Evaluation and Visualization of Driver Inattention Rating From Facial Features

Isha Dua, *Member, IEEE*, Akshay Uttama Nambi, *Member, IEEE*, C. V. Jawahar, *Member, IEEE* and Venkata N. Padmanabhan, *Fellow, IEEE*

***Abstract*—In this paper, we present AUTORATE, a system that leverages the 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. 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 0.88 with the ratings provided by 5 human annotators. We also propose soft attention mechanism in AUTORATE which improves AUTORATE's accuracy by 10%. We use temporal and spatial attention to visualize the key frame and the key action which justify the model's predicted rating. Further, we observe that personalization in AUTORATE can improve driver specific results by a significant amount.**

***Index Terms*—AUTORATE, attention based AUTORATE, driver behavior analysis, visualization, personalization.**

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 [1]. 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 [2], [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 are not present in all the vehicles. Hence, several camera-based ADAS systems have been designed. For instance, [4], [5] propose smartphone-based drowsiness detection based on analyzing features such as eye closure and yawn frequency. In [6], [7] 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 [4] 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 as the system needs to identify and extract spatio-temporal 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 [8] 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

Manuscript received June 9, 2019; revised September 30, 2019 and December 13, 2019; accepted December 15, 2019. Date of publication December 27, 2019; date of current version March 30, 2020. This article was recommended for publication by Associate Editor C. Pelachaud upon evaluation of the reviewers' comments. *(Corresponding author: Isha Dua.)*

Isha Dua and C. V. Jawahar are with the Centre for Visual Information Technology, International Institute of Information Technology Hyderabad, Hyderabad 500032, India (e-mail: isha.dua@research.iiit.ac.in).

Akshay Uttama Nambi and Venkata N. Padmanabhan are with Microsoft Research, Bengaluru 560001, India.

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

Digital Object Identifier 10.1109/TBIOM.2019.2962132

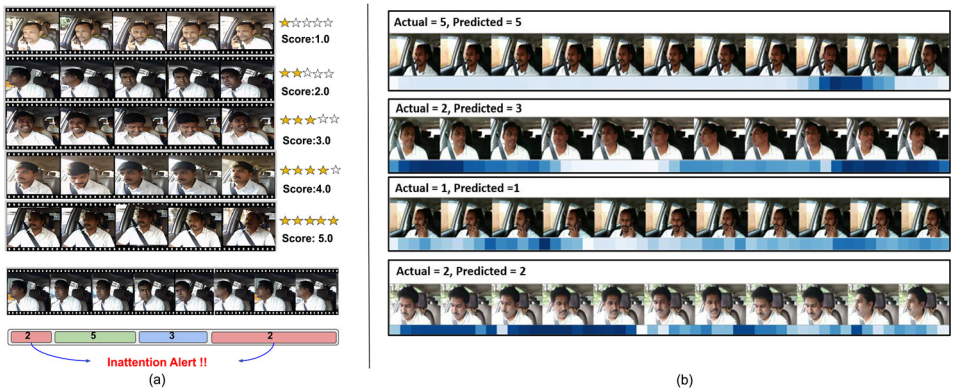


Fig. 1. AUTORATE to predict driver inattention based on specific and generic facial features. (a) The figure shows the use of AUTORATE to predict driver inattention over a long video. (b) The figure shows the visualization of attention based AUTORATE in predicting driver attention rating. Darker the color on the heat map, more is the distraction.

"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) [9], [10], 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 [11], [12]. 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 [13]: 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-F).

Further, we observe that AUTORATE fails to predict the correct rating for challenging real world videos. Challenging here means videos for which predicting specific facial features like face area, head pose, eye gaze, etc., is difficult. To address this, we propose attention-based AUTORATE architecture which learns to attend to key regions of video and ignore the rest. We then use these learned attention probabilities for spatial and temporal visualization of attention based AUTORATE model. 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. Further, 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. 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.

In this extension, we made following contributions: (1) We propose a soft attention mechanism in AUTORATE which significantly improves AUTORATE's accuracy by 10%. (2) We use temporal and spatial attention to visualize the key frames and the key actions which justify the models predicted rating. (3) We also show personalization in attention based AUTORATE model for driver specific results.

