A Neural Network-Based Driver Gaze Classification System with Vehicle Signals

Simone Dari
Research and Development
BMW Goup
Munich, Germany
simone.dari@bmw.de

Nikolay Kadrileev
Department of Electrical Engineering
Technical University Munich
Munich, Germany
kadrileev@yandex.ru

Eyke Hüllermeier
Department of Computer Science
Paderborn University
Paderborn, Germany
eyke@upb.de

***Abstract*—Driver monitoring can play an essential part in avoiding accidents by warning the driver and shifting the driver's attention to the traffic scenery in time during critical situations. This may apply for the different levels of automated driving, for take-over requests as well as for driving in manual mode. A great proxy for this purpose has always been the driver's gazing direction. The aim of this work is to introduce a robust gaze detection system. In this regard, we make several contributions that are novel in the area of gaze detection systems. In particular, we propose a deep learning approach to predict gaze regions, which is based on informative features such as eye landmarks and head pose angles of the driver. Moreover, we introduce different post-processing techniques that improve the accuracy by exploiting temporal information from videos and the availability of other vehicle signals. Last but not least, we confirm our method with a leave-one-driver-out cross-validation. Unlike previous studies, we do not use gazes to predict maneuver changes, but we consider the human-computer-interaction aspect and use vehicle signals to improve the performance of the estimation. The proposed system is able to achieve an accuracy of 92.3% outperforming earlier landmark-based gaze estimators.**

***Index Terms*—Driver Monitoring, Driver Gaze Estimation, Driver State Recognition, Driver distraction, Autonomous Driving, Naturalistic Driving Study**

I. INTRODUCTION

Automated driving has a great potential for not only preventing accidents, but also giving the driver the possibility to choose whether to drive by oneself or by the vehicle. For this purpose, driver monitoring is needed. As the National Highway Traffic Safety Administration states, accident risk correlates with the driver's visual attention. 80% of all crashes and 65% of all near-crashes included the driver looking away from the street [1]. Therefore, driver monitoring incorporated into current advanced driver assistant systems (ADAS) can potentially warn the driver in critical situations by shifting the driver's visual focus towards the actual driving task. Hence, foreseeing that the driver is not aware of an approaching risk may allow new ADAS for even more possibilities. The efficiency of such systems is also confirmed in a study by the Boston Consulting Group which revealed that current ADAS functions have the potential to avoid 30% of all crashes [2].

With recent developments in computer vision, research has progressed in presenting driver gaze estimators. While much research relied on using the head pose as a proxy for gazes, a lot of studies used Convolutional Neural Networks (CNNs) to directly map a gaze region to an inserted image. For this purpose a lot of training, computational power and a broad dataset are needed to learn all the important features from an image. Also, when conducting further research on other visual aspects such as talking or fatigue, new models have to be trained again.

Another aspect regarding the power of the proposed system lies in the number of participants in the study. Often only a small number of persons appear in the training dataset and data are collected within a short period of test time, e.g. one hour drives. Hence, *cross-driver*-testing becomes important for generalisability. But how much can we trust a system that e.g. was trained on 6 people and worked well on 3 others? Does another train-test split yield different results? These questions are typically not answered by cross-driver-testing but by *cross-validation*, a statistical tool for model validation and generalisability.

In this work, we circumvent the aforementioned disadvantages by making use of a feature-based approach. We extract eye landmarks and head pose angles of the driver in order to train a neural network that outputs the gaze region. To analyze the effectiveness of our system we perform a leave-*one-driver*-out cross-validation for every driver. Also, we do not just evaluate our model on a set of images, but also on videos of new drivers and take advantage of the underlying temporal information as well as the available vehicle signals. Data were taken from a naturalistic driving study and cover over 20 drivers from 80 different video sequences.

The main contributions of this work are: (1) A driver gaze detection system based on deep learning, using inputs from head pose angles and landmarks, which performs better than previous landmark-based detection systems and competes with current CNN-based approaches, (2) a leave-one-driver-out cross-validation of the system displaying its power for generalisability, and (3) several post-processing techniques that improve the accuracy of the system by taking temporal dependency into account, as well as vehicle signals that indicate where the driver is looking at. To the best of our knowledge, previous studies have not successfully examined their systems with cross-validation, nor have they examined the power of vehicle signals to enhance the proposed system.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

TABLE I
OVERVIEW OF RELEVANT STUDIES AND METHODS

Authors	Type*	Input	Classifiers	Dataset (Drivers)	RoIs	Best Result	Remarks
Choi et al. (2016) [3]	A	Face	modified AlexNet	35,900 (4)	8	95.0%	no cross-driver testing
Fridman et al. (2016) [4]	F	Facial Landmarks Geometrical Approach	Random Forest SVM	1,351,864 (40)	6	94.6%	only confident estimates (7.1% of all images)
Naqvi et al. (2018) [5]	A	Face, both eyes	VGGface Score fusion	6,518 (20)	16	96.3%	neighbouring regions included to count into accuracy
Vora et al. (2018) [6]	A	Face Face and context Upper half of face Whole image	AlexNet VGG16 SqueezeNet ResNet	47,515 (11)	7	95.2 %	included class *eyes closed* cross-driver-testing no cross-validation

*where A = Appearance-based model. F = Feature-based model.

