor for good driving and rating-5.

Case II in Figure 3.11 shows the predicted rating of 1 which is equivalent to the rating given by the annotators when the driver is using a phone for any length of the video or talking to passengers for most of the time. We observe high attention probabilities for generic features in two small sections of the video input. Corresponding to the first small section of the video, high probability value for 'Head Pitch', 'Head Yaw' and 'Phone' concludes that driver is using a phone (illegal activity) and not looking straight on road. We also observe high attention probability for MAR showing that the driver is talking. Corresponding to the second section of the video input, we observe that 'Head Yaw', 'Eye Gaze' and 'Phone' are the major reason for driver inattention. From the above two small sections, we conclude that driver attention rating is 1 because he is using a phone and is inattentive.

From the above two cases, we observe that attention probabilities from generic feature show the regions of driver inattention and specific facial feature block can be used to reason about the driver inattention. This can also be used to summarize the driver inattention over long video and provide the reason for the same. Further, analysis on misclassified videos helps us to understand the reason for misclassification of equivocal videos.

3.4.3 More Visualization Results

We present 4 more examples of visualization results. Darker the color on attention map, more is the relevance of frame or feature in predicting the rating. First row in visualization results show frames of

video input (Every fourth frame is plotted for ease of visualization). Second row shows the ground truth rating given by annotators and rating predicted by attention based AutoRate model. In third row, we plot attention probabilities learned for each frame using generic features (VGGFace features) as input. From fourth row onwards, we show the attention probabilities corresponding to each specific facial features. Now we present visualization results accompanied with their analysis:

3.5 Personalization

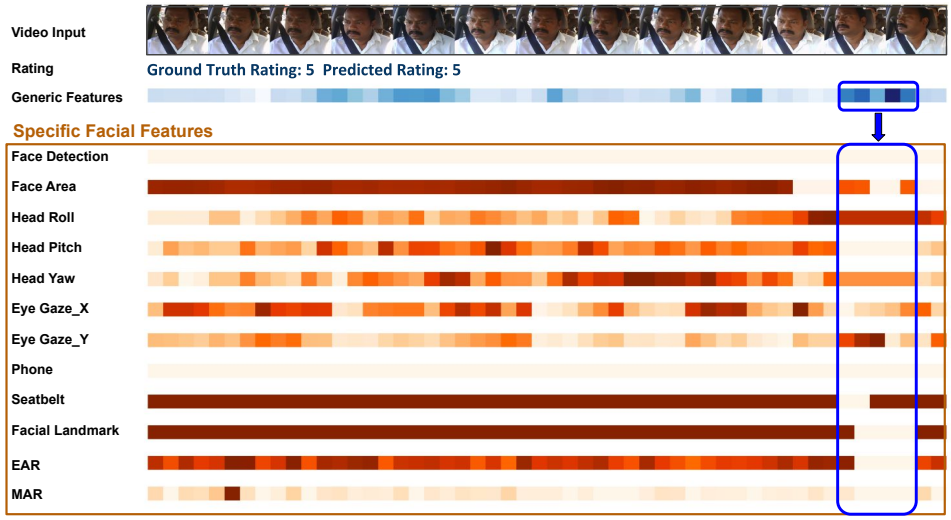
Personalization in attention based AutoRate can be used to improve driver specific results. Here, personalization refers to the generation of driver attention rating based on specific driver behaviour. This the can be achieved by finetuning the attention based AutoRate model using driver specific data. Table 3.7 demonstrates the result of 6 random drivers before and after finetuning attention based AutoRate model. The first five rows in the table show that accuracy, F1 score, mode kappa, and average kappa improves by a significant amount on finetuning the attention based AutoRate model for a specific driver, but the last row shows that it may also go down by some value if the driver-specific dataset is highly imbalanced. In the real world, an imbalanced dataset is a big problem and can be addressed by finetuning the attention based AutoRate model every few days using randomly sampled balanced set of videos.

Driver	Attention based AutoRate							
	Before Finetuning				After Finetuning			
	Acc	F_1	Mode κ	Avg κ	Acc	F_1	Mode κ	Avg κ
Driver1	0.57	0.70	0.60	0.56	0.65	0.74	0.85	0.82
Driver2	0.55	0.58	0.56	0.54	0.71	0.77	0.78	0.75
Driver3	0.63	0.69	0.36	0.37	0.63	0.74	0.46	0.49
Driver4	0.29	0.36	0.22	0.20	0.61	0.63	0.81	0.78
Driver5	0.65	0.73	0.78	0.75	0.79	0.84	0.91	0.88
Driver6	0.46	0.60	0.08	0.13	0.43	0.46	0.38	0.41

Table 3.7: Evaluation of personalization on 6 random drivers from dataset.

3.6 Summary

In this chapter, we have discussed the data collection process in detail. We have evaluated AutoRate on a real-world dataset with 30 drivers. AutoRate's automatically-generated rating has an overall agreement of 0.88 with the ratings provided by 5 human annotators on static dataset. We also show the results obtained on a model trained on static dataset and tested on driving dataset is comparable to the result obtained by training and testing on the driving dataset. In addition, we show that Attention Based AutoRate model outperforms AutoRate model by 10% accuracy on the extended dataset. Our analysis shows that Attention Based AutoRate's driver attention rating closely resembles a human annotator rating, thus enabling automated rating system. We also show the spatial and temporal visualization of Attention Based AutoRate model which helps to determine the region of inattention in videos and the key action performed that leads to this inattention. We further show personalization in attention based AutoRate for user specific accuracy. The features and code is available at `https://github.com/duaisha/AutoRate`.



(a) Case I: Driver with attention rating-5. In above figure, we observe that attention probability corresponding to generic features is high for 10% of the total video length. This means driver is attentive in majority of the video and hence the rating-5. Specific facial feature block corresponding to this inattentive region of video show high attention probability for 'Head Roll' and 'Eye Gaze'. This concludes that driver is not looking straight on road for this section of the video and hence the reason for his inattentiveness.

(b) Case II:Driver with attention rating-1. In above figure, corresponding to generic feature label we observe high attention probability in two section of video length specifying inattention for approximately 20% of the video length. Specific facial feature block corresponding to the first inattention region show high attention probability for 'Head Pitch', 'Head Yaw' and 'Phone'. Specific facial feature block corresponding to the second inattention region show high attention probability for 'Head Yaw', 'Eye Gaze' and 'Phone'. In both region, high attention probability for phone usage supports model decision to rate driver attention as 1.

Figure 3.11: Visualization results from Attention Based AutoRate. Darker the shade in the cell of attention map, higher is the impact of the feature in predicting driver attention rating.