II. RELATED WORK

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

Sensor-Based Techniques: Lee and Chung [14] 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 [15], 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 [16] and [17], 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 [4], [5]. Dong *et al.* [18] present a review of various state-of-the-art techniques proposed to detect driver drowsiness, fatigue and distraction. Rezaei and Klette [19] 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 [20]. 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.* [21] 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.* [22] describe a system to detect talking over the phone using the microphone's audio data and driver's voice features. Sheshadri *et al.* [23] 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 [24] 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.

Attention Mechanism: Attention mechanism has recently succeeded in image captioning [25], neural machine translation [26], multimedia recommendation [27] and many other tasks because it can concentrate on the effective parts of features adaptively. Vinyals *et al.* [25] propose an attention based model that automatically learns to describe the content of an image by using visual attention mechanism. Bahdanau *et al.* [26] propose an approach which allows a model to automatically soft search for parts of a source sentence that are relevant to predicting a target word, without having to specify these parts like a hard segment explicitly. Bahdanau *et al.* [26] propose a two-layer attention mechanism to extract implicit feedback. The bottom layer adaptively selects the informative implicit feedbacks on component level. The upper layer adaptively selects the informative implicit feedbacks on item level. The selected implicit feedbacks are incorporated into the classic CF model with implicit feedback. In this paper, we use attention on top of LSTM encoding to learn to select the frame using learned attention probabilities. Also, we apply the attention mechanism for late fusion. Here we learn the relevance of frame using both deep features and facial features as input.

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 three approaches for determining the rating from the given video. In the first 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 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. 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. We use these learned attention probabilities and weights learned corresponding to each specific facial features like head pose, eye gaze, etc., for temporal and spatial visualization respectively.

A. CNN (Generic Features) and GRU or (CNN + GRU)

As recent works [28] 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 [29] trained on the ImageNet [30] dataset, to extract

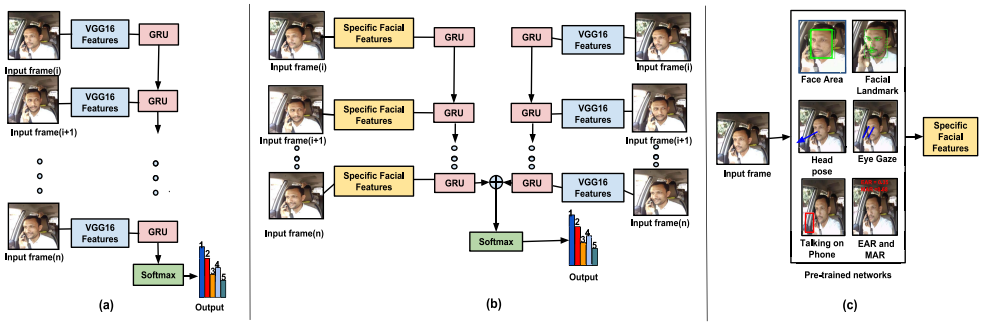


Fig. 2. Design choices. (a) CNN + GRU design (b) AUTORATE's design and (c) Specific facial feature block.

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 [31] 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 [29] 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 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 [18]. We use state-of-the-art pre-trained models to extract these specific facial features from a sequence of frames.

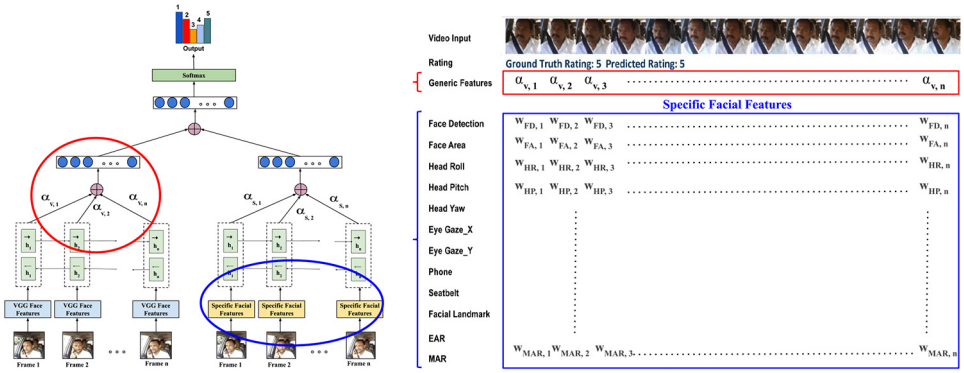
Figure 2(c) shows the facial feature extraction block for each frame. We now discuss the key facial features and describe how these are extracted from an input image:

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 to extract facial landmarks from an image [32], [33], [34]. In this work, we employ a pre-trained Face Alignment Network (FAN) [30] to extract facial landmarks.

