IET Computer Vision

Special Issue: Video Analytics

‘Owl’ and ‘Lizard’: patterns of head pose and eye pose in driver gaze classification

Lex Fridman1, Joonbum Lee1, Bryan Reimer1, Trent Victor2
1Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts 02142, USA
2SAFER, Chalmers University of Technology, Gothenburg SE-412 96, Sweden
E-mail: fridman@mit.edu

Abstract: Accurate, robust, inexpensive gaze tracking in the car can help keep a driver safe by facilitating the more effective study of how to improve (i) vehicle interfaces and (ii) the design of future advanced driver assistance systems. In this study, the authors estimate head pose and eye pose from monocular video using methods developed extensively in prior work and ask two new interesting questions. First, how much better can they classify driver gaze using head and eye pose versus just using head pose? Second, are there individual-specific gaze strategies that strongly correlate with how much gaze classification improves with the addition of eye pose information? The authors answer these questions by evaluating data drawn from an on-road study of 40 drivers. The main insight of the study is conveyed through the analogy of an ‘owl’ and ‘lizard’ which describes the degree to which the eyes and the head move when shifting gaze. When the head moves a lot (‘owl’), not much classification improvement is attained by estimating eye pose on top of head pose. On the other hand, when the head stays still and only the eyes move (‘lizard’), classification accuracy increases significantly from adding in eye pose. The authors characterise how that accuracy varies between people, gaze strategies, and gaze regions.

1 Introduction

The classification of driver's visual attention allocation is an area of increasing relevance in the pursuit of accident reduction. The allocation of visual attention away from the road has been linked to accident risk [1, 2] and a drop in situational awareness as uncertainty in the environment increases [3]. Driver distraction is often construed as a key source of attention divergence from the roadway and the topic of numerous scientific studies and design guidelines [4, 5].

Furthermore, as the level of vehicle automation continues to increase through advanced driver assistance systems as well as other higher forms of automation, freeing available resources from the primary operational task, drivers are expected to be increasingly allowed to glance away from the roadway for greater periods. When the need arises to orient the driver to the roadway, different alerting strategies may be advantageous. Such work would suggest that a real-time estimation of driver's gaze could be coupled with an alerting system to enhance safety [6]. Gaze tracking from video in the driving context is a difficult problem especially due to rapidly varying lighting conditions. Other challenges, common to other domains, include unpredictability of the environment, presence of eyeglasses or sunglasses occluding the eye, partial occlusion of the pupil due to squinting, vehicle vibration, image blur, poor video resolution, and so on. We consider the challenging case of uncalibrated monocular video because it has been and continue to be the most commonly available form of video in driving datasets due to its low equipment and installation costs.

From the perspective of image processing, gaze estimation can be divided into two components: head pose estimation and eye pose estimation. Due to all the factors above, the latter is more difficult than the former. In fact, gaze classification performance can be good based on head pose alone [7], because it frequently correlates with eye pose, but not always. ‘Eye pose’ and ‘head pose’ are terms used throughout this paper to mean the relative orientation of the pupil in the eye socket and the relative orientation of facial features on the head, respectively. This use of ‘pose’ is made broader in order to allow for the non-linear modelling discussed in Section 4.2.

In this paper, we seek to characterise when eye pose significantly contributes to gaze classification and when it does not. Specifically, we ask two questions:

i. Contribution of eye pose: How much better can we classify driver gaze using (a) head and eye pose together versus (b) using head pose alone?

ii. Classification of different gaze strategies: Are there individual-specific gaze strategies that strongly correlate with how much gaze classification improves with the addition of eye pose information?

These two questions are answered by analysing data drawn from an on-road study of 40 drivers performing secondary tasks of varying complexity. The inter-person classification and gaze strategy variation is discussed using the analogy of an ‘owl’ and ‘lizard’ (introduced previously in [8, 9]) which describes the degree to which the eyes and the head move when shifting gaze. When the head moves a lot (‘owl’), not much classification improvement is attained by estimating eye pose on top of head pose. On the other hand, when the head stays still and only the eyes move (‘lizard’), classification accuracy increases significantly from adding in eye pose. Examples of the two strategies are shown in Fig. 1. We propose an end-to-end driver gaze classification system based on monocular video and use it to explore the importance of eye pose for classification performance as we move along the spectrum of people from ‘owl’ to ‘lizard’.