The paper is organized as follows. Section 2 summarizes the results from studies related to this work. Section 3 describes the dataset that is used, and Section 4 introduces the gaze detection system with all its elements. Section 5 gives an overview of the results and Section 6 discusses the detection system. Section 7 concludes the paper with a summary and an outlook on future research.

II. RELATED WORK

In the past a high number of studies was conducted to develop accurate gaze classification systems from images and videos. Experiments were either conducted in driving simulators [7], [8] or in a real-world environment [9]. Eye-tracking glasses were used as well [10].

However, in order to develop a robust and reliable system, data from naturalistic driving studies are needed. Over the past few years two different vision-based approaches have evolved for gaze classification, which are shortly discussed here: there are appearance-based and feature-based approaches. A more detailed overview is given in [11].

Appearance-based models make use of image intensities and usually feed an image directly into classifier, which is a Convolutional Neural Network (CNN). Its output is then gaze region. Popular classifiers are VGG16 or ResNet. Even though they also work with lower resolutions, appearance-based models require a lot of training data with a larger number of subjects. The trained system can then only be applied on faces from the same camera angle. Also, one cannot fully control from what the classifier is learning exactly.

In feature-based approaches certain features are derived in a pre-processing step, either through appearance-based models such as facial landmarks or through geometrical approaches where geometrical properties of the eyes are exploited. These features are then summarized into a feature vector and fed into a classifier (e.g. Support Vector Machines).

The most promising studies and results were mostly derived from appearance-based approaches. An overview is given in Table I.

A. Relevant Studies

The first appearance-based gaze classification system was introduced by Choi et al. in 2016 [3]. They used a modified AlexNet and achieved an accuracy of 95 %. The highest error rate was found in the classes *inner mirror* with *front*. When looking into the inner mirror without moving the head the classifier fails as the small eyes of the drivers make it difficult to distinguish properly between these two classes. The camera was attached to the front windshield in front of the driver.

In 2018, Vora et al. [6] compared different CNNs and different kinds of input images to these CNNs. They used ResNet, VGG16, AlexNet and SqueezNet architectures with the pre-trained weights from ImageNet [12]. As inputs they either used the upper part of the face, the whole face, the whole face with more background information and the whole image containing the driver face. The study covered 47,515 test images from 11 drivers driving 30 to 60 minutes. The best result for all CNN architectures was achieved using the upper half of the face. The camera was attached behind the rearview mirror.

Naqvi et al. [5] considered a total number of 17 regions of interest (RoIs) instead of the usual 6 to 9 different RoIs. The authors used three VGG16-face CNNs, which they simultaneously applied to both eyes and to the whole face. They used a score-level fusion from all these three CNNs in order to predict the right RoI. Their gaze classifier achieved an accuracy of 96.3%. However, they summarized neighbouring regions together for that. A model-based approach was introduced as well, but its accuracy was slightly lower. The camera was mounted near the speedometer in front of the driver.

Fridman et al. [4] compared a head pose estimator to a gaze classifier. They also extracted information from the pupil and the iris geometrically and fed those features into a random forest classifier. They reported both, a general model and a leave-one-driver-out cross-validation. For the cross-validation they achieved an accuracy of 65.0%. After taking only confident estimations into account they reached an accuracy of 94.6% which corresponds to 2.1 frames instead

Fig. 1. Example images from the dataset with the same driver in different vehicles. The two left images show the same gaze region, but from different head poses. Images c), d) and e) show that for the same driver different initial head poses are making calibration necessary.

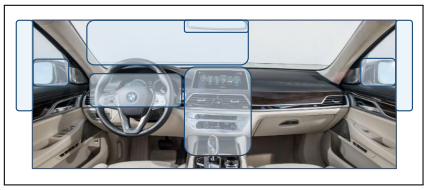

Fig. 2. The approximate regions of interest considered in the present study: *front*, *left mirror*, *left shoulder*, *right mirror*, *right shoulder*, *center console* and *speedometer*.

of 30 frames per second. Their camera was attached to the dashboard of the vehicle in a little off-axis position.