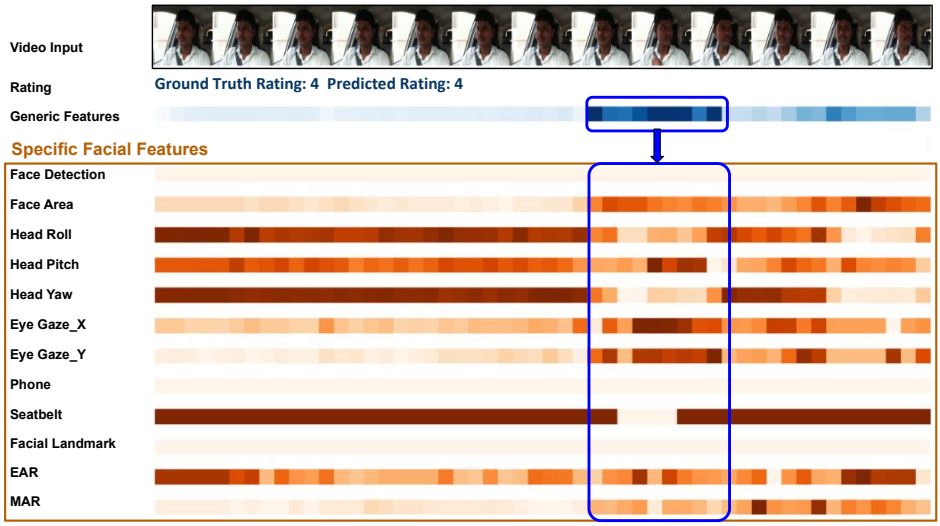


Figure 3.12: Driver with attention rating-4. In above figure, we observe that attention probability corresponding to generic features is high for 20% of the total video length. This means driver is attentive in 80% of the video and hence the predicted rating-4 which is equivalent to ground truth rating. Specific facial feature block corresponding to this inattentive region of video shows high attention probability for Eye Gaze X, Eye Gaze Y and Head Pitch which further concludes that driver is not looking straight on road for this section of the video. This further justifies driver attention rating as 4. We also see high probability value for seat-belt, which defines him a good driver.

Figure 3.13: Driver with attention rating-3 and predicted rating-3. In above figure, we observe that attention probability corresponding to generic features is high for more than 40% of the total video length. This means driver is attentive for less than 60% of the video and hence the predicted rating-3 which is equivalent to ground truth rating. Specific facial feature block corresponding to this inattentive region of video show high attention probability for Eye Gaze X, Eye Gaze Y, Face Area and Head Roll which concludes that driver is not looking straight on road for this section of the video. This further justifies driver attention rating as 3.

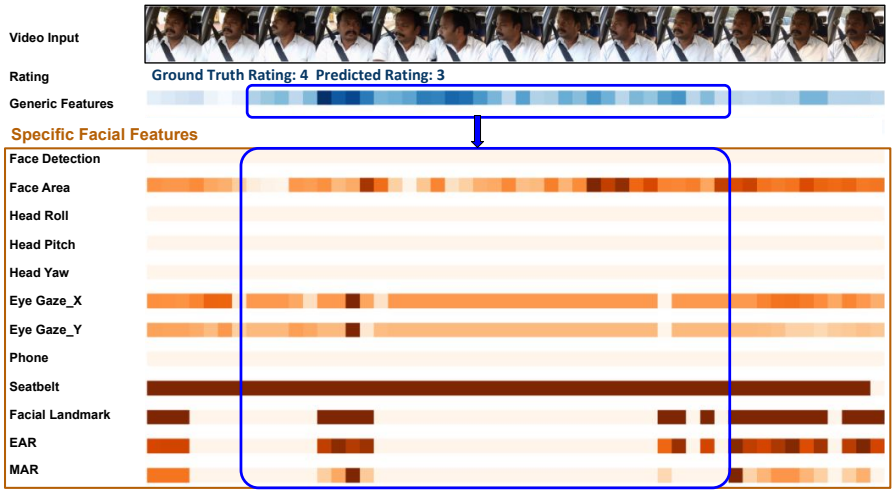


Figure 3.14: Driver with attention rating-4 and predicted rating-3. In above figure, we observe that attention probability corresponding to generic features is very high for 10% of the total video and decently high for 60% of the total video length. This means driver is properly attentive for only 30% of the video. Specific facial feature block corresponding to this inattentive region of video show high attention probability for Face Area, Eye Gaze X, and Eye Gaze Y but fails to predict Head Roll, Head Pitch and Head Yaw value which is important as video input shows a lot of variation in head pose. This means failure in specific facial feature effects final attention rating.

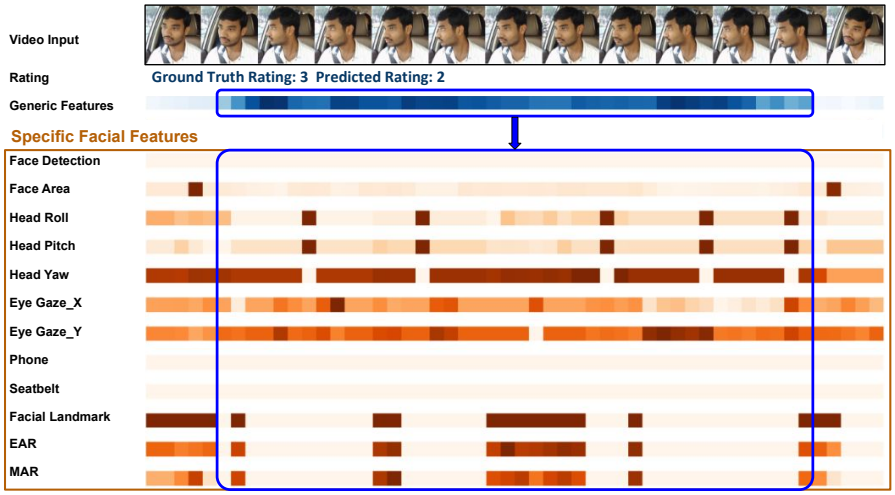


Figure 3.15: Driver with attention rating-3 and predicted rating-2. In above figure, we observe that attention probability corresponding to generic features is high for 80% of the total video length. This means driver is attentive for 20% of the video and hence the predicted rating-2 which is not equivalent to ground truth rating-3. Specific facial feature block corresponding to this inattentive region of video show high attention probability for Head Yaw, Eye Gaze X, and Eye Gaze Y. Visually the generic and specific features looks aligned with the video input but the network still fails to predict correct rating. This happens because the input video sample is ambiguous and it is difficult for even human annotators to rate it as 2 or 3.