2 Related work

The problem of gaze tracking from monocular video has been investigated extensively across many domains [10, 11]. We build on this work to characterise the individual contribution of head movement and eye movement to gaze classification accuracy. The building blocks of our image processing pipeline are: face alignment, head pose estimation, and pupil detection. We apply cutting-edge algorithms from these fields to answer two questions posed by our work (see Section 1) on a large on-road driving dataset.
The algorithm in [12] uses an ensemble of regression trees for super-real-time face alignment. Our face feature extraction algorithm draws upon this method as it is built on a decade of progress on the face alignment problem (see [12] for a detailed review of prior work). The key contribution of the algorithm is an iterative transform of the image to a normalised coordinate system based on the current estimate of the face shape. Also, to avoid the non-convex problem of initially matching a model of the shape to the image data, the assumption is made that the initial estimate of the shape can be found in a linear subspace.

Head pose estimation has a long history in computer vision. Murphy-Chutorian and Trivedi [13] describe 74 published and tested systems from the last two decades. Generally, each approach makes one of several assumptions that limit the general applicability of the system in driver gaze detection. These assumptions include: (i) the video is continuous, (ii) initial pose of the subject is known, (iii) there is a stereo vision system available, (iv) the camera has frontal view of the face, (v) the head can only rotate on one axis, and (vi) the system only has to work for one person. While the development of a set of assumptions is often necessary for the classification of a large number of possible poses, our approach skips the head pose estimation step [i.e. the computation of a vector in three-dimensional (3D) space modelling the orientation of the head] and goes straight from the detection of a facial feature to a classification of gaze to one of the six glance regions. Prior work has shown that such a classification set is sufficient for the in-vehicle environment, even under rapidly shifting lighting conditions [7].

Pupil detection approaches have been extensively studied. Methods usually track corneal reflection, distinct pupil shape in combination with edge-detection, characteristic light intensity of the pupil, or a 3D model of the eye to derive an estimate of an individual's pupil, iris, or eye position [14]. Our approach uses an adaptive CDF-based method [15] in conjunction with face alignment that significantly narrows the search space.

Studies of the correlation between head and eye movement have shown inter-person variation in the degree to which the head serves as a proxy for gaze [7, 9]. For example, a previous work tested drivers' head movements while looking at the 'road' and the 'center stack' and found that drivers' horizontal range of head movements varied from 5 to 20 degrees across individuals [9]. This paper makes this variation more explicit by characterising classification performance with and without eye pose information.

3 Dataset

Training and evaluation is carried out on a dataset of 40 subjects drawn from a larger driving study of 80 subjects that took place on a local interstate highway (see [16] for detailed experimental methods). For each subject, the collection of data was carried out in one of the two vehicles: 2013 Chevrolet Equinox or Volvo XC60 (randomly assigned). The drivers performed a number of secondary tasks of varying difficulty including using the voice interface in the vehicle to enter addresses into the navigation system and using the voice interface as well as manual controls to select phone numbers from a stored phone list.

Both vehicles were instrumented with an array of sensors for assessing driver behaviour. The sensor set included a camera positioned on the dashboard of each vehicle that was intended to capture the driver's face for annotation of glance behaviour. The cameras were positioned off-axis to the driver and in slightly different locations in the two vehicles (based on features of the dashboard, etc.). As each driver positioned the seat (electronic in both vehicles) differently, the relative position of the driver in relation to the camera varied somewhat by subject and across each driver over time (i.e. drivers move continuously in the seat, etc.). The camera was an Allied Vision Tech Guppy Pro F-125, capturing greyscale images at a resolution of 800 × 600 and speed of 30 fps. The data was double manually annotated for driver glances transitions during secondary task periods (at a resolution of sub-200 ms) into one of 11 classes (road, center stack, instrument cluster, rearview mirror, left, right, left blindspot, right blindspot, passenger, uncodable, and other). As detailed in [16], any discrepancies between the two annotators were mediated by an arbitrator. This method of double annotation and mediation of driver gaze has been shown to produce very accurate annotations that can be effectively used as ground truth for supervised learning approaches [17].

In this paper, a broad random subset of data was drawn from the initial experiment and the 'left' and 'left blind spot' classes/'right', 'right blind spot', 'passenger' classes were collapsed, respectively, in to 'left' and 'right'. Periods that were labelled 'uncodable' and 'other' were excluded. Subject pruning was completed to ensure that every subject under consideration has sufficient training data for each of the six glance regions (road, center stack, instrument cluster, rearview mirror, left, and right). The threshold for 'sufficient training' was that each subject had at least 120 frames of video (where pupils were detected) for each of the six gaze regions.