Other feature-based approaches have been used earlier, but the quality of the extracted features was rather poor as both head poses and pupils were roughly estimated from a few landmarks [13]. Before, it was also common to derive gaze region from the head pose [8], [14] [15]. However, results have shown that especially when the driver was only looking from the corners of the eyes instead of turning the head, classification accuracy was decreasing, i.e. for the center stack region and the inner mirror region the accuracies were lower. Only the authors in [4] directly address this complexity in their work. Example images are given in Fig. 1 a-c).

Cross-driver testing has been applied just recently. Drivers in the training set must not appear in the testing set, which occurred in [3]. Both the authors in [6] and [5] also applied the method from [3] onto their datasets with less success. The authors from [5] and [6] used both cross-driver testing. However, only the authors in [4] also applied a leave-one-driver-out cross-validation to their dataset fully unveiling that their results do not stem from a random choice of drivers in the training and test set.

III. DATASET

The dataset considered in this work was collected within a naturalistic driving study. Subjects were driving in sensor-equipped vehicles for several months with current ADAS. By doing so, naturalistic driving is guaranteed.

An RGB-camera was attached to the A-pillar recording the driver in an off-axis position. Fig. 1 shows example images from the dataset. There are two datasets, one for training the classifier and another one for validating the post-processing techniques. In total, there were 75 video snippets from 20 subjects (5 female, 15 males) in the training dataset and 5 more video snippets with 5 other subjects (3 female, 2 male) in the second dataset. All subjects were aged between 18 and 60 years. Driver videos were of size 980×540 with 15 frames per second. The videos displayed snippets from cruises on the highway and on the country road with the car always moving. The drivers did not wear any glasses.

A. Annotation

Annotation was performed by two annotators. The first annotator labelled the videos while the second annotator controlled the labelled images per class and checked for outliers.

TABLE II
OVERVIEW OF CLASS DISTRIBUTION IN THE TRAINING DATA SET

	front	**lm**	**im**	**cc**	**sp**	**rm**	**ls**	**rs**
Set 1	2,296	967	713	332	228	356	98	72
in %	45.4	19.1	14.0	6.5	4.5	7.0	1.9	1.4
weight	1	1	2	3	2	2	2	2
Set 2	5,870	253	138	2,758	796	120	66	5
in %	58.69	2.52	1.37	27.57	7.96	11.9	0.65	0.005

B. Regions of Interest

Similarly to previous studies, we consider the following eight gazes regions that we also refer to as gazes, RoIs or classes: *front*, *left mirror (lm)*, *left shoulder (ls)*, *right mirror (rm)*, *right shoulder (rs)*, *center console (cc)* and *speedometer (sp)*. The RoIs are depicted in Fig. 2. While [6] and [4] summarized the gazes to the outside as *left* and *right* we stick to the above classes. After training they can be aggregated. The distribution of all classes is shown in Table II under Set 1 (for training) and Set 2 (for validation). In total, there are 15,068 images. Since the driver spends most time looking into the front direction [15], attention was payed to not label all gazes from the *front* class.

IV. METHOD

In order to evaluate videos, we train a classifier and later use vehicle signals and the underlying temporal information to improve the accuracy of the system.

The pipeline for the classifier is presented in Fig. 3. It consists of the following parts: a) face detection, b) estimation of head pose angles, c) estimation of facial and eye landmarks, d) calculation of confidence values and e) the network architecture. Afterwards, different post-processing techniques are applied. At first, a face detector is used for all images. We use the Single Shot MultiBox Detector introduced by [16] in 2016. All other steps are explained in the following parts.

A. Head Pose Angles and Calibration

Euler angles are used to describe the rotations of the head in a 3-dimensional coordinate system. In the head coordinate system as depicted in Fig. 4 (left), the rotation around the x-axis is called *pitch* (i.e. from up to down), around the y-axis *yaw* (i.e. from left to right), and around the z-axis *roll* (i.e. from left to right shoulder). We use the method described in [17] in which for a cropped face angles are returned as output.

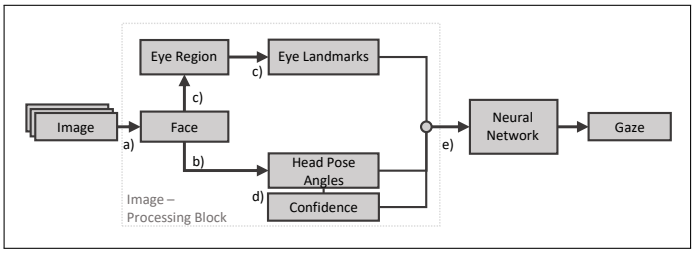

Fig. 3. An overview of the gaze detection system with its pre-processing steps is displayed.