Chapter 4

Driver Gaze Mapping on Road

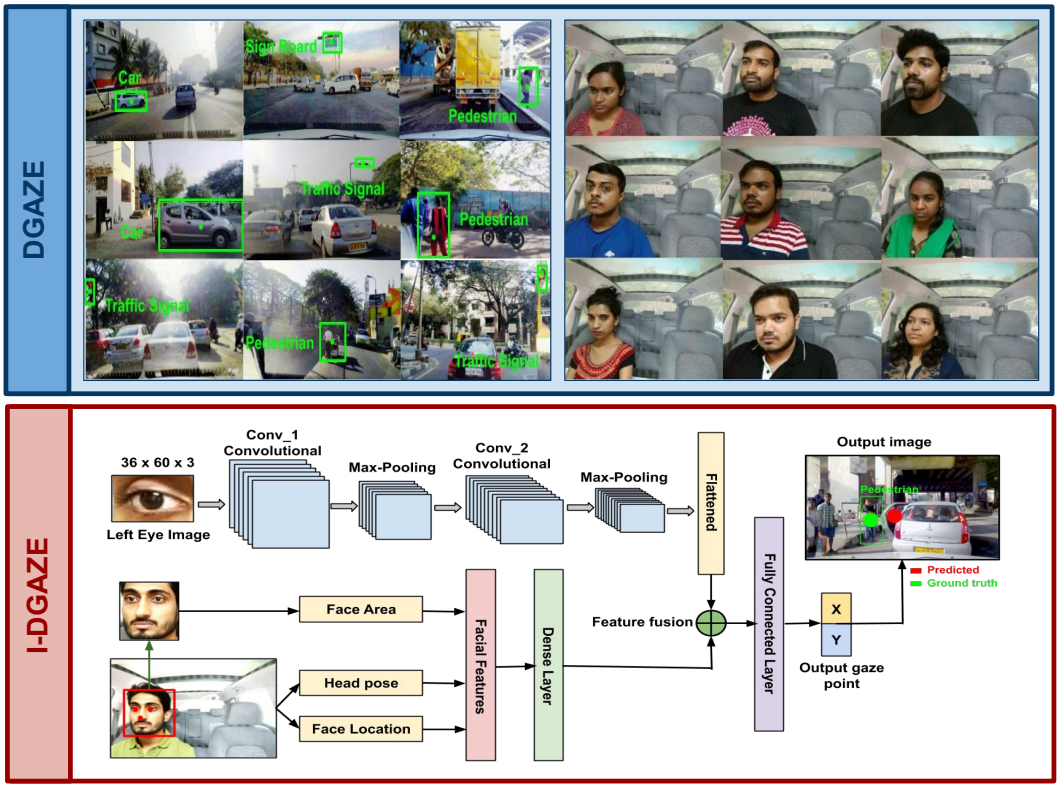


Figure 4.1: In this work, we develop DGAZE the driver gaze mapping dataset that includes both driver and road view to capture driver gaze on road using low cost mobile phone cameras. Using DGAZE, we train I-DGAZE for prediction of gaze point on road.

In this chapter, we propose the novel technique for driver attention monitoring using a fusion of driver and road videos. Traditionally, various approaches are proposed for driver attention monitoring, where the aim is to analyze the driver's state and behavior to determine driver attention on the road. These approaches focus on detecting driver drowsiness, driver fatigue, and rating the driver's attention.

These advanced systems can contribute to reducing fatal road accidents by a factor of more than 30%. However, relatively few vehicles on the road today have these systems, and their share of the market is growing at only 2 to 5 percent annually. The major problem preventing massive implementation of these systems is that these systems are found only in the high price vehicle segment. This implies that improved driving safety is restricted to the top purchasing power consumers. In this chapter, we propose a technique for driver gaze mapping on-road using late fusion of driver and road videos.

Driver gaze mapping on road is closely associated with eye gaze tracking. Eye gaze tracking is Driver gaze mapping on the road can be executed by employing various eye gaze tracking algorithms. Eye gaze tracking is popularly used across a range of different research fields ranging from psychological research to medical diagnosis, neuro-marketing applications, and beyond. With recent advancements in eye gaze tracking, this has become a commercially stable device to track driver gaze. The driver gaze can be used for the analysis of driver attention on the road, length of the driver gaze on a specific object, the order in which visual elements are fixated upon. These eye-gaze trackers can be used for avoiding various accidents due to driver inattention. But the major disadvantage of using these wearable eye-gaze trackers is that they range from a few thousand to few lakh rupees, which is not affordable. The other downside is that as these trackers are mounted on lightweight eyeglasses, we cannot obtain an unobstructed driver image. We can only get the road view. Thus, these images cannot be used to create a dataset. So, even if we achieve an acceptable error for our gaze tracking algorithm, we still need to buy these costly trackers to predict gaze at test time. We tackle this problem by collecting a large dataset containing both road view and driver view using low-cost mobile phone cameras. So, that at test time, we can use a mobile camera installed with an eye gaze tracking algorithm.

Traditionally, there exists several publicly available eye gaze dataset. Gaze Capture is the accessible eye gaze dataset containing 2.5M frames, but this eye-tracking data is collected for gaze points on mobile devices and tablets. On the other hand, we collect the gaze data for target objects on road videos projected on the wall. The driver gaze prediction on projected videos simulates driver gaze prediction on the actual road. We collected the data in simulation for the following reasons (1) Eye gaze trackers like EyeTribe, Tobii EyeX are costly, and it is not feasible to collect data by pointing the objects on the road. (2) It is also hard to verify if the driver is following instructions while looking at the objects on the road. (3) It is not feasible to collect large scale data for training the deep neural network for predicting driver gaze on the road. While in a simulation setting, we can obtain the data easily by showing the annotated object on the screen, and a large amount to data can be collected for training purposes.

Our main contributions are threefold: (1) We introduce the DGAZE dataset for predicting driver gaze on the road. The dataset is collected using low cost and low power devices like a mobile phone camera. (2) We propose the baseline architecture I-DGAZE for predicting driver gaze on-road using a fused convolutional neural network. The network use eye image as input to the first branch and facial features along with face location as input to the second branch of the fused CNN. (3) We further improve the results of I-DGAZE using calibration.

4.1 DGAZE: Driver Gaze Mapping Dataset

In this section, we present the DGAZE dataset for the prediction of driver eye gaze on the road. Each sample includes an image of a road and a corresponding image of a driver looking on the road, as shown in Figure 4.1. Our dataset has 100,000 images of 103 unique objects on the road which belong to 7 object classes. We also include images corresponding to 9 calibration points for each participant. The dataset is collected with 20 drivers of age group 20-30 and height between 150 cm to 180 cm on average. The dataset has both male and female participants with and without spectacles. In the section below, we describe the process of data collection using low-cost mobile phone cameras and the statistics of the DGAZE dataset.

4.1.1 Dataset Collection

We now discuss in detail about the DGAZE dataset collection procedure in simulation using low-cost mobile phone devices.

4.1.1.1 Dataset collection setup

The data is collected in a lab which is set up to closely mimic driving conditions. Figure 4.2 shows that the driver sits in car-like setup in data collection room. A video of a road is projected in front of a seated subject. The road videos are collected from dashboard-mounted cameras of cars driving on actual roads. The backdrop behind the person depicts the interior of a car, in order to make it more realistic. The distance between the driver and the projected video is adjusted to match real driving conditions. The video of the subject and the projection is collected simultaneously. In order to do so, a mobile phone is set up on a tripod between the person and the projected video. The front camera is used to record the person while the back camera records the projected video. The application used for recording

Figure 4.2: DGAZE Collection Setup: Dataset is collected in lab setting which has close proximity with actual driving setting. Mobile phone camera attached to tripod stand similar to mobile phone camera mounted on wind shield of the car collects both driver view and projected road view at same frame per seconds.

is previously used for AutoRate[20] Data collection. The frames are then extracted from these videos to create the image dataset.