As shown in Table 1, the resulting dataset contains 1,351,864 images each annotated as belonging to one of the six glance regions. The algorithm described in Section 4 is used for face detection, face alignment, and pupil detection. The gaze classification approach requires a face and a pupil to be successfully detected in the image. The filtering procedure is discussed in detail in Section 4. Therefore, in the evaluation we include only the images where a face and a pupil are detected. As the table shows, on average, a face is detected in 79.4% of images.
Around 61.6% of images pass the full image processing pipeline where both a face and a pupil are detected.

4 Gaze classification pipeline

The steps in the gaze region classification pipeline are: (i) face detection, (ii) face alignment, (iii) pupil detection, (iv) feature extraction and normalisation, (v) classification, and (vi) decision pruning. If the system passes the first three steps, it will lead to a gaze region classification decision for every image fed into the pipeline. In step 6, that decision may be dropped if it falls below a confidence threshold (see Section 4.5). The three face images in Fig. 2 are examples of the result achieved in the first four steps of the pipeline: going from a raw video frame to the extracted face features and pupil position. As mentioned in Section 1, the relative orientation of facial features serves as a proxy for ‘head pose’ and the relative orientation of pupil position serves as a proxy for ‘eye pose’. We discuss each of the six steps in the pipeline in the following sections.

4.1 Face detection

The environment inside the car is relatively controlled in that the camera position is fixed and the driver torso moves in a fairly contained space. Thus, a camera can be positioned such that the driver’s face is always fully or almost fully in the field of view. However, the lighting conditions are sometimes drastically variable (e.g. quickly passing under a bridge, reflection of the sun on the camera lens, etc.) and thus there are frequent cases where the intensity distribution of the image does not allow for successful detection of the face (i.e. false reject). Every detection step in the pipeline is tuned to have a low false accept rate (FAR). A false accept error early in the pipeline propagates and can result in drastically incorrect head pose and eye pose estimation. In the context of video-based driver gaze classification, a high false reject rate is more acceptable than a high FAR.

The face detector in our pipeline uses a histogram of oriented gradients combined with a linear support vector machine (SVM) classifier, an image pyramid, and sliding window detection scheme implemented in the DLIB C++ library [18]. The performance of this detector has lower FAR than the widely-used default Haar-feature-based face detector available in OpenCV [19] and thus is more appropriate for our application.

4.2 Face alignment and head pose

Both face alignment and head pose estimation are extremely well studied problems in computer vision [13, 20]. We investigated several cutting edge methods from each domain, and chose the ones that worked best for monocular video with highly varying lighting conditions.

Face alignment in our pipeline is performed on a 68-point multi-PIE facial landmark mark-up used in the iBUG 300W dataset [21]. These landmarks include parts of the nose, upper edge of the eyebrows, outer and inner lips, jawline, and parts in and

Table 1 Dataset statistics for the total number of video frames annotated, the number of frames where faces were detected, and the number of frames where pupil were detected. Each of these pruning steps are discussed in Section 4

<table>
<thead>
<tr>
<th>Pruning steps</th>
<th>Total frames remaining</th>
<th>Fraction of original, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>0. total frames annotated</td>
<td>1,351,864</td>
<td>100</td>
</tr>
<tr>
<td>1. frames with faces detected</td>
<td>1,073,380</td>
<td>79.4</td>
</tr>
<tr>
<td>2. frames with pupils detected</td>
<td>833,049</td>
<td>61.6</td>
</tr>
</tbody>
</table>
around the eye. The selected landmarks are shown as black dots in Fig. 2. The algorithm for aligning the 68-point shape to the image data uses a cascade of regressors as described in [12] and implemented in [18]. The two characteristics of this algorithm most important to driver gaze localisation is: (i) it is robust to partial occlusion and self-occlusion and (ii) its running-time is significantly faster than 30 fps rate of incoming images.

Face alignment produces estimates for facial feature positions in the image. These features can be mapped directly to a gaze region using methods that fall under the non-linear regression category defined in [13]. They can also be mapped to a 3D model of the head. The resulting 3D–2D point correspondence can be used to compute the orientation of the head. This is categorised under geometric methods in [13]. Then the yaw, pitch, and roll of the head can be used as features for a gaze region classifier. We implemented both methods and found the former (non-linear classification) to be more robust to errors in the face alignment and pupil detection steps of the pipeline. The geometric approach uses OpenCV’s SolvePNP solution of the PnP problem [22]. The non-linear classification approach is discussed further in Section 4.5.

4.3 Pupil detection

As described in Section 1, the problem of accurate pupil detection is more difficult than the problem of accurate face alignment, but both are not always robust to poor lighting conditions. Therefore, the secondary task of pupil detection is to flag errors in the face alignment step that preceded it. As Table 1 shows, the face is detected in 79.4% video frames but only 61.6% of the original frames pass the pupil detection step.

