Driver Gaze Tracking and Eyes Off the Road Detection System

Francisco Vicente, Zehua Huang, Xuehan Xiong, Fernando De la Torre, Wende Zhang, and Dan Levi

***Abstract*—Distracted driving is one of the main causes of vehicle collisions in the United States. Passively monitoring a driver's activities constitutes the basis of an automobile safety system that can potentially reduce the number of accidents by estimating the driver's focus of attention. This paper proposes an inexpensive vision-based system to accurately detect Eyes Off the Road (EOR). The system has three main components: 1) robust facial feature tracking; 2) head pose and gaze estimation; and 3) 3-D geometric reasoning to detect EOR. From the video stream of a camera installed on the steering wheel column, our system tracks facial features from the driver's face. Using the tracked landmarks and a 3-D face model, the system computes head pose and gaze direction. The head pose estimation algorithm is robust to nonrigid face deformations due to changes in expressions. Finally, using a 3-D geometric analysis, the system reliably detects EOR.**

The proposed system does not require any driver-dependent calibration or manual initialization and works in real time (25 FPS), during the day and night. To validate the performance of the system in a real car environment, we conducted a comprehensive experimental evaluation under a wide variety illumination conditions, facial expressions, and individuals. Our system achieved above 90% EOR accuracy for all tested scenarios.

***Index Terms*—Driver monitoring system, eyes off the road detection, gaze estimation, head pose estimation.**

I. Introduction

DRIVER distractions are the leading cause of most vehicle crashes and near-crashes. According to a study released by the National Highway Traffic Safety Administration (NHTSA) and the Virginia Tech Transportation Institute (VTTI) [16], 80% of crashes and 65% of near-crashes involve some form of driver distraction. In addition, distractions typically occurred within three seconds before the vehicle crash. Recent reports have shown that from 2011 to 2012, the number of people injured in vehicle crashes related to distracted driving has increased 9% [1]. In 2012 alone, 3328 people were killed due to distracted driving crashes, which is a slight reduction from the 3360 in 2011.

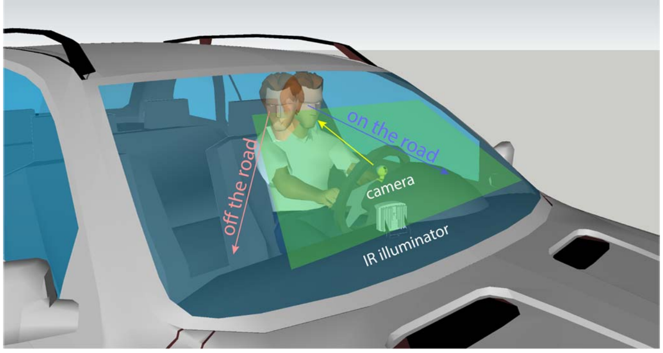


Fig. 1. Eyes off the road (EOR) detection system.

Distracted driving is defined as any activity that could divert a person's attention away from the primary task of driving. Distractions include texting, using a smartphone, eating and drinking, adjusting a CD player, operating a GPS system or talking to passengers.

This is particularly challenging nowadays, where a wide spectrum of technologies have been introduced into the car environment. Consequently, the cognitive load caused by secondary tasks that drivers have to manage has increased over the years, hence increasing distracted driving. According to a survey [14], performing a high cognitive load task while driving affects driver visual behavior and driving performance. References [22] and [36] reported that drivers under high cognitive loads showed a reduction in the time spent examining mirrors, instruments, traffic signals, and areas around intersections.

Especially concerning is the use of hand-held phones and other similar devices while driving. NSTHA [16] has reported that texting, browsing, and dialing cause the longest period of drivers taking their Eyes Off the Road (EOR) and increase the risk of crashing by three fold. A recent study [41] shows that these dangerous behaviors are wide-spread among drivers, 54% of motor vehicle drivers in the United States usually have a cell phone in their vehicles or carry cell phones when they drive.