4.1.1.2 Object Annotation

Eye gaze trackers are mostly used to collect the driver gaze data on road but it is very costly and ranges from few thousands to lakh rupees. Instead, we annotated the object of interest on the road video with a bounding box and asked the participant to look at the center of the box at those specific points. The objects in the video are annotated using the dlib [15] correlation tracker implementation based on Accurate Scale Estimation for Robust Visual Tracking[15]. The tracking algorithm works in real-time by tracking the objects that change in both translation and scaling throughout a video sample.

4.1.1.3 Dataset collection

DGAZE dataset is collected in a lab setting with 20 drivers including both male and female. The height of the drivers range from 150 to 180 cm and includes drivers with and without spectacles. An 18-minute long video is projected for each driver on the wall in front of the driver seat. The video has 9 samples for calibration and 103 samples with annotated 7 unique objects on road. These objects

Figure 4.3: Samples from DGAZE dataset corresponding to seven unique objects annotated on road. Note the significant variation in the size of the object, distance of the object from the driver and illuminance variation on road. These variations help the I-DGAZE model to train well on the dataset.

include car, motorbike, pedestrian, etc. Figure 4.3 shows the object annotated with a bounding box and its center point corresponding to the label of each object annotated on road. A mobile phone camera captures both the driver video and the projected road video. The front and back videos are collected at different frame rates and the original road video has 25 frames per second. We align the original road video (output) to the driver view (input) by dropping a frame at each frame-drop location in the longer video. The following equation is used to compute the frame-drop:

$$fd = \frac{max(fc(rv), fc(dv))}{diff(fc(rv), fc(dv))} \tag{4.1}$$

where fd denotes frame drop, fc denotes the frame count, rv is the road video, and dv denotes the driver video. Note that $diff$ is abbreviation for difference.

4.1.2 Dataset Statistics

In this section, we provide detailed statistics of DGAZE dataset and its comparison with existing datasets. DGAZE dataset has both driver and road views collected with 20 drivers. It has 103 annotated objects consisting of 7 unique objects. The objects annotated includes pedestrian, cars, motorbikes, auto-rickshaw, traffic signals, and signboards. Figure 4.3 shows some sample frames from the DGAZE

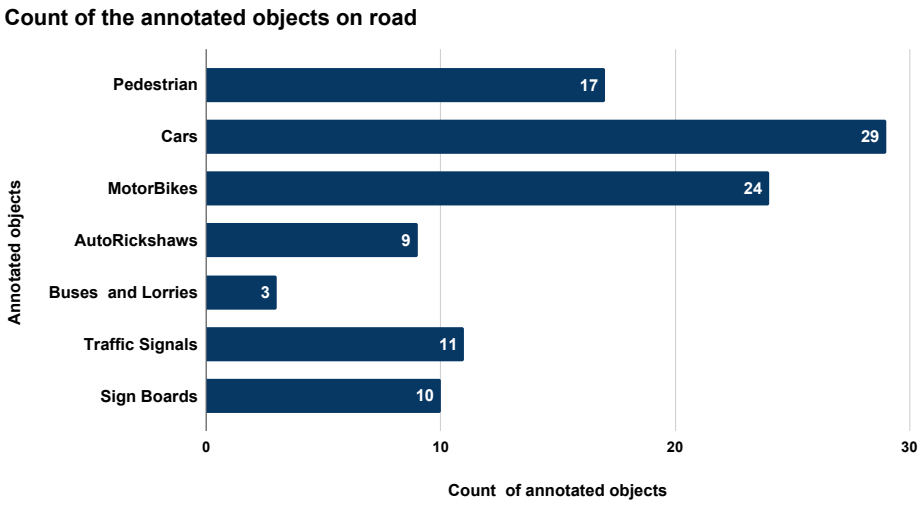
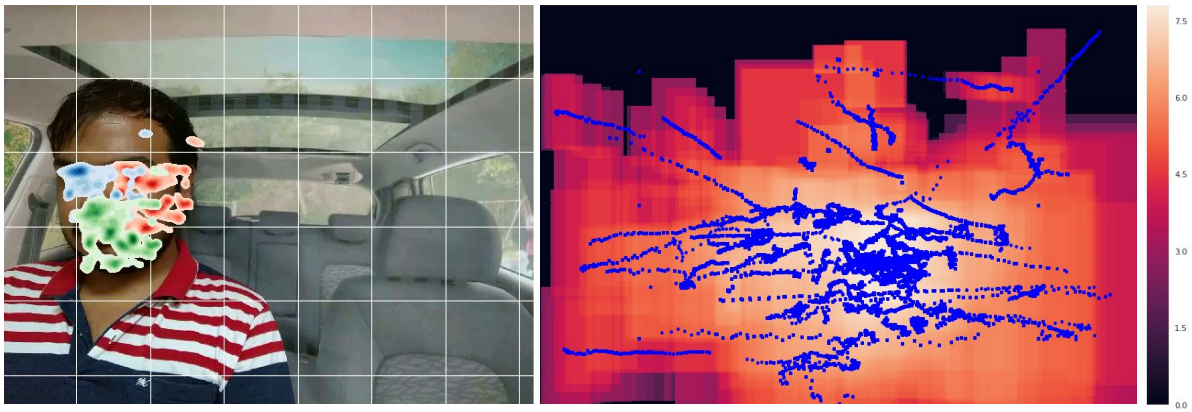


Figure 4.4: Number of annotated objects in views corresponding to each unique object on road

dataset corresponding to the seven unique objects. Figure 4.4 shows the number of annotated objects in the views corresponding to each unique object listed above. We have additionally collected 9 stationary calibration points for each driver. This is then used to extract images, which amounts to a total of 100,000 images. The road images have varied illumination as the images are captured from morning to evening in the real cars on actual roads. 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 4.5(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 4.5(a) shows how the position of the eyes and mouth vary in the dataset samples, We observe that there is a good variation of eye and mouth positions, but they remain within realistic bounds for drivers.

Table 4.1 shows the comparison of DGAZE dataset with other existing datasets. Works such as [37, 67, 55, 38, 58, 74, 27, 32] 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.



(a) (b)

Figure 4.5: 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.

4.2 I-DGAZE: Gaze Prediction Architecture

We now present I-DGAZE architecture for predicting driver gaze on road. The network is a two-branch late fusion convolutional neural network with eye image as input to one branch and facial features (like head pose, face location and distance of driver face from mobile phone camera) as input to the other branch. We first detect the face and facial landmark locations using a deep convolutional neural network trained for the corresponding tasks. We then extract the head pose features using a CNN trained on the PRIMA [23] dataset for predicting yaw, pitch and roll of the driver's face. The late fusion of these specific facial features from one branch along with eye image features from the second branch is used to predict driver gaze on road. In Section 4.2.1, we describe the facial features required for I-DGAZE and the algorithms used to extract these features. In section 4.2.2, we describe the I-DGAZE architecture.