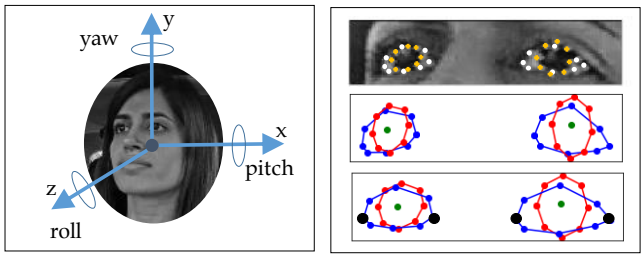

Fig. 4. Left: The head pose angles are displayed. Right: For an RGB-image (top) the eye landmarks are computed. After rotation and normalisation (bottom), the black landmarks are the same for all instances and left out in the input feature vector.

As one notices in Fig. 1 c), d) and e), all drivers are looking to the front direction. However, the position of the driver towards the camera is slightly different. This difference can even become larger for different drivers and different seating positions. Therefore, a calibration is needed first. The most occurring head pose angles from all head pose angles in a video is computed, as this typically belongs to the every driver's initial pose. For our classifier, we only use the yaw and pitch values. From the initial angles α_{yaw}, α_{pitch}, we compute the calibration shift α_{yaw}^{calib}, α_{pitch}^{calib} per angle by finding the angles that allocate most of the mass in the interval of $[15° \pm 10°]$ and subtract them from all the other angles. By dividing by 90°, the newly obtained angles $\tilde{\alpha}_{yaw}$, $\tilde{\alpha}_{pitch}$ are in $[-1,1]$. Larger values may occur, however, the face detector usually fails in finding a face for such extreme positions.

B. Facial and Eye Landmarks

Both facial and eye landmarks are used. The facial landmark method from [18] is a CNN-based classifier that outputs 68 coordinates for a given face. In total, there are 7 coordinates per eye. As this number is too small, we use these landmarks to detect a bounding box for the eye regions and feed this box again into a new detection system [19]. This detector outputs 8 landmarks for the eyelid, 8 landmarks on the iris edge, one landmark for the iris center and another one for the eyeball center. We use all of them except for the eyeball center as input features for the classifier. During pre-processing, landmarks are centered at the inner eye corner, both x and y-coordinates of the landmarks are normalized by the eye width and rotated such that both eye corners are aligned horizontally to the x-axis. Then, the landmarks from the corners of the eye are removed as they are the same for all input. The landmarks before and after rotation are shown in Fig. 4. In total we use 15 landmarks, i.e. 30 coordinates per eye as an input to the classifier.

C. Confidence Metric and Neural Network

Due to the off-axis position in the setup, parts of the driver's face become occluded while turning to the right side. However, the landmarks for the occluded eye are still computed. In order to circumvent training with these wrong values, we introduce a confidence metric into the model, which depends on the yaw angle α_{yaw}. The confidences ν_{left} and ν_{right} for the left and the right eye are defined as follows:

$$\nu_{left} = \begin{cases} \max(0, 1 - s_1 \alpha_{yaw}^2) & \text{if } \alpha_{yaw} \geq 0 \\ \max(0, 1 - s_2 \alpha_{yaw}^2) & \text{if } \alpha_{yaw} < 0 \end{cases}, \quad (1)$$

$$\nu_{right} = \begin{cases} \max(0, 1 - s_1 \alpha_{yaw}^2) & \text{if } \alpha_{yaw} < 0 \\ \max(0, 1 - s_2 \alpha_{yaw}^2) & \text{if } \alpha_{yaw} \geq 0 \end{cases}. \quad (2)$$

For our calculations we choose $s_1 = 0.2$ and $s_1 = 0.8$.

All inputs, i.e. the eye landmarks, the head pose angles and the confidence metrics, are fed into the neural network architecture. The left eye, the right eye and the head pose angles are processed separately in two fully-connected layers with a Sigmoid activation function. Then the intermediate outputs of both eyes are multiplied with the corresponding yaw confidence parameters ν_{left} and ν_{right}. All processed features from eyes and head pose are stacked into one vector and fed into two more fully-connected layers followed by the softmax activation function that produces a vector with probabilities for each class. In Fig. 5 the described architecture is shown.

As the classes are heavily imbalanced, we choose a weighted cross-entropy loss function $\mathscr{L}$ with the weights w_i for the class $i \in \{1,\ldots,8\}$ as follows:

$$\mathscr{L}(\mathbf{W}) = -w_i \log \mathbf{y_i}(\mathbf{x}, \mathbf{W})$$

where $\mathbf{y_i}(\mathbf{x}, \mathbf{W})$ is the i-th output of the softmax-layer, $\mathbf{W}$ specifies the weights of the neural network [20]. The weights w_i which yielded the best results are displayed in Table II.