We use a CDF-based method [15] to extract the pupil from the image of the right eye, and adjust the extracted pupil blob using morphological operations of erosion and dilation. The six steps in this process are as follows:

i. Extract the right eye from the face image based on the right eye features computed as part of the face alignment step.

ii. Remove all pixels that fall outside the boundaries of the polygon defined by the six eye features.

iii. Rescale the intensity such that the 98-percentile intensity becomes 0.0 and 2-percentile intensity becomes 1.0.

iv. Define a CDF intensity threshold and convert the greyscale image to a binary image. Each pixel intensity above the threshold becomes 1, and otherwise becomes 0.

v. Perform an ‘opening’ morphology transformation (described in [23]). This operation is useful for removing small holes in large blobs.

vi. Perform a ‘closing’ morphology transformation (described in [23]). This operation is useful for removing small objects and smoothing the shape of large blobs.

The above steps have three parameters: the CDF threshold, the opening window size, and the closing window size. These parameters are dynamically optimised for each image over a discrete set of values in order to maximise the size of the largest resulting blobs under one constraint: the largest blob must be circle-shaped (i.e. have similar height and width). More specifically, each of the three parameters take on three values and using exhaustive search we find the set of parameter values that results in the largest circular blob.

The pupil detection process also includes pruning procedures based on whether the eye is sufficiently open and whether there is a possible error in the preceding face alignment step. These are:

i. An eye shape height that is <10% of its width is considered ‘closed’ and is removed from the pipeline.

ii. When a sufficiently large blob is not found in the eye region (<5 pixels in area), it is assumed that the face alignment did not properly localise the eye and the image is removed from the pipeline.

4.4 Feature extraction and normalisation

The driver spends more than 90% of their time looking forward at the road and this fact was used in [7] to normalise the position of facial features relative to the average bounding box of the face associated with the ‘road’ gaze region. This required an initial 120 s period of automated calibration. In this paper, we remove the need for calibration and instead normalise the facial features based on the bounding box of the eyes and nose for the current frame only. Fig. 1 shows departure of the head and eyes away from their ‘reference’ positions. The normalisation step linearly transforms the facial landmarks such that the landmarks of the eyes and nose fit a unit square. After this transformation, the relative orientation of the facial landmarks becomes the feature vector for the gaze classification step. The bounding box of the eyes and nose was experimentally found to be the most robust normalising region. This is due to the fact that the greatest noise in the face alignment step was associated with the features of the jawline, the eyebrows, and the mouth. The position of the pupil is normalised to the bounding box of the eye rotated such that the two eye corners lie on a horizontal line.

4.5 Classification and decision pruning

SciKit-learn implementation of a random forest classifier [24] is used to generate a set of probabilities for each class from a single feature vector. The probabilities are computed as the mean predicted class probabilities of the trees in the forest. The class probability of a single tree is the fraction of samples of the same class in a leaf. A random forest classifier of depth 25 with an ensemble of 2000 trees is used for all experiments in Section 5. The class with the highest probability is the one that the system assigns to the image as the ‘decision’. The ratio of the highest probability to the second highest probability is termed the ‘confidence’ of the decision. A confidence of 1 is the minimum. There is no maximum. The effect of this threshold is explored in [7]. For the experiments in Section 5 a confidence threshold of 10 is used, which means that any decisions with a confidence >10 are accepted and the others are ignored. A random forest classifier was used because it achieved a much higher accuracy than k-nearest neighbours and linear SVM classifiers. RBF-kernel SVM achieved a slightly higher accuracy but at the cost of over a 100-fold increase in training time.

5 Results

5.1 Gaze region classification

We evaluate the gaze classification pipeline described in Section 4 on the dataset of 40 drivers described in Section 3. In all the experiments and discussions that follow, the key comparison is between classification performed using head pose alone and classification performed using head pose and eye pose together. The classification problem has six classes, one for each of the six gaze regions: (i) road, (ii) center stack, (iii) instrument cluster, (iv) rearview mirror, (v) left, and (vi) right.

The pipeline starts at an annotated frame from the raw video. As previously described, each frame is double annotated and mediated ensuring that the gaze region annotations can reliably serve as ground truth for the cross validation training and testing. There are a total of 1,351,864 annotated images. As shown in Table 1, 833,049 of those images pass through the face detection, face alignment, and pupil detection steps of the pipeline. As discussed in Section 4.5, we further reduce this number during testing by only considering decisions with a confidence above the confidence threshold of 10. On average, only 7.1% of all decisions are deemed confident in this way, resulting in a decision rate of 2.3 Hz. This selection is distributed evenly through time among cases where a face is successfully detected. If we consider the fraction of original raw video frames that lead to a confident gaze classification decision, then the overall effective decision rate is 1.3 Hz.