4.2.1 Facial features

We extract several facial features from the driver face such as head pose, face location, face area, etc to help with eye gaze prediction. We detail the algorithms used for these features below.

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

Table 4.1: Comparison of DGAZE with other eye gaze mapping dataset

4.2.1.1 Face Area

Driver gaze prediction depends on the distance of the driver's face from the mobile phone camera mounted on the windshield of the car. It also depends on the height of the driver which varies a lot, especially between man and woman. Since the distance between the driver and the camera is not directly apparent from the driver image, we use the face area as a proxy feature. The assumption is that the face area does not vary considerably between people. Thus, a large area can be considered as the face being closer to the camera and vice versa. To obtain the Face Area, we first detect the face using a state of the art face detection algorithm[26] and then use the output from the face detection network to compute the face area.

4.2.1.2 Face Location

Face location is used to determine the relative position of the driver in the car with respect to the mobile phone camera mounted on the windshield of the car. The face location of the driver depends on the facial landmark locations, which is obtained by using the several existing techniques ranging from active appearance model to Convolutional Neural Networks to extract [5], [25], [30]. In this work, we employ pretrained Face Alignment Network (FAN)[17] to extract facial landmarks.

4.2.1.3 Head Pose

Head pose estimation is a key feature for determining where the driver is looking. Techniques for predicting driver head pose ranges from traditional techniques such as PnP (Perspective-n-Point) algo-

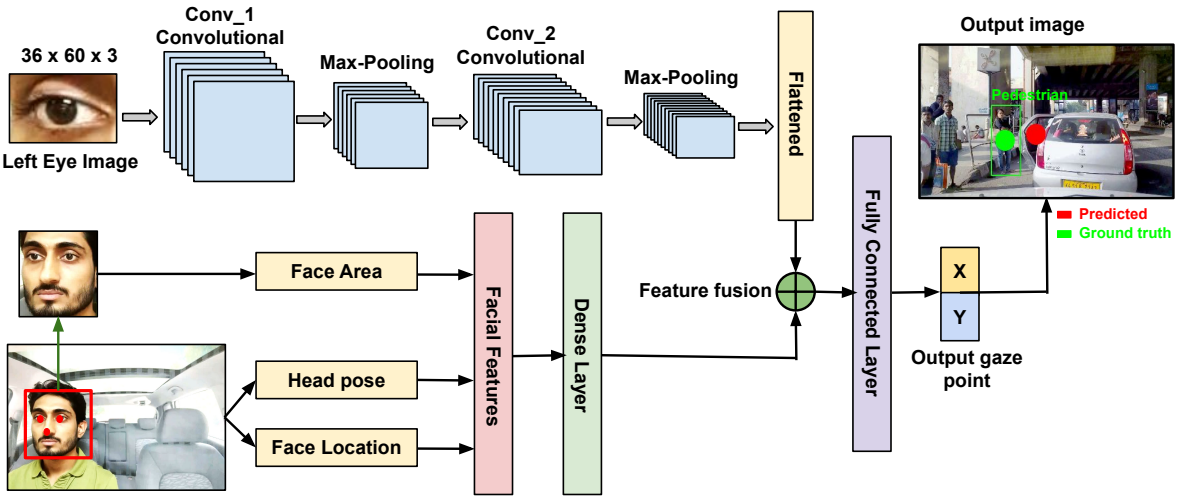


Figure 4.6: I-DGAZE Architecture to predict driver gaze on road. The network is a two branch late fusion convolutional neural network with input to one branch as eye image and input to other branch as facial features like head pose, face location and distance of driver face from mobile phone camera.

rithms [41] to deploying CNN for predicting driver's head pose. The pre-trained CNN network like Deepgaze[45] is trained using dataset such as PRIMA [23], AFLW[36] and AFW[31].

4.2.2 I-DGAZE Architectire

Figure 4.6 shows an overview of our proposed architecture I-DGAZE. The architecture is a two-branch fused convolutional neural network. The input to one branch is the left eye image of dimension (36 x 60 x 3). The eye region is obtained using the tight crop around the facial key landmarks detected around the eyes using the above-detailed algorithm. Here, the assumption is eye gaze and pupil location of both the eyes moves in aligned fashion. This image is then passed through Le-Net style model used in [34] where the image is first passed through a first convolutional layer with 20 channels followed by max-pooling and then passed to another convolutional neural network with 50 channels followed by max-pooling and then the output from max-pool is flattened to get 4550-dimensional feature vector. In the second branch of I-DGAZE, we use extracted face detection features to obtain the face area. Also, we extract facial landmark locations on the driver face image to compute the driver location in the car relative to the car seat and windshield mounted mobile phone camera. Along with face area and face location, the head pose is a key factor to determine driver's pose while driving which in turn helps in predicting driver gaze on road. The features face area, face location, and head pose together are mapped to 16-dimensional feature vector which is then fused with 4550 features extracted from the first branch.

These 4466 features (first branch + second branch) are combined and mapped to the 512 hidden units of a fully connected layer. The output of the network is the x and y coordinates of the predicted eye gaze on the road. We compute the loss between predicated coordinate values and actual coordinate values using mean absolute error. The model is trained until the error reaches its minimal value or error saturates. Further, we observed that eye gaze is a subjective task and changes from person to person. So, after training the fused convolutional neural network, we finetune the network for each driver using the data collected corresponding to each calibration point which reduced the error to 45 pixels.

4.3 Experimental Evaluation and Results

In this section, we thoroughly evaluate the performance of I-DGAZE on the DGAZE dataset. I-DGAZE outperforms other state-of-the-art models in predicting driver gaze. The error for driver gaze is 94.5 pixels without calibration and is 85 pixels with calibration. We also present the qualitative results for the I-DGAZE architecture and an ablation study to evaluate the various components of I-DGAZE. For training the model, the dataset is divided into 90% train, 5% validation and 5% test data. Further, we also use 1000 images collected corresponding to the calibration points for finetuning the I-DGAZE architecture.

4.3.1 Evaluation Metrics

We evaluated our results using mean absolute error between the ground truth gaze point and the predicted gaze point as used previously in [27][32]. The error between the prediction and ground truth is noted in pixels. 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.