D. Post-processing Method

In order to test the classifier not just on images but videos, two post-processing tools are introduced, which make use of the temporal information and the vehicle signals. For a video consisting of m frames, we obtain an $(8 \times m)$-matrix from the output probabilities of the classifier.

Signal Filter: As vehicle signals are also available, we are interested in using these in order to improve the accuracies of the system. Signals related to the left or right side might be related to the blinkers. Signals correlated with the gaze *center console* are signals from the center stack region.

We assume that the driver is looking into that direction a few moments before and a few moments after. Let us assume signal s correlates with a gaze region. Let signal $s \in \{0, 0.5, 1\}$ be discrete. Then we define a kernel of size k that is a linearly spaced symmetric vector, and its samples are evenly drawn

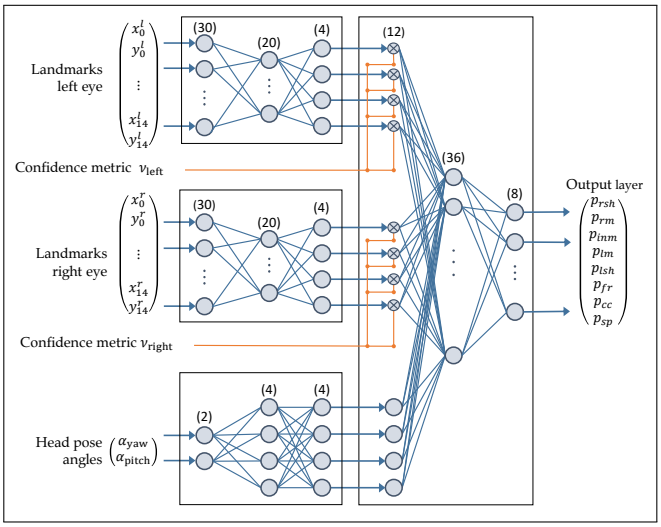

Fig. 5. The classifier used in the detection system is a fully connected neural network. At first there are three fully connected networks for each eye and and the head pose angles. The output from the eye networks is multiplied with the corresponding confidence metric. The number of neurons per layer is given in parentheses.

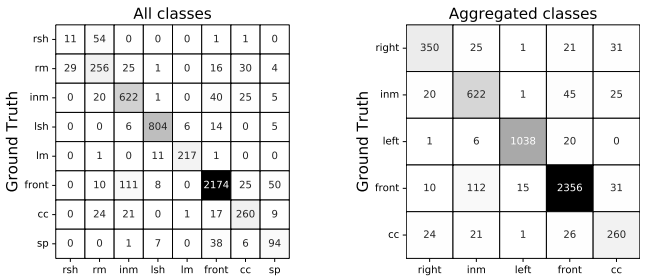

Fig. 6. Left: Confusion matrix of all test results during leave one-driver-out cross-validation. The accuracy is 87.1%. Right: Results are shown for aggregated classes. The accuracy increases to 91.4 %.

from the interval $[0.1,\ldots,1]$. We multiply this kernel with the signals available in $[s_1,\ldots,s_k]$ and obtain another vector with values in $[0,1]$. This vector can then be applied on the probabilities of the corresponding time window.

Moving Average Filter: For the Moving Average (MA) filter, we consider a window of size $2n+1$. For a given frame, we take a look at the n frames before and after that frame. The algorithm then finds the most represented class within this window and labels the frame in the middle with this class. If it comes to a situation that two or more classes have the same amount of labels inside the sliding window, the algorithm uses probabilities to break a tie. It sums up probabilities inside the window for each most represented class. Then, the class with the largest sum of probabilities is assigned to the frame in the middle of the window. Other filter models such as the Hidden-Markov-Model or the Viterbi method were considered as well, but the best results were achieved with the aforementioned filters.

V. RESULTS

A. Cross-Validation

As mentioned before, cross-driver testing is a crucial part in computer vision tasks, especially in appearance-based models. Drivers must not appear in the test set when they appeared during training. In a feature-based approach like in this work, cross-driver testing may not be needed as normalized arrays are extracted from images. However, we believe that choosing a random train-test-split might cause for instable results. Therefore, a leave-one-*driver*-out cross-validation of the system is carried out to assure its performance.