Monitoring driver activities forms the basis of a safety system that can potentially reduce the number of crashes by detecting anomalous situations. In [29], authors showed that a successful vision-based distracted driving detection system is built upon reliable EOR estimation, see Fig. 1. However, building a real-time EOR detection system for real driving scenarios is very challenging for several reasons: (1) The system must operate during the day and night and under real world illumination conditions; (2) changes in drivers' head pose and eye movements

Manuscript received February 4, 2014; revised August 26, 2014; accepted December 22, 2014. Date of publication March 3, 2015; date of current version July 31, 2015. The work of F. Vicente and Z. Huang was supported by General Motors. The work of F. De la Torre and X. Xiong was supported in part by Grant CPS-0931999, by NSF IIS-1116583, and by internal funds from the Robotics Institute. The Associate Editor for this paper was K. Wang.

F. Vicente, Z. Huang, X. Xiong, and F. De la Torre are with the Carnegie Mellon University, Pittsburgh, PA 15213-3815 USA (e-mail: fvicente@andrew.cmu.edu; huazehuang@gmail.com; xiong828@gmail.com; ftorre@cs.cmu.edu).

W. Zhang is with the General Motors Company, Warren, MI 48090 USA (e-mail: wende.zhang@gm.com).

D. Levi is with the General Motors Company, Herzliya 46733, Israel (e-mail: dan.levi@gm.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2015.2396031

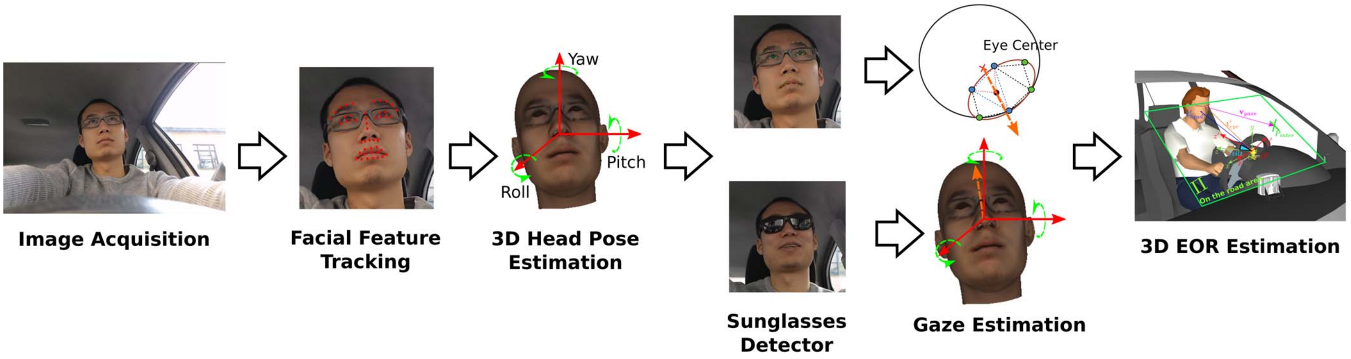


Fig. 2. Overview of the eyes off the road (EOR) detection algorithm.

result in drastic changes in the facial features (e.g., pupil and eye corners) to be tracked; (3) the system must be accurate for a variety of people across multiple ethnicities, genders, and age ranges. Moreover, it must be robust to people with different types of glasses. To address these issues, this paper presents a low-cost, accurate, and real-time system to detect EOR. Note that EOR detection is only one component of a system for detecting and alerting distracted drivers.

Fig. 2 illustrates the main components of our system. The system collects video from a camera installed on the steering wheel column and tracks facial features, see Fig. 1. Using a 3D head model, the system estimates the head pose and gaze direction. Using 3D geometric analysis, our system introduces a reliable method for EOR estimation. Our system works at 25 FPS in MATLAB and does not require any specific driver dependent calibration or manual initialization. It supports glasses (including sunglasses) and operates during the day and night. In addition, the head pose estimation algorithm uses a 3D deformable head model that is able to handle driver facial expressions (i.e., yawning and talking), allowing reliable head pose estimation by decoupling rigid and non-rigid facial motion. Experiments in a real car environment show the effectiveness of our system.

II. Previous Work

Driver monitoring has been a long standing research problem in computer vision. It is beyond the scope of the paper to review all existing systems, but we provide a description of the most relevant work in academia and industry. For a complete overview of existing systems, we refer the reader to [14].