4.3.2 Qualitative and Quantitative Results

Figure 4.7 presents the qualitative results of the predicted driver gaze fixation. From left to right: column 1 shows 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. The distance between the two is 100 pixels which is used to identify the region of driver attention on road but to determine the fine level prediction, we fine-tune the model using images corresponding to 9 calibration points. In column 5, we show the

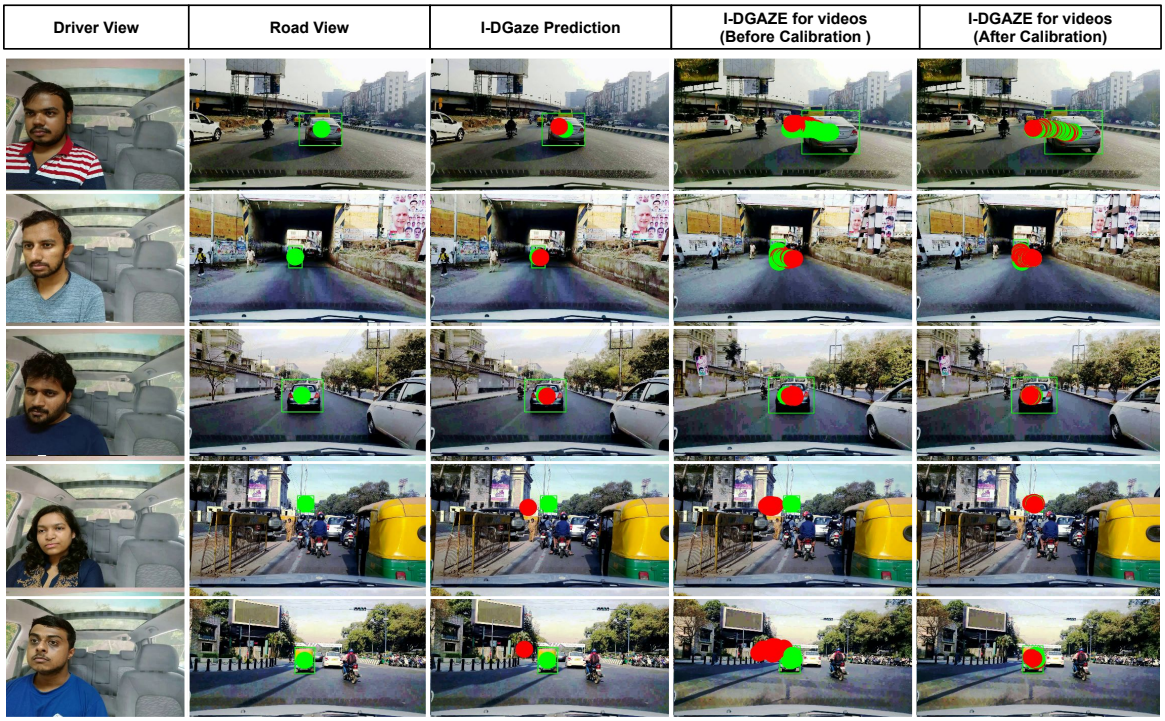


Figure 4.7: Qualitative assessment of the predicted driver gaze fixation. From left to right: driver image, road image, I-DGAZE for gaze prediction in images, I-DGAZE for gaze prediction in multiple frames without calibration and I-DGAZE for gaze prediction in multiple frames with 9-point calibration.

result of I-DGAZE model on video sample and we observe that the ground truth and prediction never overlaps each other and results in avg error of 95 pixels on average. Column 6 shows the result of I-DGAZE after fine-tuning the model using images corresponding to calibration points for one driver. The error got reduced to 10 pixels after calibration and hence the eye gaze model gives fine gaze prediction on road.

Table 4.2 shows the quantitative results obtained by training I-DGAZE model on DGAZE dataset and its comparison with other state of the art eye gaze models. We compare our results with methodologies proposed in Turker Gaze[68], where they use pixel level face features as input and use Support Vector Regression to estimate gaze point on screen. We also compare our results with MPII Gaze [74], which has state-of-the-art results for eye gaze estimation in wild and we compare it with Eye tracking for everyone which predicts user gaze on phone and tablet. We observe that I-DGAZE outperforms all the above approaches as it is fusion of high resolution pixel level eye image in one branch and specific facial features relevant to the gaze prediction in other branch. We obtain an improvement of pixels over other network. Table 4.3 shows the ablation study of our model by removing one facial feature at a time. We observe that the results obtained using I-DGAZE with all the facial features outperforms highlighting the importance of all facial features in predicting driver gaze on road.

Method	Without Calibration			With Calibration		
	Train Error	Val Error	Test Error	Train Error	Val Error	Test Error
Turker Gaze[68]	300	350	338.5	140.5	232	240
MPII Gaze[74]	103.4	115	117	80.4	95	97.5
iTracker[32]	90	92	98	40	65	68
I-DGAZE	**81.3**	**95**	**94.5**	**25**	**40**	**45**

Table 4.2: Comparison of I-DGAZE with existing gaze prediction methods

4.4 Summary

This chapter presents driver attention monitoring using the fusion of driver and road videos. We propose the DGAZE dataset, which includes both driver and road view collected using mobile phone camera mounted on tripod stand mimicking the actual car driving setting. The dataset can be used for various applications like driver gaze fixation on the road, driver gaze analysis with a change in distance

Method	Without Calibration			With Calibration		
	Train Error	Val Error	Test Error	Train Error	Val Error	Test Error
LEye + HP(Y+P)	97	109.04	107.11	62	90.4	97.5
LEye + HP(Y+P+R)	91.08	102.77	103.78	60	89.87	95
LEye + HP + FL	86.37	99.41	99.7	38	50	58
I-DGAZE	**81.3**	**95**	**94.5**	**25**	**40**	**45**

Table 4.3: Ablation study of I-DGAZE model. Here, LEye = Left Eye, HP = Headpose, Y = Yaw, P = Pitch, R = Roll, FL = Face Landmark Location and FA = Face Area

of an object in front, etc. We also propose I-DGAZE, a late fusion convolutional neural network with eye image as input to one branch of the system and facial features as input to the second branch of the network. I-DGAZE predicts gaze on the road with an error of 94.5 pixels. We also provide user-specific results by fine-tuning the model using images collected corresponding to 9 calibration points. This work can be used to analyze driver behavior monitoring by following driver gaze on the road. The code, data, and models are available at https://github.com/duaisha/DGAZE.git.

Chapter 5

Conclusions and Future Directions

In this chapter, we present conclusions of the thesis and future directions of this work:

5.0.1 Conclusions

We propose two novel approaches for driver attention analysis on the road. One is driver attention rating to assess the quality of driver attention on the road, and the other is driver gaze mapping on the road to prevent accidents due to driver distraction while driving.

In the first part of the thesis, we present the driver attention rating, a smartphone-based system for predicting the quality of driver attention on the road. It employs several state of the art pre-trained models to extract generic features like VGGFace and specific facial features like facial landmarks, face area, head pose, eye gaze, eye-blink, and yawning. We finetune the YOLO model for phone usage detection. These features are then combined using different feature aggregation techniques like AutoRate and Attention-based AutoRate. AutoRate is a multibranch convolution neural network with generic features as input to one branch and specific features as input to the other branch. It is used to predict driver attention rating on the road. We use the kappa coefficient, an evaluation metric, to compute the inter-rater agreement. We observe that the proposed model's automatically-generated rating has an overall agreement of 0.88 with the ratings provided by 5 human annotators on the static dataset. We also show the results obtained on a model trained on the static dataset and tested on driving dataset is comparable to the result obtained by training and testing on the driving dataset. Besides, we show that the Attention-based AutoRate model outperforms the AutoRate model by 10% accuracy. We then use the learned attention probabilities to show the spatial and temporal visualization of Attention Based AutoRate mode to determine the region of inattention in videos and the key action performed that leads to this inattention. We further show personalization in the proposed model for the user-specific accuracy.