2. *Eye Closure & Yawns:* Several studies have identified behavioral measures such as eye closure and yawn frequency to detect drowsiness [36]. AUTORATE leverages facial landmarks to detect eye closure and yawns [37]. Specifically, to detect eye closure we use the eye aspect ratio (EAR) 5 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 [38], we employ a pre-trained CNN due to its robustness. The pre-trained network *viz.,* Deepgaze [39] is trained using datasets such as Prima [40], AFLW [41], and AFW [41] 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 [37]. We use the modified LENET-5 architecture proposed in Appearance based gaze estimation in the wild [43]. 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



(a) Attention based AUTORATE model (b) Visualization of attention probabilities

Fig. 3. Attention based AUTORATE model used for temporal and spatial visualization of videos.

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.

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 [44] network trained on COCO dataset [45], 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 [46] 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 GRU, which is a variant of LSTM that can model long-term dependencies in the data [47]. 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 [48]. 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.

C. Attention Based AUTORATE

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 3(a), 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 ($\alpha_1 \cdots \alpha_n$) as shown in Equation (1).

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

where, k is S or V. S stands for specific features in right branch of Figure 3(a), V stands for VGGFace features in left branch of Figure 3(a), 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.

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 1 in Section III-C, 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.

D. 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. Figure 3(a) 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 ($\alpha_{V,1} \cdots \alpha_{V,n}$) learned corresponding to each frame from generic feature (VGGFace feature) branch of attention based AUTORATE model. The attention probability for temporal visualization is marked with red circle in Figure 3(a) and red box with generic feature label in Figure 3(b). 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(a) and blue box under the label specific facial feature block in Figure 3(b). Each row under the label specific facial feature block in Fig. 3(b) 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 ($w_{HR,1} \cdots w_{HR,50}$), 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 seatblet respectively.

E. Personalization

As driver attention rating is a very subjective problem and features (head pose, eye gaze, face area, etc.) responsible for predicting this rating may vary from one driver to another. This means the head pose value for two drivers looking at the same object may vary. To this end, we finetune attention based AUTORATE model for a specific user to provide personalized driver attention rating. For this, we first cluster the dataset for each of the 30 drivers by computing the Euclidean distance between the VGGFace features of the first frame from two videos. If the difference is less than the threshold value δ (In our experiments, we fix δ as 0.40), a video is assigned to the same driver. We then select the drivers with a nearly balanced dataset and finetune the attention based AUTORATE model to provide driver specific attention rating.

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 [37]: (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.

TABLE I
DATASET DESCRIPTION WITH TRAIN AND TEST SPLIT

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

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 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.

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 approximately 3200 videos each of 10 seconds. Table I shows the training and test split in the merged dataset.

B. Implementation Details

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 [49] 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

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

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" [46]. For facial landmark detection, we used FAN(Face Alignment Network) [30] 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" [39] 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" [43]. 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 [44] network trained previously on COCO dataset [45]. 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 II 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.

TABLE II
DETAILED DESCRIPTION OF THE FEATURE EXTRACTION

Feature	Description	Output Type	Dimension
Face Detection [46]	Face detected or not	Binary (0 or 1)	1
Face Area [46]	X,Y location on frame of size 500 x 500	0 to 500	4
Facial Landmark [35]	X,Y location on frame of size 500 x 500	0 to 500	128
Head Pose [39]	Yaw, Pitch and Roll	-90° to +90°	3
Eye Gaze [43]	Gaze in X and Y direction	Real Values	2
EAR [50]	Ratio of height to width of eye	Real Values	1
MAR [50]	Ratio of height to width of mouth	Real Values	1
Phone Detection [44]	Using phone or not	Binary(0 or 1)	1
Seat Belt Detection [44]	Wearing seatbelt or not	Binary(0 or 1)	1

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.