Broadly speaking, there are two approaches to estimate gaze direction: Techniques that only use the head pose and those that use the driver's head pose and gaze. For systems that rely only on head pose estimation, an extensive report on the topic can be found in [34]. Lee *et al.* [30] proposed an algorithm for yaw and pitch estimation based on normalized histograms of horizontal and vertical edge projections combined with an ellipsoidal face model and a Support Vector Machine (SVM) classifier for gaze estimation. Chutorian *et al.* [33] proposed a driver head pose estimation algorithm based on Localized Gradient Orientation (LGO) histograms in combination with Support Vector Regressors (SVR). The algorithm was further developed in [35] by introducing a head tracking module built upon 3D motion estimation and a mesh model of the driver's head. Recently, Rezaei and Klette [37] introduced a new algorithm for distracted driving detection using an improved 3D head pose estimation and Fermat-point transform. All the described approaches reported to work in real time.

Systems that use head pose and gaze estimation are grouped into hardware and software based approaches. Ishikawa *et al.* [25] proposed a passive driver gaze tracking system using Active Appearance Models (AAMs) for facial feature tracking and head pose estimation. The driver's pupils were also tracked and a 3D eye-model was used for accurate gaze estimation from a monocular camera. Smith *et al.* [39] relied on motion and color statistics to robustly track driver head and facial features. Using the assumption that the distance from the driver's head to the camera is fixed, the system recovered the three dimensional gaze of the eyes using a simplified head model without any calibration process.

Hardware-based approaches to driver head pose and gaze estimation rely on near-infrared (IR) illuminators to generate the bright pupil effect [7], [26], [27], [32]. The bright pupil effect allows for low-cost pupil detection, which simplifies localization of the driver's pupil using only computer vision techniques. Ji and Yang [26], [27] described a system for driver monitoring using eye, gaze, and head pose tracking based on the bright pupil effect. The pupils are tracked using a Kalman filter; the system uses image features around the pupil in combination with a nearest neighbor classifier for head pose estimation. The gaze is estimated by extracting the displacement and direction from the center of the pupil to the glint and using linear regression to map to nine gaze directions. This system is not person-independent and must be calibrated for every system configuration and driver. Batista [7] used a similar system but provided a more accurate gaze estimation using ellipse fitting for the face orientation. These near-IR illumination systems work particularly well at night, but performance can drop dramatically due to contamination introduced by external light sources and glasses [8], [21]. While the contamination due to artificial lights can easily be filtered with a narrow band pass filter, sunlight contamination will still exist. Additionally, the hardware necessary to generate the bright eye effect will hinder system integration into the car dashboard.

In industry, systems based on near-IR are the most common. The Saab Driver Attention Warning System [6] detects visual inattention and drowsy driving. The system uses two miniature IR cameras integrated with Smart Eye technology to accurately

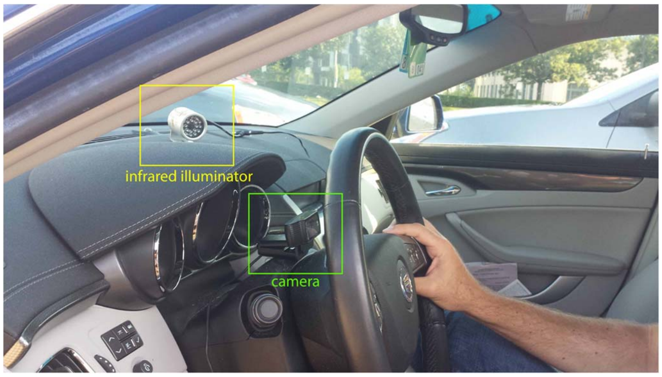


Fig. 3. Camera and IR illuminator position.

estimate head pose, gaze, and eyelid status. When a driver's gaze is not located inside the primary attention zone (which covers the central part of the frontal windshield) for a predefined period, an alarm is triggered. Nevertheless, no further details about the performance of the system in real driving scenarios were reported. Toyota has equipped their high-end Lexus models with their Driver Monitoring System [24]. The system permanently monitors the movement of the driver's head when looking from side to side using a near-IR camera installed on the top of the steering wheel column. The system is integrated into Toyota's pre-crash system, which warns the driver when a collision is probable.