In the second part of the thesis, we present driver gaze mapping on the road using the fusion of driver and road video collected using a mobile phone mounted on the windshield of the car. To collect dataset for mapping driver gaze, we require costly eyeglass trackers and eye-gaze trackers to collect data in real road settings and on the computer screen. To this end, we propose the DGAZE dataset, which includes both driver and road view collected using mobile phone camera mounted on tripod stand mimicking the actual car driving setting. The dataset can be used for various applications like driver gaze fixation on the road, driver gaze analysis with a change in distance of the object in front, etc. We also propose I-DGAZE, a late fusion convolutional neural network with eye image as input to one branch of the network and facial features as input to the other branch of the network. I-DGAZE predicts gaze on the road with an error of 94.5 pixels. We also provide user-specific results by fine-tuning the model using collected images corresponding to 9 calibration points. This work can be used for driver attention monitoring by following the driver gaze on the road. In summary, we made the following notable contributions to the driver attention monitoring community:

1. We propose a rating system for driver attention analysis in the range of 1 to 5, where rating-1 means very careless and rating-5 means very attentive. The rating system combines all the features responsible for driver inattention on-road rather than analyzing them independently.

2. We introduce the kappa coefficient, an evaluation metric to obtain an inter-rater agreement between the subjective ratings by different annotators.

3. We introduce the DGAZE dataset for driver gaze mapping on the road. The dataset is collected in a lab setting using a mobile phone camera. It consists of both driver and road view along with gaze point on the road. Eye gaze trackers and eyeglass trackers are very expensive, and hence, collection of such dataset is challenging.

5.0.2 Future Directions

During the work for this thesis, we have identified the following future directions for extending this work:

Real-time driver attention rating: Presently, the total time taken to extract facial features is approximately 6 minutes/video on an average. The time used to predict driver rating using Attention-based AutoRate is 1 second approximately. The total time taken at run time is 6 minutes and 1 seconds. Predicting this driver attention rating on the road in real-time is the future direction of this work.

Driver attention rating using more features: Predicting driver attention on-road using the fusion of driver videos, road videos, and other sensor information is the new challenge in this direction. Currently, we use facial features extracted from driver videos to predict driver attention on the road.

Driver attention analysis by predicting driver gaze on the road: The driver gaze on the road can be used for analyzing the change in driver attention with the distance of the object in front of the car. Avoiding vehicle-pedestrian collision by predicting driver gaze on the road is one such application of this task.

Related Publications

Journal

1. I. Dua, A. U. Nambi, C. V. Jawahar and V. N. Padmanabhan, ”Evaluation and Visualization of Driver Inattention Rating from Facial Features,” in IEEE Transactions on Biometrics, Behavior, and Identity Science, 2019.

Conference

1. 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, 2019 [Oral Paper].

2. I. Dua, T. A. John, R. Gupta, C. V. Jawahar, “DGAZE: Driver Gaze Mapping on Road,”(Under review).

Other related Indian efforts

1. A. U. Nambi, S. Bannur, I. Mehta, H. Kalra, A. Virmani, V. N. Padmanabhan, R. Bhandari, and B. Raman, ”Hams: Driver and driving monitoring using a smartphone,” in Proceedings of the 24th Annual International Conference on Mobile Computing and Networking, 2018.

2. Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker and C V Jawahar , ”IDD: A Dataset for Exploring Problems of Autonomous Navigation in Unconstrained Environments,” in IEEE Winter Conference on Applications of Computer Vision (WACV 2019).

3. C. V. Jawahar and V. N. Padmanabhan, ”Technology interventions for road safety and beyond”, Communications of the ACM, 2019

Important Links:

1. Features, and source code for driver attention rating: https://github.com/duaisha/AutoRate.

2. Dataset, and source code for driver gaze mapping: https://github.com/duaisha/DGAZE.git.

Bibliography

[1] Global status report on road safety 2018. https://www.who.int/violence$_i njury_p revention/road_s afety_s tatus/2018/en/$.

[2] Honda CR-V SUV. https://venturebeat.com/2017/03/09/this-small-suv-knows-when-you-get-sleepy-and-can-wake-you-up/.

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

[4] Receive Warnings About Your Level of Alertness While Driving With Honda's Driver Attention Monitor. http://www.hiltonheadhonda.com/blog/how-does-the-honda-driver-attention-monitor-work/.

[5] B. Ahn, Y. Han, and I. S. Kweon. Real-time facial landmarks tracking using active shape model and lk optical flow. In 2012 9th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), 2012.

[6] S. Begum. Intelligent driver monitoring systems based on physiological sensor signals: A review. In 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013), 2013.

[7] L. M. Bergasa, D. Almería, J. Almazán, J. J. Yebes, and R. Arroyo. Drivesafe: An app for alerting inattentive drivers and scoring driving behaviors. In 2014 IEEE Intelligent Vehicles Symposium Proceedings, 2014.

[8] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In International Conference on Computer Vision, 2017.

[9] G. A. P. C., F. García, A. de la Escalera, and J. M. Armingol. Driver monitoring based on low-cost 3-d sensors. IEEE Transactions on Intelligent Transportation Systems, 2014.

[10] J. Chen and Q. Ji. 3d gaze estimation with a single camera without ir illumination. 2008 19th International Conference on Pattern Recognition, 2008.

[11] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. CoRR, 2014.

[12] G. G. Chrysos, E. Antonakos, S. Zafeiriou, and P. Snape. Offline deformable face tracking in arbitrary videos. 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), 2015.

[13] J. Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological bulletin, 1968.

[14] C. Craye and F. Karray. Driver distraction detection and recognition using rgb-d sensor. arXiv preprint arXiv:1502.00250, 2015.

[15] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsberg. Accurate scale estimation for robust visual tracking. In Proceedings of the British Machine Vision Conference, 2014.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR09, 2009.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR09, 2009.

[18] T. A. Dingus, S. Klauer, V. L. Neale, A. Petersen, S. E. Lee, J. Sudweeks, M. A. Pérez, J. M. Hankey, D. J. Ramsey, S. Gupta, C. Bucher, Z. R. Doerzaph, J. Jermeland, and R. R. Knipling. The 100-car naturalistic driving study phase ii - results of the 100-car field experiment. 2006.

[19] Y. Dong, Z. Hu, K. Uchimura, and N. Murayama. Driver inattention monitoring system for intelligent vehicles: A review. IEEE Transactions on Intelligent Transportation Systems, 2011.

[20] 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), 2019.

[21] A. Fridman, P. Langhans, J. Lee, and B. Reimer. Driver gaze region estimation without use of eye movement. IEEE Intelligent Systems, 2015.

