AutoRate: How attentive is the driver?

Isha Dua[1], Akshay Uttama Nambi[2], C.V.Jawahar[1], Venkat Padmanabhan[2]
[1]International Institute of Information Technology Hyderabad, India
[2]Microsoft Research, India
isha.dua@research.iiit.ac.in, jawahar@iiit.ac.in, {t-snaksh, padmanab}@microsoft.com

***Abstract*— 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 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, in order to develop a robust driver attention monitoring system all the factors affecting driver's attention needs to be analyzed holistically. In this paper, we present `AutoRate`, a system that leverages front camera of a windshield-mounted smartphone to monitor driver's 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 perform extensive evaluation of `AutoRate` 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. Further, we compute the agreement between `AutoRate`'s rating and human annotator rating using kappa coefficient. `AutoRate`'s automatically-generated rating has an overall agreement of 0.87 with the ratings provided by 5 human annotators on the static dataset.

I. Introduction

Driver inattention is one of the leading causes for road accidents in the world. According to National Highway Traffic Safety Administration (NHTSA), 15% of crashes in the U.S. in 2015 were due to driver inattention [2]. Driver inattention occurs when the drivers divert their attention from the driving task to focus on other activity. 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. 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 such as fatigue, drowsiness and distraction, have been evaluated independently. For instance, some high end cars like Honda CR-V and Accord [1], [3] constantly monitor steering wheel input and raise alerts when the driver is frequently veering out of the lane. However, these solutions are expensive and

Fig. 1: `AutoRate` to predict driver inattention 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

are not present in all the vehicles. Hence, several camera-based ADAS systems have been designed. For instance, [44], [6] propose smartphone-based drowsiness detection based on analyzing features such as eye closure and yawn frequency. In [45], [42] various algorithms have been proposed to detect driver's gaze information to assess driver distraction, e.g., eyes off the road.

Thus far, most of the techniques proposed [44] 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 analyze all the factors affecting the driver's attentiveness. Such a system could be used to provide a quantitative rating of driver attention.

Designing 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

978-1-7281-0089-0/19/$31.00

as the system needs to identify and extract spatiotemporal information across 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 paper we propose `AutoRate`, a camera-based system to automatically determine the driver's attention rating. 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 `AutoRate` is to derive a driver's attention 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 [5] to detect inattention as sensors are intrusive. Further, 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 spatiotemporal features that affect the driver's attention. `AutoRate` is trained and tested using an extensive real-world dataset comprising over 2900 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.

Since the objective of `AutoRate` is to derive a rating that is indistinguishable from that of a human annotator, we need to obtain ratings from human annotators. 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) [39], [34], 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 evaluate `AutoRate` 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 (Section V-B). (2) **Agreement-based:** In this approach, we compute the kappa coefficient (κ) that measures inter-rater agreement between raters [41], [10]. 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 (Section V-C). (3) **Turing test based [38]:** 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 (See section V-E).

Our main contributions are, (1) We gather driver video data (both static and driving setting) that can be used for building a comprehensive driver attention system. (2) We propose a method to exploit spatial and temporal facial features from the video data to automatically rate driver attention in the range of 1 to 5, where 1 implies least attentive and 5 implies most attentive. (3) We propose a novel method for evaluating our model, so as to incorporate the subjective nature of decision making rather than avoiding the ambiguity.

II. Related work

Prevalent work on driver attention can be broadly classified into sensor-based and camera-based techniques.

Sensor-based techniques: Lee et al. [23] 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 [11], 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 [47] and [7], 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, `AutoRate` 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 [44], [6]. Dong et al. [13] present a review of various state-of-the-art techniques proposed to detect driver drowsiness, fatigue and distraction. Rezaei et al. [31] 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 [18]. 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. [40] 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. [36] describe a system to detect talking over the phone using the microphone's audio data and driver's voice features. Sheshadri et al. [33] 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, `AutoRate` 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. `AutoRate` goes beyond existing works [29] to derive a driver attention rating, 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.

III. `AutoRate` Design

We now present the design of `AutoRate` to determine driver's attention rating. The objective of `AutoRate` is to derive a rating (in the range of 1 to 5) that is equivalent to a rating provided by a human annotator.