Another commercial system is FaceLAB [2], a stereo-based eye tracker that detects eye movement, head position and rotation, eyelid aperture and pupil size. FaceLAB uses a passive pair of stereo cameras mounted on the car dashboard. The system has been used in several driver assistance and inattention systems, such as [17]–[19]. However, stereo-based systems are too expensive to be installed in mass-produced cars and they require periodic re-calibration because vibrations cause the system calibration to drift over time. Similarly, Smart Eye [3] uses a multicamera system that generates 3D models of the driver's head, allowing it to compute her gaze direction, head pose, and eyelid status. This system has been evaluated in [5]. Unfortunately, it is prohibitively expensive for mass dissemination in commercial cars and it imposes strong constraints with respect to the necessary hardware to be installed. As a result, it is unfeasible to install this system in regular cars. Other commercial systems include the ones developed by Delphi Electronics [15] and SensoMotoric Instruments GmbH [4].

III. System Description

This section describes the main components of our system. There are six main modules: Image acquisition, facial feature detection and tracking, head pose estimation, gaze estimation, EOR detection, and sunglasses detection. Fig. 2 shows the system block diagram and algorithm flow.

A. Image Acquisition

The image acquisition module is based on a low-cost CCD camera (in our case, a Logitech c920 Webcam) placed on top of the steering wheel column, see Fig. 3. The CCD camera was placed over the steering wheel column for two reasons: (1) It facilitates the estimation of gaze angles, such as pitch, which is relevant for detecting when the driver is texting on a phone (a major threat to safety). (2) From a production point of view, it is convenient to integrate a CCD camera into the dashboard. On the downside, when the wheel is turning there will be some frames in which the driver's face will be occluded by the steering wheel.

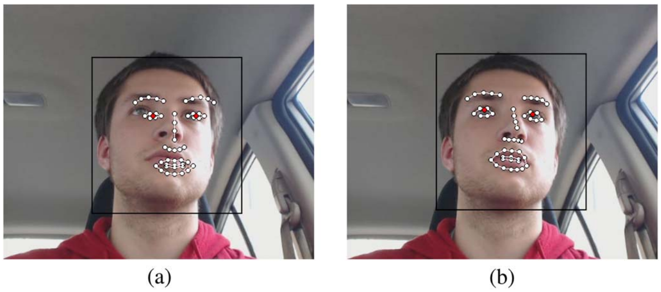


Fig. 4. a) Mean landmarks, $\mathbf{x}_0$, initialized using the face detector. Black outline indicates face detector. b) Manually labeled image with 51 landmarks.

For night time operation, the system requires an illumination source to provide a clear image of the driver's face. Moreover, the illumination system cannot impact the driver's vision. To this end, an IR illuminator was installed on the car dashboard, see Fig. 3. Note that the proposed system does not suffer from the common drawbacks of near-IR based systems [7], [26], [27], because it does not rely on the bright pupil effect. To adapt our CCD camera to IR illumination, it was necessary to remove the IR filter from the CCD camera, making the CCD more sensitive to IR illumination (i.e., sunlight, artificial illumination). As shown in Fig. 5, this effect is not noticeable in real driving scenarios.

B. Facial Feature Detection and Tracking

Parameterized Appearance Models (PAMs), such as Active Appearance Models (e.g., [12], [13]) and Morphable Models [9], are popular statistical techniques for face tracking. They build an object appearance and shape representation by computing Principal Component Analysis (PCA) on a set of manually labeled data. Fig. 4(a) illustrates an image labeled with p landmarks ($p = 51$ in this case). Our model includes two extra landmarks for the center of the pupils. However, there are several limitations of PAMs that prevent to use them for detection and tracking in our system. First, PAMs typically optimize many parameters (about 50–60), which makes them very prone to local minima. Second, PAMs work very well for person-specific subjects but do not generalize well to other untrained subjects because they use a linear model of shape and appearance [13]. Third, the shape model typically cannot model asymmetric expressions (e.g., one eye open and another closed, or an asymmetric smile). This is due to the fact that in most training datasets, these expressions do not occur.