C. Evaluation Metrics

We now describe the various metrics used for evaluation.

F_1 score: It is a measure of test's accuracy and is defined as harmonic mean of the precision and recall.

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

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 [11]: It measures agreement between two annotators and defined as,

$$\kappa = \frac{P_o - P_e}{1 - P_e},$$
$$P_o = \sum_{i}^{R}\sum_{j}^{R} w_{ij}x_{ij}, \quad P_e = \sum_{i}^{R}\sum_{j}^{R} w_{ij}m_{ij}, \tag{3}$$

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 [12], 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,

$$w_d = 1 - \frac{d^2}{(R-1)^2}, \tag{4}$$

where d is the difference between ratings.

V. RESULTS

We now present our evaluation of the CNN + GRU architecture and of the AUTORATE for driver attention rating. We

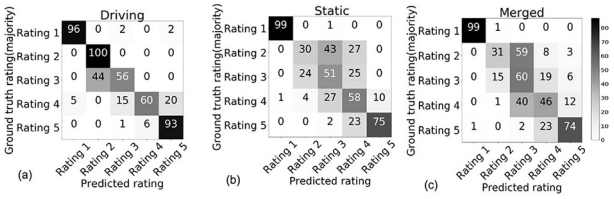


Fig. 4. Confusion matrix obtained for Attention Based AutoRate for (a) driving, (b) static and (c) merged datasets. For each ground truth rating, the row of numbers represents the percentage of predicted ratings from 1 to 5. The higher the percentage the darker the shade of the cell.

also show the efficacy of our model on datasets captured under various conditions. We further demonstrate the visualization of attention based AUTORATE model and its comparison with AUTORATE on 3200 videos. We also show the evaluation of personalization in attention based AUTORATE model.

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 0.90 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 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 5 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 4 shows the confusion matrix for AUTORATE, where each cell of confusion matrix shows the percentage of predicted rating.

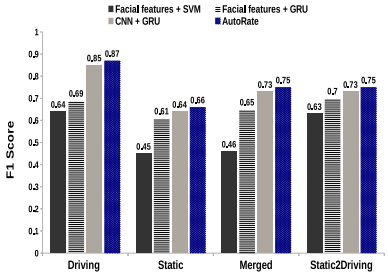


Fig. 5. F1 score for four methods for all three dataset(driving, static, merged) and static2driving.

TABLE III
AGREEMENT BETWEEN ATTENTION BASED AUTORATE AND MAJORITY/AVERAGE RATINGS USING KAPPA COEFFICIENT

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

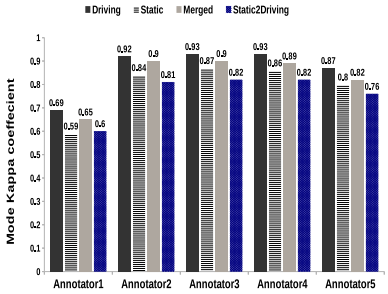


Fig. 6. Agreement between AUTORATE ratings and human annotators across datasets.

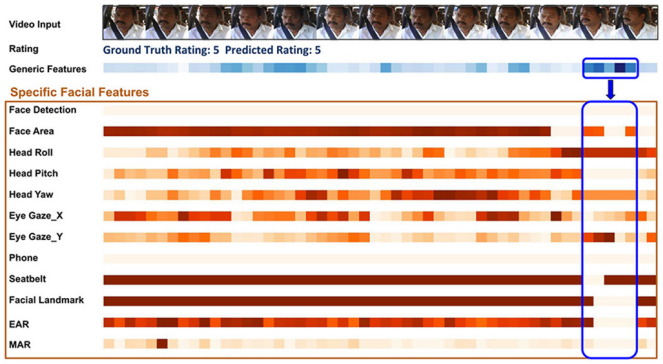
The off-diagonal values are high for adjacent rating levels indicating the ambiguity in the ground truth which results in majority of misclassifications.

C. Agreement Based Evaluation

We now present evaluation based on the kappa coefficient to quantify the agreement between 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).

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 6 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



(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.