`Rating-1` represents *inattentive* and distracted driving, e.g., talking over the phone or with other passengers for the most part 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. Note that our rating of driver attention is based on the driver behavior, and *not* on their driving. An assessment of driving would likely need additional sensing streams to detect sharp braking, jerks, honking, etc., and would be quite challenging to do (and even more subjective) if attempted based just on the driver-facing video.

We present two approaches for determining the rating from the given video, in the first approach we use pre-trained CNN to extract the generic features and then apply a GRU across the frames to get a final representation of the entire video snippet. In the second 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.

A. CNN (generic features) and GRU or (CNN + GRU)

As recent works [27] 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 [35] trained on the ImageNet [12] dataset, to extract bottleneck features. Figure 2(a) 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.

B. `AutoRate` Architecture

Figure 2(b) 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) [8] 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. We now discuss the building blocks of `AutoRate`'s architecture.

1) Feature identification and extraction: As mentioned earlier, `AutoRate` extracts two types of features, (i) generic features and (ii) specific facial features.

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 III-A above, with a pre-trained VGG16 [35] convolutional network being used to extract a low-dimensional feature representation (or bottleneck features) of the frames.

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

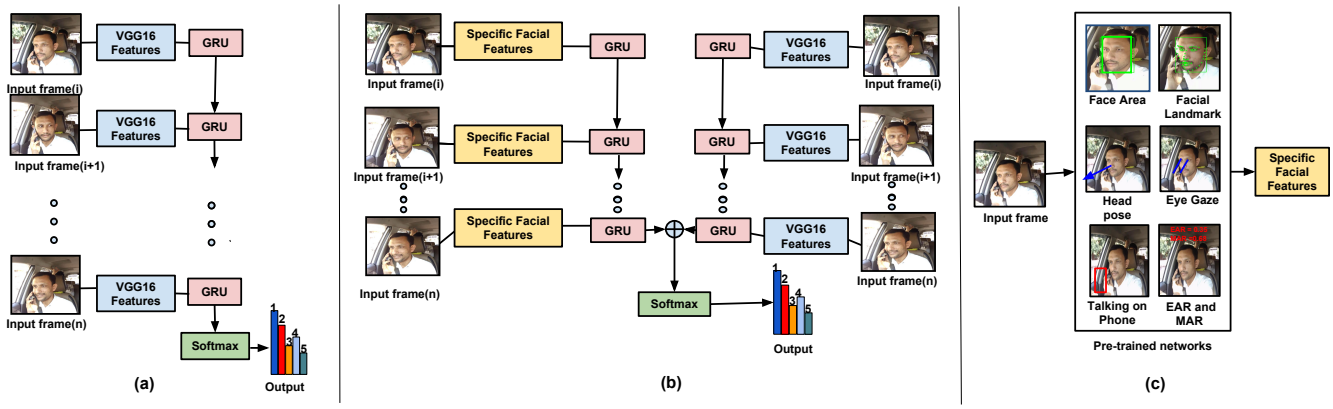

Fig. 2: Design choices. (a) CNN + GRU design (b) AutoRate's design and (c) Specific facial feature block.

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 [13]. We use state-of-the-art pre-trained models to extract these specific facial features from a sequence of frames. Figure 2(c) 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 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 [19], [4], [15]. In this work, we employ a pre-trained Face Alignment Network (FAN) to extract facial landmarks.

2. Eye closure & yawns: Several studies have identified behavioral measures such as eye closure and yawn frequency to detect drowsiness [32]. AutoRate leverages facial landmarks to detect eye closure and yawns [25]. Specifically, to detect eye closure we use the eye aspect ratio (EAR) [37] metric, which is the ratio of the height of the eye to its width.

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. 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 [26], we employ a pre-trained CNN [21] due to its robustness. The pre-trained network *viz.,* Deepgaze [28] is trained using datasets such as Prima [14], AFLW [24], and AFW [20] to handle large pose variations.

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 [25]. We employ a standard LeNet-5 [22] network that takes an eye patch as the input and outputs 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 [46] 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 [30] network trained on COCO dataset [43], 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.

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.,* Tiny Faces [17] that can deal with extreme illumination, blurring, pose variation, and occlusion.