The dataset (Set 1) and the corresponding weights from Table II are used. The weights remain the same during cross-validation. The SGD optimizer is used with a learning rate of 0.001 and a weight decay of 0.0005. After 210 epochs the learning rate was changed to 0.0001 for fine-tuning, and the model with best accuracy was chosen[1].

During training an average accuracy of 94.0% was achieved with 92.4% at the lowest and 96.4% at the highest. For the test sets the overall accuracy of all sets together is 87.1% and 91.4% for the aggregated classes, i.e. the classes *left shoulder* and *left mirror* were aggregated to *left* (resp. for the right side) and the classes *front* and *speedometer* were aggregated to *front*. The confusion matrix is given in Fig. 6. The aggregated accuracy per driver was varying between 75.6% and 100%. For 6 drivers the aggregation resulted in a performance increase of 5% points. The results are displayed in Fig. 7 (left). The weighted precision and weighted recall values are shown in Fig. 7 (right). Precision varies most for the class *right* while the other classes are more concentrated around their average with almost no outliers. When looking at the recall rates, all classes but the class front exhibit outliers below 20%. Most variance is found for the class *inner mirror*. The other classes display more concentrated recall rates at a higher average recall rate above 89%. The highest recall rate with the lowest variance was achieved for the *front* class. For both recall and precision, the weighted and unweighted average values are centered around 88%.

B. Results with Post-processing

In order to test the post-processing techniques, we consider continuously annotated images. The dataset (Set 2) from Table II is used. All vehicle signals related to the center stack region are considered.

We chose the kernel size $k = 31$ around a signal, which corresponds to one second before and after a signal alters[2]. The Signal filter is applied on all available signals in the center stack region. The results are shown in Table III under signals. The overall accuracy increases from 86.5% to 90.5%. The accuracy for *center console* even increases from 74.9% to 90%. We then apply the MA filter with different values for n additionally to the results from the signal filter. Table III shows that the best results are achieved with $n = 2$, while

[1]We only report the best results. The Adam-optimizer, other weight configurations and hyperparameters were also used.

[2]Other values were examined as well with worse results.

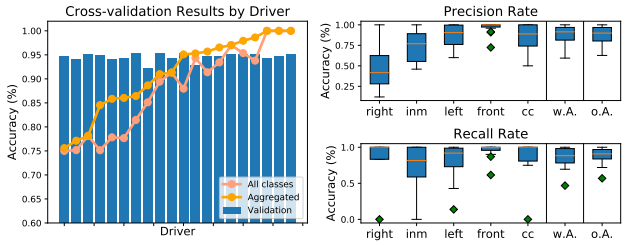

Fig. 7. Left: The accuracy during cross-driver validation is reported for every driver. The columns depict the accuracy reached during training while the graphs report the accuracy of the system applied to the driver left out in training. Right: Cross-validation results for precision and recall are reported, as well as the weighted and unweighted values.

TABLE III
COMPARISON OF RESULTS WITH FILTER METHODS

	front	lm	inm	cc	sp	rm	ls	rs	A.*
MA; $n = 5$	91.4	90.1	84.8	91.4	85.4	87.5	93.9	100	90.8
MA; $n = 2$	93.0	94.9	87.0	92.0	88.3	90.8	93.9	100	92.3
Signals; $k = 31$	91.7	93.2	81.8	90.0	85.95	80.8	83.3	80.0	90.5
Baseline	92.0	92.3	82.6	74.9	86.8	80.8	83.3	80.0	86.5

*A. = weighted accuracy.

$n = 5$ does not yield a considerable improvement. The overall accuracy for $n = 2$ is then 92.3%.

C. Comparison with Other Classifiers

1) Comparison 1: We compare our system with the classification system from Vora et al. (2018) [6], which unlike our system is an appearance-based model linking images to regions directly. The authors originally trained on 47,515 frames from 11 different drivers in higher resolution. In their setup the camera was installed below the rearview mirror resulting in another camera angle. We make use of several videos of their dataset that were published online in a lower resolution. In total, there are 6 videos available with 4 different drivers and 8,130 frames that we labelled according to our best judgement. The classifier needed to be trained again from scratch as the confidence metrics needed to be adjusted for the new camera position. Three drivers appeared in the training set and two drivers appeared in the test set, with one driver appearing in both sets due to the limited availability of data. No post-processing techniques were applied. The results are displayed in Table IV with the results that they presented in their study. It can be shown that the feature-based classifier performs better for the classes *left*, *center console* and *speedometer*, while the values for the other classes are similarly high compared to the appearance-based values. Also, when comparing to the results from the previous sections, the performance for the classes *right*, *inner mirror* increases considerably.