Fig. 7. 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.

conclude that Attention Based AUTORATE is doing quite well in mimicking a human annotator.

D. Attention Based AUTORATE

In this section, we show the visualization results on the attention based AUTORATE model followed by personalization and then a rigorous ablation study using different feature combinations.

1) Visualization: Figure 7 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 7. 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 7 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.

2) Personalization: Personalization in attention based AUTORATE can be used to improve driver specific results. Table VI 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.

TABLE IV
COMPARISON OF ATTENTION BASED AUTORATE MODEL WITH AUTORATE [51] 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
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 [51]	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 V
PERFORMANCE OF ATTENTION BASED AUTORATE MODEL AS A RESULT OF STRIPPING INPUT FEATURES ONE BY ONE

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 VI
EVALUATION OF PERSONALIZATION ON 6 RANDOM DRIVERS FROM DATASET

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

E. Ablation Study

We performed extensive ablation studies on our Attention-based AutoRate model which we present in Table IV. 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 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 IV 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 V 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.

F. Turing Test Evaluation

We now report on a *Turing test* [13], 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.

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 0.88 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.

REFERENCES

[1] *NHTSA Distracted Driving*. Accessed: Sep. 25, 2019. [Online]. Available: https://www.nhtsa.gov/risky-driving/distracted-driving

[2] *Honda CR-V SUV*. Accessed: Sep. 27, 2019. [Online]. Available: https://venturebeat.com/2017/03/09/this-small-suv-knows-when-you-get-sleepy-and-can-wake-you-up/

[3] *Receive Warnings About Your Level of Alertness While Driving With Honda's Driver Attention Monitor*. Accessed: Sep. 25, 2019. [Online]. Available: http://www.hiltonheadhonda.com/blog/how-does-the-honda-driver-attention-monitor-work/

[4] C.-W. You *et al.*, "CarSafe: A driver safety app that detects dangerous driving behavior using dual-cameras on smartphones," in *Proc. ACM Conf. Ubiquitous Comput.*, 2012, pp. 671–672.

[5] L. M. Bergasa, D. Almeria, J. Almazán, J. J. Yebes, and R. Arroyo, "DriveSafe: An app for alerting inattentive drivers and scoring driving behaviors," in *Proc. IEEE Intell. Veh. Symp.*, Jun. 2014, pp. 240–245.

[6] Y. Yun, I. Y. H. Gu, M. Bolbat, and Z. H. Khan, "Video-based detection and analysis of driver distraction and inattention," in *Proc. Int. Conf. Signal Process. Integr. Netw. (SPIN)*, Feb. 2014, pp. 190–195.

[7] Y. Wang, T. Zhao, X. Ding, J. Bian, and X. Fu, "Head pose-free eye gaze prediction for driver attention study," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Feb. 2017, pp. 42–46.

[8] S. Begum, "Intelligent driver monitoring systems based on physiological sensor signals: A review," in *Proc. 16th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2013, pp. 282–289.

[9] V. Vapnik and R. Izmailov, "Learning using privileged information: Similarity control and knowledge transfer," *J. Mach. Learn. Res.*, vol. 16, no. 2, pp. 2023–2049, 2015.

[10] V. Sharmanska, D. Hernández-Lobato, J. M. Hernández-Lobato, and N. Quadrianto, "Ambiguity helps: Classification with disagreements in crowdsourced annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2194–2202.

[11] A. J. Viera and J. M. Garrett, "Understanding interobserver agreement: The kappa statistic," *Family Med.*, vol. 37, no. 5, pp. 360–363, 2005.

[12] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit," *Psychol. Bull.*, vol. 70, no. 4, p. 213, 1968.

[13] A. M. Turing, "I.—Computing machinery and intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, Oct. 1950. [Online]. Available: https://doi.org/10.1093/mind/LIX.236.433

[14] B.-G. Lee and W.-Y. Chung, "A smartphone-based driver safety monitoring system using data fusion," *Sensors*, vol. 12, no. 12, pp. 17536–17552, 2012.

[15] C. Craye and F. Karray, "Driver distraction detection and recognition using RGB-D sensor," 2015. [Online]. Available: arXiv:1502.00250.