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 employ a Gated Recurrent Unit (GRU), which is a variant of LSTM that can model long-term

Dataset	Driving		Static		Merged	
	Train	Test	Train	Test	Train	Test
Rating-1	249	65	313	66	555	138
Rating-2	68	8	126	25	191	36
Rating-3	55	13	247	57	312	60
Rating-4	91	28	308	55	409	73
Rating-5	557	103	425	98	972	211
Total	1020	217	1419	301	2439	518

TABLE I: Dataset description with train and test split.

dependencies in the data [16]. A GRU has two gates *viz.,* a reset gate r, and an update gate z. The reset gate determines how to combine the input with the previous memory, and the update gate defines how much of the previous memory to keep. In general, GRUs train faster and perform better than LSTMs when the training data is less [9]. In `AutoRate`, a GRU layer has 256 neurons and the feature vector (V_i) from each frame is fed to a GRU layer. Finally, the output layer is a softmax classifier resulting in 5-way classification corresponding to the 5 rating levels.

IV. Experimental Evaluation

In this section, we describe the real-world dataset collected and the metrics used to evaluate `AutoRate`.

A. Datasets

We considered two datasets [25]: (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.

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 1200 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 1200 snippets in all) whereas rating-1 has fewer than 100 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.

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

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 III. Table I shows the training and test split for the static dataset.

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 over 2900 videos each of 10 seconds. Table I shows the training and test split in the merged dataset.

B. Evaluation metrics

We now describe the various metrics used for evaluation.

$\mathbf{F}_1$ score: It is a measure of test's accuracy and is defined as harmonic mean of the precision and recall.

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R}, \tag{1}$$

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 arthematic mean to obtain the overall precision. The Recall (R) is computed analogously, by considering the ground truth classes instead of the predicted classes.

Kappa coefficient (κ) between two annotators [41]: It measures agreement between two annotators and defined as,

$$\begin{aligned} \kappa &= \frac{P_o - P_e}{1 - P_e}, \\ P_o &= \sum_{i}^{R}\sum_{j}^{R} w_{ij} x_{ij}, \quad P_e = \sum_{i}^{R}\sum_{j}^{R} w_{ij} m_{ij}, \end{aligned} \tag{2}$$

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 [10], 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,

$$w_d = 1 - \frac{d^2}{(R-1)^2}, \tag{3}$$

where d is the difference between ratings.

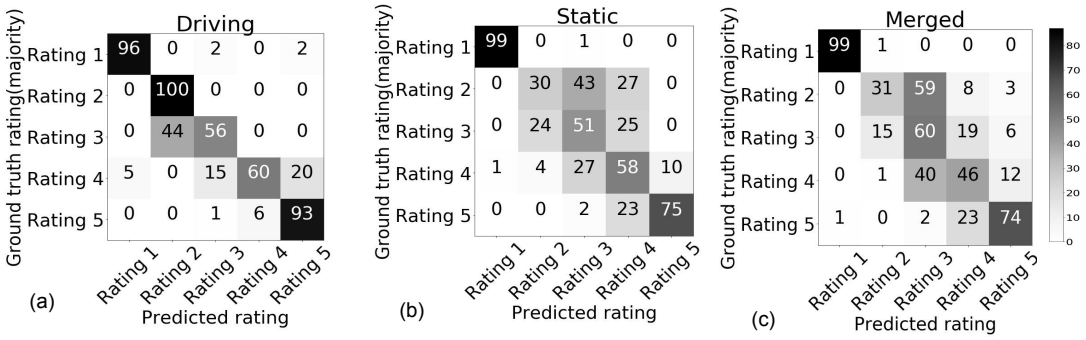

Fig. 3: Confusion matrix obtained for AutoRate, for (a) driving, (b) static and (c) merged datasets. For each ground truth rating, the row of numbers represents the percentage of predicted ratings from 1 to 5. The higher the percentage the darker the shade of the cell.

V. Results

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

A. 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 0.88 and in static dataset the agreement is 0.91. The higher kappa coefficient in the static dataset can be attributed to the unambiguous and deliberate actions (e.g., simulated inattentive driving) performed by the driver in a static vehicle. Furthermore, the kappa coefficient for the merged dataset is 0.94. 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 AutoRate using the three approaches noted in Section I, 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.