2) Comparison 2: Next, we compare the results from the proposed gaze estimation system to the results that Fridman et al. [4], [14] achieved. The camera in their approach was also attached between the rearview mirror and the center stack region. Similarly to the proposed approach, they made use of a leave-one-driver-out cross-validation with a feature-based approach using all 68 facial landmarks. The information from the eyes was retrieved through a geometrical approach. They only included *confident* predictions, i.e. the ratio of the highest probability to the second highest probability needed to exceed a threshold of 10. In this way, they achieved an accuracy of 91.4% on 7.1% of all images. Before they reported an accuracy of 65.0%. Applying the same threshold to our dataset leads to an overall accuracy of 95.6% on 85.6% of all images. The proposed systems then performs better for the classes *left* and *front*. Results are given in Table IV.

VI. DISCUSSION

The results from cross-validation support the hypothesis that splitting drivers randomly in a test and train set for cross-driver testing can result in considerably different values as seen in the range of the achieved accuracy per driver and in the range of the quantiles for the per-class precision rates (Fig. 7).

The confusion matrices from Fig. 6 show that the system has difficulties to properly separate the classes *inner mirror* and *center console* from the region *front* which may indicate that classes distant to the camera position at the A-pillar are harder to distinguish for the system. This is also reflected in lower precision values for *right*, *center console* and *inner mirror* and lower values with a higher variance in recall for *inner mirror*. The comparisons with [4] and [6] have also confirmed that a considerable proportion of performance can be attributed to the camera position. The new camera position below the rearview mirror produces better results for the proposed method and classes are separated better.

Compared to Fridman et al. [4] who used a similar approach and reported results from the leave-one-driver-out cross-validation, the proposed system is achieving a higher accuracy by over 30%-points. While they can only identify a small share of images (7.1%) as confident predictions the proposed approach is able to *confidently* classify 85.6% of all images at a higher accuracy, which indicates that their geometrical approach for eye information retrieval is not as precise as the eye landmarks.

However, the comparisons of the performance values with the two other methods are only valid to a limited extent due to the changes in the underlying datasets.

The post-processing techniques have confirmed the hypothesis that vehicle signals can help to indicate where the driver is looking. For the case of the center console an improvement was achieved for a time window of 2 seconds. General temporal dependence was shown for gazes within a window of 5 frames corresponding to 0.33 seconds. Choosing a higher window size for the MA filter resulted in worse results as only the few frames before and after a frame are important frames to consider.

When looking at drivers with worse performance values, some remarkable observations can be made as well: either there are only a few instances in a class while testing or the participant has too small eyes. As the pipeline of the introduced system relies on different parts, errors in the

TABLE IV
COMPARISON TO OTHER CLASSIFIERS

	right	**inm**	**left**	**front**	**cc**	**sp**	**A.****
Method from [6]*	100	99.9	94.0	97.7	90.4	89.2	95.2
Proposed method	97.3	97.2	100	84.7	93.6	93.5	94.3
Method from [4]*	94.6	94.7	97.6	69.5	95.5	-	91.4
Proposed method	88.3	94.1	98.8	96.8	88.3	-	95.1

*Results as reported in [4] and [6].
**A. = unweighted accuracy.

prediction along the pipeline may occur. The head pose estimation works reliably with a reported average mean squared error of 6.5° per angle [17], whereas for very small eyes some eye landmarks are estimated falsely, eventually leading to a misclassification.

Regarding the imbalanced dataset the usage of the weighted cross-entropy has showed some effect: the high recall value with its low variance for the class *front* can be attributed to the low weights assigned in the system. An image is only assigned to this class if the system is certain about it. This is also the reason why the recall rate is varying more for the other classes, e.g. *inner mirror*.
In general the open question remains on how to accurately assess the performance of an estimator with imbalanced class sizes. Here, we reported the accuracy and the weighted precision and recall values as the most commonly used performance metrics. Yet, other performance measures could be considered as well, and might be even more suitable.

VII. CONCLUSION

Driver gaze classification systems have the power to assist the driver during all the different automation levels towards fully automated driving. Especially when presenting new systems, cross-validation for generalisability is needed to be applied in order to expose full transparency. We have introduced a feature-based and cross-validated gaze detection system that is able to compete with current state-of-the-art appearance-based systems. We have also shown that vehicle signals and other post-processing methods improve the performance of such systems and indicate where the driver is looking to balancing out false predictions. Accuracies increase from 86.5% to 92.3%. The advantages of a feature-based system are quite obvious: the extracted features can be used for further research e.g. for fatigue detection. Also, we have shown that far less data were needed to build a robust system. We have not investigated further on the roots of some error rates. A possible reason may lie in the frames belonging to transition regions. One approach solving this might be the consideration of set-value instead of single value predictions.