[22] N. Gourier, D. Hall, and J. L. Crowley. Estimating face orientation from robust detection of salient facial features. In ICPR International Workshop on Visual Observation of Deictic Gestures, 2004.

[23] N. Gourier, D. Hall, and J. L. Crowley. Estimating face orientation from robust detection of salient facial structures. In FG NET WORKSHOP ON VISUAL OBSERVATION OF DEICTIC GESTURES, 2004.

[24] D. W. Hansen and A. E. C. Pece. Eye tracking in the wild. Computer Vision and Image Understanding, 2005.

[25] Z. He, M. Kan, J. Zhang, X. Chen, and S. Shan. A fully end-to-end cascaded cnn for facial landmark detection. In 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017), 2017.

[26] P. Hu and D. Ramanan. Finding tiny faces. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July.

[27] Q. Huang, A. Veeraraghavan, and A. Sabharwal. Tabletgaze: A dataset and baseline algorithms for unconstrained appearance-based gaze estimation in mobile tablets. ArXiv, 2015.

[28] C. V. Jawahar and V. N. Padmanabhan. Technology interventions for road safety and beyond. Communications of the ACM, 2019.

[29] J. Jo, S. J. Lee, K. R. Park, I.-J. Kim, and J. Kim. Detecting driver drowsiness using feature-level fusion and user-specific classification. Expert Systems with Applications, 2014.

[30] 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.

[31] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, 2011.

[32] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. M. Bhandarkar, W. Matusik, and A. Torralba. Eye tracking for everyone. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems 25. 2012.

[34] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In Proceedings of the IEEE, 1998.

[35] B.-G. Lee and W.-Y. Chung. A smartphone-based driver safety monitoring system using data fusion. Sensors, 2012.

[36] P. M. R. Martin Koestinger, Paul Wohlhart and H. Bischof. Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization. In Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies, 2011.

[37] C. McMurrough, V. Metsis, J. Rich, and F. Makedon. An eye tracking dataset for point of gaze detection. In ETRA, 2012.

[38] K. A. F. 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 ETRA, 2014.

[39] G. Munoz. How fast is a blink of an eye? https://sciencing.com/fast-blink-eye-5199669.html.

[40] A. U. Nambi, S. Bannur, I. Mehta, H. Kalra, A. Virmani, V. N. Padmanabhan, R. Bhandari, and B. Raman. Hams: Driver and driving monitoring using a smartphone. In Proceedings of the 24th Annual International Conference on Mobile Computing and Networking, 2018.

[41] S. Ohayon and E. Rivlin. Robust 3d head tracking using camera pose estimation. In 18th International Conference on Pattern Recognition (ICPR'06), 2006.

[42] 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.

[43] S. J. Pan and Q. Yang. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22:1345–1359, 2010.

[44] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In British Machine Vision Conference, 2015.

[45] M. Patacchiola and A. Cangelosi. Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. Pattern Recognition, 2017.

[46] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[47] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. arXiv preprint arXiv:1612.08242, 2016.

[48] M. Rezaei and R. Klette. Look at the driver, look at the road: No distraction! no accident! In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[49] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: database and results. Image Vision Comput., 2016.

[50] A. Sahayadhas, K. Sundaraj, and M. Murugappan. Detecting driver drowsiness based on sensors: A review. In Sensors, 2012.

[51] K. Seshadri, F. Juefei-Xu, D. K. Pal, M. Savvides, and C. P. Thor. Driver cell phone usage detection on strategic highway research program (shrp2) face view videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015.

[52] V. Sharmanska, D. Hernández-Lobato, J. M. Hernández-Lobato, and N. Quadrianto. Ambiguity helps: Classification with disagreements in crowdsourced annotations. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[53] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), 2015.

[54] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[55] 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.

[56] T. Song, X. Cheng, H. Li, J. Yu, S. Wang, and R. Bie. Detecting driver phone calls in a moving vehicle based on voice features. In IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications, 2016.

[57] T. Soukupová. Real-time eye blink detection using facial landmarks. 2016.

[58] Y. Sugano, Y. Matsushita, and Y. Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[59] A. M. Turing. Computing machinery and intelligence. In Parsing the Turing Test. 2009.

[60] G. Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[61] R. Valenti, N. Sebe, and T. Gevers. Combining head pose and eye location information for gaze estimation. IEEE Transactions on Image Processing, 2012.

[62] V. Vapnik and R. Izmailov. Learning using privileged information: similarity control and knowledge transfer. Journal of machine learning research, 2015.

[63] G. Varma, A. Subramanian, A. M. Namboodiri, M. K. Chandraker, and C. V. Jawahar. Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1743–1751, 2018.

[64] F. Vicente, Z. Huang, X. Xiong, F. De la Torre, W. Zhang, and D. Levi. Driver gaze tracking and eyes off the road detection system. IEEE Transactions on Intelligent Transportation Systems, 2015.

[65] A. J. Viera, J. M. Garrett, et al. Understanding interobserver agreement: the kappa statistic. Fam Med, 2005.

[66] Y. Wang, T. Zhao, X. Ding, J. Bian, and X. Fu. Head pose-free eye gaze prediction for driver attention study. 2017 IEEE International Conference on Big Data and Smart Computing (BigComp), 2017.

[67] U. Weidenbacher, G. Layher, P.-M. Strauss, and H. Neumann. A comprehensive head pose and gaze database. IET, 2007.

[68] P. Xu and K. A. Ehinger. Rich feature hierarchies for accurate object detection and semantic segmentation. 2015.

[69] J. Yang, Q. Liu, and K. Zhang. Stacked hourglass network for robust facial landmark localisation. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on, 2017.

[70] T. yi Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, and C. L. Zitnick. Microsoft coco: Common objects in context.

[71] D. H. Yoo and M. J. Chung. A novel non-intrusive eye gaze estimation using cross-ratio under large head motion. Computer Vision and Image Understanding, 2005.

[72] C.-W. You, M. Montes-de Oca, T. J. Bao, N. D. Lane, H. Lu, G. Cardone, L. Torresani, and A. T. Campbell. Carsafe: a driver safety app that detects dangerous driving behavior using dual-cameras on smartphones. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing, 2012.

[73] Y. Yun, I. Y. H. Gu, M. Bolbat, and Z. H. Khan. Video-based detection and analysis of driver distraction and inattention. 2014 International Conference on Signal Processing and Integrated Networks (SPIN), 2014.

[74] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Appearance-based gaze estimation in the wild. In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[75] Y. Zhao, L. Görne, I.-M. Yuen, D. Cao, M. Sullman, D. Auger, C. Lv, H. Wang, R. Matthias, L. Skrypchuk, et al. An orientation sensor-based head tracking system for driver behaviour monitoring. Sensors, 2017.

[76] Z. Zhu and Q. Ji. Eye gaze tracking under natural head movements. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 1:918–923 vol. 1, 2005.

[77] Z. Zhu, Q. Ji, and K. P. Bennett. Nonlinear eye gaze mapping function estimation via support vector regression. 18th International Conference on Pattern Recognition (ICPR'06), 2006.