B. 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 4 shows the F1 score

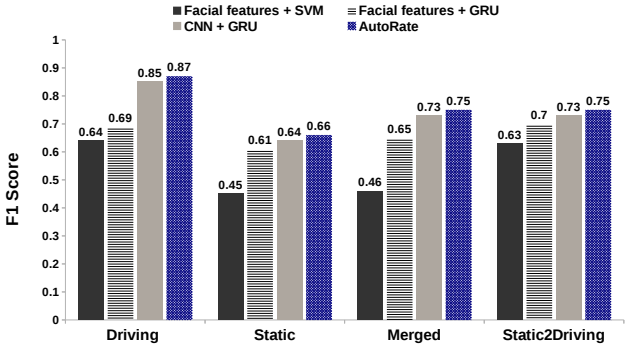

Fig. 4: F1 score for four methods for all three dataset(driving, static, merged) and static2driving i.e. trained on static and test on driving data

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 1200 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 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.

C. Agreement based evaluation

We now present evaluation based on the kappa coefficient to quantify the agreement between AutoRate's predicted rating with the individual human annotator rating. Table III shows the agreement between 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

Datasets	AutoRate vs Majority	AutoRate vs Average
Driving	0.89	0.72
Static	0.87	0.82
Merged	0.9	0.83
Stat2Driving	0.83	0.77

TABLE II: Agreement between AutoRate and Majority/Average ratings using kappa coefficient.

compute kappa coefficient between the human rating and AutoRate's rating using Equation 2.

It can be seen that for the driving dataset, AutoRate has an overall agreement of 0.89 and 0.72 with the mode and average ratings, respectively. Note that, for the same driving dataset, among the 5 annotators the agreement was 0.89. Further, AutoRate has around 0.9 agreement for mode and 0.83 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 5 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 0.89. Given that the agreement among the human annotators in driving dataset was itself low (i.e., 0.88), we conclude that AutoRate is doing quite well in mimicking a human annotator.

Method	Acc	F_1	Mode κ	Avg κ
HP + SVM	0.41	0.54	0.25	0.21
HP + EG + SVM	0.38	0.43	0.22	0.19
EB + yawning + SVM	0.33	0.36	0.12	0.11
HP + EG +EB + yawning + SVM	0.41	0.46	0.31	0.28
All facial features + SVM	0.68	0.63	0.68	0.62
All facial features + GRU	0.69	0.7	0.8	0.74
CNN + GRU	0.73	0.73	0.81	0.75
AutoRate	**0.75**	**0.75**	**0.83**	**0.77**

TABLE III: Comparison of model trained on static dataset and tested on driving dataset for various 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).

D. Ablation study

We also conducted an ablation study by combining various features used to train the model on static dataset and test on real dataset. Table III, shows the results of combination of various features such as head pose, eye gaze, eye blink and yawning. It shows that individual features are not enough for determining driver inattention accurately. Our system uses a combination of these features along with deep features to detect driver inattention, which gives better results as shown in Table III.

E. Turing test evaluation

We now report on a *Turing test* [38], where a new human *evaluator* is presented ratings from another human annotator and from AutoRate. The job of the evaluator is to tell which rating came from the human vs. from AutoRate. If the evaluator cannot reliably tell which rating came from whom, then AutoRate would have done a good job in rating driver's attention. Note that the focus here is on having

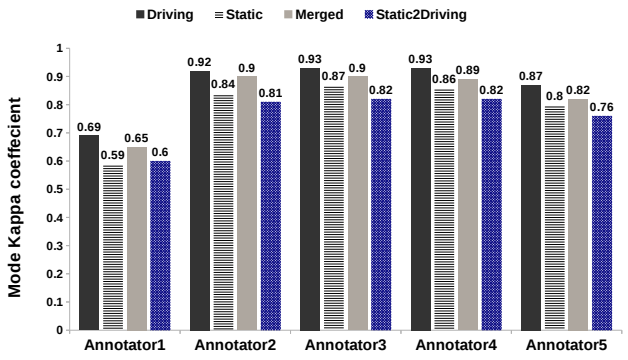

Fig. 5: Agreement between AutoRate ratings and human annotators across datasets.

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 AutoRate. AutoRate'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 AutoRate rating. Each evaluator decided which of the ratings across the 233 snippets came from a human and which from 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 AutoRate ratings. Thus, the 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.