ACKNOWLEDGMENT

The authors like to thank Nico Epple and Valentin Protschky for helpful remarks and comments on the content of this paper.

REFERENCES

[1] S. Klauer, T. A. Dingus, V. L. Neale, J. Sudweeks, and D. Ramsey, "The Impact of Driver Inattention On Near Crash/Crash Risk: An Analysis Using the 100-Car Naturalistic Driving Study Data," Tech. Rep. April, U.S. Department of Transportation, 2006.
[2] X. Mosquet, M. Andersen, and A. Arora, "A Roadmap to Safer Driving," *Auto Tech Review*, vol. 5, no. 7, pp. 20–25, 2015.
[3] I. H. Choi, S. K. Hong, and Y. G. Kim, "Real-time categorization of driver's gaze zone using the deep learning techniques," *2016 International Conference on Big Data and Smart Computing, BigComp 2016*, pp. 143–148, 2016.
[4] L. Fridman, J. Lee, B. Reimer, and T. Victor, "'Owl' and 'Lizard': Patterns of head pose and eye pose in driver gaze classification," *IET Computer Vision*, vol. 10, no. 4, pp. 308–313, 2016.
[5] R. A. Naqvi, M. Arsalan, G. Batchuluun, H. S. Yoon, and K. R. Park, "Deep learning-based gaze detection system for automobile drivers using a NIR camera sensor," *Sensors (Switzerland)*, vol. 18, no. 2, 2018.
[6] S. Vora, A. Rangesh, and M. M. Trivedi, "Driver Gaze Zone Estimation Using Convolutional Neural Networks: A General Framework and Ablative Analysis," *IEEE Transactions on Intelligent Vehicles*, vol. 3, no. 2, pp. 254–265, 2018.
[7] D. D. Salvucci and A. Liu, "The time course of a lane change: Driver control and eye-movement behavior," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 5, no. 2, pp. 123–132, 2002.
[8] A. Doshi and M. M. Trivedi, "Head and eye gaze dynamics during visual attention shifts in complex environments," *Journal of Vision*, vol. 12, no. 2, pp. 1–16, 2012.
[9] O. Lappi, P. Rinkkala, and J. Pekkanen, "Systematic observation of an expert driver's gaze strategy-An on-road case study," *Frontiers in Psychology*, vol. 8, no. APR, pp. 1–15, 2017.
[10] T. Taylor, A. K. Pradhan, G. Divekar, M. Romoser, J. Muttart, R. Gomez, A. Pollatsek, and D. L. Fisher, "The view from the road: The contribution of on-road glance-monitoring technologies to understanding driver behavior," *Accident Analysis and Prevention*, vol. 58, pp. 175–186, 2013.
[11] A. Kar and P. Corcoran, "A Review and Analysis of Eye-Gaze Estimation Systems , Algorithms and Performance Evaluation Methods in Consumer Platforms," *IEEE Access*, pp. 16495–16519, 2017.
[12] J. Deng, W. Dong, R. Socher, L.-J. Li, and L. Fei-Fei, "No Title," in *2009 IEEE Conference on Computer Vision and Patter n Recognition*, pp. 248–255, 2009.
[13] A. Tawari, K. H. Chen, and M. M. Trivedi, "Where is the driver looking: Analysis of head, eye and iris for robust gaze zone estimation," *2014 17th IEEE International Conference on Intelligent Transportation Systems, ITSC 2014*, pp. 988–994, 2014.
[14] L. Fridman, P. Langhans, J. Lee, and B. Reimer, "Driver Gaze Region Estimation without Use of Eye Movement," *IEEE Intelligent Systems*, vol. 31, no. 3, pp. 49–56, 2016.
[15] J. Lee, M. Muñoz, L. Fridman, and T. Victor, "Investigating the correspondence between driver head position and glance location," *PeerJ Computer Science*, vol. 4, p. e146, 2019.
[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9905 LNCS, pp. 21–37, 2016.
[17] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 2018-June, pp. 2155–2164, 2018.
[18] 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)," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-Octob, pp. 1021–1030, 2017.
[19] S. Park, X. Zhang, A. Bulling, and O. Hilliges, "Learning to find eye region landmarks for remote gaze estimation in unconstrained settings," in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications - ETRA '18*, (New York, New York, USA), pp. 1–10, ACM Press, 2018.
[20] S. Panchapagesan, M. Sun, A. Khare, S. Matsoukas, A. Mandal, B. Hoffmeister, and S. Vitaladevuni, "Multi-task learning and weighted cross-entropy for DNN-based keyword spotting," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 08-12-Sept, no. September 2016, pp. 760–764, 2016.