All of the plots in this section share the same experiment setup. For each user, we train the six-class classifier on all 39 others
users. The training data for each of the six classes is balanced by random sub-sampling [25]. The testing is performed on the data for the one user by balancing the classes through super-sampling. This helps ensure that the per-class accuracy is not skewed by the greater representation of ‘road’ versus the other five classes in the dataset. The process is repeated 100 times for each of the 40 users. The plots with error bars indicate the standard deviation of accuracy among the 100 runs for each user.

Fig. 3 shows the confusion matrix for classification using both head pose and eye pose. The overall accuracy achieved is 94.6%. Most of the errors in classification are incorrectly labelling an image as ‘road’ when it is one of the other five gaze regions. Fig. 4 compares the accuracy in this confusion matrix with that achieved by a system that only uses head pose information. The overall accuracy achieved by such a system is 89.2%. One of the questions posed by this paper is: how much we gain by considering eye pose on top of head pose? The answer in our final optimised system is 5.4% accuracy. As Fig. 4 shows, the biggest gain of 8.7% is achieved for the center stack region. This can be interpreted to mean that people are more likely to use only their eyes when glancing down to the center stack or that the head pose associated with the center stack is similar to the head pose of other gaze regions like ‘road’, ‘instrument cluster’, and ‘rearview mirror’ as Fig. 3 suggests.

5.2 User-specific classification and gaze strategies

As shown in Section 5.1, adding in eye pose to head pose increases gaze classification accuracy by 5.4%. However, that does not tell the full story because some user-trained classifiers benefit more from eye pose than others. Fig. 5 shows the variation in accuracy among users before and after adding in eye pose to the classification feature set. For many users, 100% accuracy is achieved, while for many others accuracy drops to below 80% and even to as low as 40%.

Using the Pearson correlation coefficient as a guide, we programmatically explored over a million pairs of variables in search of an answer to the question of what explains this difference in classification performance between users. Some variables correlate with per-region accuracy but not overall. For example, average magnitude of off-centre head movement and pupil movement are good predictors of classification accuracy for the ‘right’ region with head pose alone and with head and eye pose together, respectively. We were not able to find a measure of an individual that correlated highly with overall classification accuracy, but there are a few variables that correlate with the increase in accuracy achieved by adding in eye pose. The most interesting and intuitive one is a metric we refer to as ‘owlness’. It is defined as

\[ M = \frac{d_s}{d_s + d_p} \]
where $d_b$ and $d_p$ are the distance of the nose tip and right pupil, respectively, from their average position in the background model. Due to the normalisation of the features, both distances are in the range $[0, 1]$. An $M$ value of 0 means that a shift in gaze involves only the eyes (‘lizard’). An $M$ value of 1 means that a shift in gaze involves only the head (‘owl’).

Fig. 6a shows the relationship between the ‘owlness’ metric and the per-user increase in accuracy achieved. The measure of ‘owlness’ for each user is computed by averaging the result of (1) for each image that passes the face detection and pupil detection steps in the pipeline. We partition users into three groups: ‘owl’, ‘lizard’, and ‘mixed’ based on the value of $M$. Fig. 6b shows how effectively these partitions separate the users who gain classification accuracy from eye pose and those who do not. In this figure, the ‘owls’ see no effect or a decrease in accuracy, while the ‘lizards’ see a significant increase in accuracy.

### 6 Conclusion

This paper investigates the contribution of head pose and eye pose to gaze classification accuracy for different gaze strategies. We answer two questions: (i) how much does an eye pose contribute? and (ii) how can the inter-user accuracy variation be explained? For the former, we show that eye pose adds a 5.4% increase in average accuracy (from 89.2 to 94.6%) with an effective average rate of 1.3 decisions per second. For the latter, we propose an ‘owlness’ metric that decomposes gaze into head movement and eye movement and computes the relative magnitude of each. This metric is used to explain the inter-person variation in impact of eye pose on gaze classification accuracy.

### 7 Acknowledgments

Support for this work was provided by the Santos Family Foundation, the New England University Transportation Center, and the Toyota Class Action Settlement Safety Research and Education Program. The views and conclusions being expressed are those of the authors, and have not been sponsored, approved, or endorsed by Toyota or plaintiffs’ class counsel. Data was drawn from studies supported by the Insurance Institute for Highway Safety (IIHS).

### References