VI. Conclusion

In this paper, we have proposed AutoRate, a smartphone-based system for driver attention rating. AutoRate employs deep learning techniques that combine generic and specific facial features towards deriving driver's attention rating. We have evaluated AutoRate on a real-world dataset with 30 drivers. AutoRate's automatically-generated rating has an overall agreement of 0.87 with the ratings provided by 5 human annotators on static dataset. In addition, 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. Our analysis shows that AutoRate's driver attention rating closely resembles a human annotator rating, thus enabling automated rating system.[2]

[2]The features and code is available at https://github.com/duaisha/AutoRate

REFERENCES

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

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

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

[4] 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)*, pages 541–543, Nov 2012.

[5] S. Begum. Intelligent driver monitoring systems based on physiological sensor signals: A review. In *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, pages 282–289, Oct 2013.

[6] L. M. Bergasa, D. Almera, J. Almazn, J. J. Yebes, and R. Arroyo. Drivesafe: An app for alerting inattentive drivers and scoring driving behaviors. In *2014 IEEE Intelligent Vehicles Symposium Proceedings*, pages 240–245, June 2014.

[7] G. A. P. C., F. Garca, A. de la Escalera, and J. M. Armingol. Driver monitoring based on low-cost 3-d sensors. *IEEE Transactions on Intelligent Transportation Systems*, 15(4):1855–1860, Aug 2014.

[8] 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*, abs/1406.1078, 2014.

[9] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

[10] J. Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

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

[12] 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.

[13] Y. Dong, Z. Hu, K. Uchimura, and N. Murayama. Driver inattention monitoring system for intelligent vehicles: A review. *IEEE Transactions on Intelligent Transportation Systems*, 12(2):596–614, June 2011.

[14] 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.

[15] 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)*, pages 200–207, May 2017.

[16] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997.

[17] P. Hu and D. Ramanan. Finding tiny faces. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[18] 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*, 41(4):1139–1152, 2014.

[19] 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)*, pages 1867–1874, 2014.

[20] 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*, pages 2144–2151. IEEE, 2011.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[22] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.

[23] B.-G. Lee and W.-Y. Chung. A smartphone-based driver safety monitoring system using data fusion. *Sensors*, 12(12):17536–17552, 2012.

[24] 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.

[25] 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*, MobiCom '18, pages 840–842, New York, NY, USA, 2018. ACM.

[26] S. Ohayon and E. Rivlin. Robust 3d head tracking using camera pose estimation. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 1, pages 1063–1066, 2006.

[27] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Trans. on Knowl. and Data Eng.*, 22(10):1345–1359, Oct. 2010.

[28] M. Patacchiola and A. Cangelosi. Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. *Pattern Recognition*, 71:132 – 143, 2017.

[29] 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)*, June 2018.

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

[31] 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*, pages 129–136, 2014.

[32] A. Sahayadhas, K. Sundaraj, and M. Murugappan. Detecting Driver Drowsiness Based on Sensors: A Review. *Sensors*, 12(12):16937–16953, 2012.

[33] 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*, pages 35–43, 2015.

[34] V. Sharmanska, D. Hernndez-Lobato, J. M. Hernndez-Lobato, and N. Quadrianto. Ambiguity helps: Classification with disagreements in crowdsourced annotations. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2194–2202, June 2016.

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

[36] 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*, pages 1–9, April 2016.

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

[38] A. M. Turing. Computing machinery and intelligence. In *Parsing the Turing Test*, pages 23–65. Springer, 2009.

[39] V. Vapnik and R. Izmailov. Learning using privileged information: similarity control and knowledge transfer. *Journal of machine learning research*, 16(2023-2049):2, 2015.

[40] 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*, 16(4):2014–2027, 2015.

[41] A. J. Viera, J. M. Garrett, et al. Understanding interobserver agreement: the kappa statistic. *Fam Med*, 37(5):360–363, 2005.

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

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

[44] 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*, pages 671–672. ACM, 2012.

[45] Y. Yun, I. Y. H. Gu, M. Bolbat, and Z. H. Khan. Video-based detection and analysis of driver distraction and inattention. In *2014 International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 190–195, Feb 2014.

[46] 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)*, pages 4511–4520, June 2015.

[47] 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*, 17(11):2692, 2017.