[16] Y. Zhao *et al.*, "An orientation sensor-based head tracking system for driver behavior monitoring," *Sensors*, vol. 17, no. 11, p. 2692, 2017.

[17] G. A. C. Peláez, F. García, A. de la Escalera, and J. M. Armingol, "Driver monitoring based on low-cost 3-D sensors," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 4, pp. 1855–1860, Aug. 2014.

[18] Y. Dong, Z. Hu, K. Uchimura, and N. Murayama, "Driver inattention monitoring system for intelligent vehicles: A review," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 2, pp. 596–614, Jun. 2011.

[19] M. Rezaei and R. Klette, "Look at the driver, look at the road: No distraction! No accident!" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 129–136.

[20] 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 Syst. Appl.*, vol. 41, no. 4, pp. 1139–1152, 2014.

[21] 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 Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 2014–2027, Aug. 2015.

[22] 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 *Proc. 35th Annu. IEEE Int. Conf. Comput. Commun. (IEEE INFOCOM)*, Apr. 2016, pp. 1–9.

[23] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 35–43.

[24] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko, "Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7699–7707.

[25] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 3156–3164.

[26] D. Bahdanau, K. Cho, and Y. Bengio. (May 2016). *Neural Machine Translation by Jointly Learning to Align and Translate*. [Online]. Available: http://arxiv.org/abs/1409.0473

[27] J. Chen, H. Zhang, X. He, L. Nie, W. Liu, and T.-S. Chua, "Attentive collaborative filtering: Multimedia recommendation with item- and component-level attention," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval (SIGIR)*, 2017, pp. 335–344. [Online]. Available: http://doi.acm.org/10.1145/3077136.3080797

[28] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2009.191

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

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, 2009, pp. 248–255.

[31] K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, 2014.

[32] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 1867–1874.

[33] B. Ahn, Y. Han, and I. S. Kweon, "Real-time facial landmarks tracking using active shape model and LK optical flow," in *Proc. 9th Int. Conf. Ubiquitous Robots Ambient Intell. (URAI)*, Nov. 2012, pp. 541–543.

[34] Z. He, M. Kan, J. Zhang, X. Chen, and S. Shan, "A fully end-to-end cascaded CNN for facial landmark detection," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 200–207

[35] 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 *Proc. Int. Conf. Comput. Vis.*, 2017.

[36] A. Sahayadhas, K. Sundaraj, and M. Murugappan, "Detecting driver drowsiness based on sensors: A review," *Sensors*, vol. 12, no. 12, pp. 16937–16953, 2012. [Online]. Available: http://www.mdpi.com/1424–8220/12/12/16937

[37] S. N. A. U. Nambi *et al.*, "Hams: Driver and driving monitoring using a smartphone," in *Proc. 24th Annu. Int. Conf. Mobile Comput. Netw. (MobiCom)*, 2018, pp. 840–842. [Online]. Available: http://doi.acm.org/10.1145/3241539.3267723

[38] S. Ohayon and E. Rivlin, "Robust 3D head tracking using camera pose estimation," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, vol. 1, 2006, pp. 1063–1066.

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

[40] N. Gourier, D. Hall, and J. L. Crowley, "Estimating face orientation from robust detection of salient facial features," in *Proc. ICPR Int. Workshop Vis. Observation Deictic Gestures*, 2004.

[41] M. Köstinger, 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 *Proc. 1st IEEE Int. Workshop Benchmarking Facial Image Anal. Technol.*, 2011, pp. 2144–2151.

[42] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2879–2886.

[43] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4511–4520.

[44] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 6517–6525.

[45] T. Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. ECCV*, 2014.

[46] P. Hu and D. Ramanan, "Finding tiny faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1522–1530.

[47] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735

[48] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. NIPS Workshop Deep Learn.*, Dec. 2014.

[49] G. Munoz. *How Fast Is a Blink of an Eye?* Accessed: Sep. 25, 2019. [Online]. Available: https://sciencing.com/fast-blink-eye-5199669.html

[50] T. Soukupová and J. Cech "Real-time eye blink detection using facial landmarks," 2016.

[51] I. Dua, A. U. Nambi, C. V. Jawahar, and V. N. Padmanabhan, "AutoRate: How attentive is the driver?" in *Proc. IEEE Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–8.