To address the limitations of PAMs, Xiong and De la Torre proposed the Supervised Descent Method (SDM) [44], which is a discriminative method for fitting PAMs. There are two main differences from the traditional PAMs. First, it uses a non-parametric shape model that is better able to generalize to untrained situations (e.g., asymmetric facial gestures). Second, SDM uses a more complex representation (SIFT descriptor [31] around the landmarks). This provides a more robust representation against illumination, which is crucial for detecting and tracking faces in driving scenarios.

Given an image $\mathbf{d} \in \mathbb{R}^{m\times 1}$ of m pixels, $\mathbf{d}(\mathbf{x}) \in \mathbb{R}^{p\times 1}$ (see footnote for notation)[1] indexes p landmarks in the image. $\mathbf{h}$ is a non-linear feature extraction function (in our case SIFT features) and $\mathbf{h}(\mathbf{d}(\mathbf{x})) \in \mathbb{R}^{128p\times 1}$ because SIFT features have 128 dimensions. During training, we will assume that the correct p landmarks are known, and we will refer to them as $\mathbf{x}_*$ [see Fig. 4(b)]. Also, to reproduce the testing scenario, we ran the face detector on the training images to provide an initial configuration of the landmarks $(\mathbf{x}_0)$, which corresponds to an average shape [see Fig. 4(a)]. Then, face alignment can be formulated as minimizing the following function over $\Delta\mathbf{x}$

$$f(\mathbf{x}_0 + \Delta\mathbf{x}) = \|\mathbf{h}\left(\mathbf{d}(\mathbf{x}_0 + \Delta\mathbf{x})\right) - \boldsymbol{\phi}_*\|_2^2 \tag{1}$$

where $\boldsymbol{\phi}_* = \mathbf{h}(\mathbf{d}(\mathbf{x}_*))$ represents the SIFT values in the manually labeled landmarks. In the training images, $\boldsymbol{\phi}_*$ and $\Delta\mathbf{x}$ are known.

One could use Newton's method to minimize Eq. (1). Newton's method makes the assumption that a smooth function, $f(x)$, can be well approximated by a quadratic function in a neighborhood of the minimum. If the Hessian is positive definite, the minimum can be found by solving a system of linear equations. The Newton updates to minimize Eq. (1) would be:

$$\mathbf{x}_k = \mathbf{x}_{k-1} - 2\mathbf{H}^{-1}\mathbf{J}_{\mathbf{h}}^{\top}(\boldsymbol{\phi}_{k-1} - \boldsymbol{\phi}_*) \tag{2}$$

where $\boldsymbol{\phi}_{k-1} = \mathbf{h}(\mathbf{d}(\mathbf{x}_{k-1}))$ is the feature vector extracted at the previous set of landmark locations, $\mathbf{x}_{k-1}$, $\mathbf{H}$ and $\mathbf{J}_{\mathbf{h}}$ are the Hessian and Jacobian evaluated at $\mathbf{x}_{k-1}$. Note that the SIFT operator is not differentiable and minimizing Eq. (1) using first or second order methods requires numerical approximations (e.g., finite differences) of the Jacobian and the Hessian. However, numerical approximations are very computationally expensive. Furthermore, $\boldsymbol{\phi}_*$ is known in training but unknown in testing. SDM addresses these issues by learning a series of descent directions and re-scaling factors (done by the Hessian in the case of Newton's method) such that it produces a sequence of updates $(\mathbf{x}_{k+1} = \mathbf{x}_k + \Delta\mathbf{x}_k)$ starting from $\mathbf{x}_0$ that converges to $\mathbf{x}_*$ in the training data. That is, SDM learns from training data a sequence of generic descent directions $\{\mathcal{R}_k\}$ and bias terms $\{\mathbf{b}_k\}$

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \mathcal{R}_{k-1}\boldsymbol{\phi}_{k-1} + \mathbf{b}_{k-1} \tag{3}$$

such that the succession of $\mathbf{x}_k$ converges to $\mathbf{x}_*$ for all images in the training set. For more details on SDM, see [44].

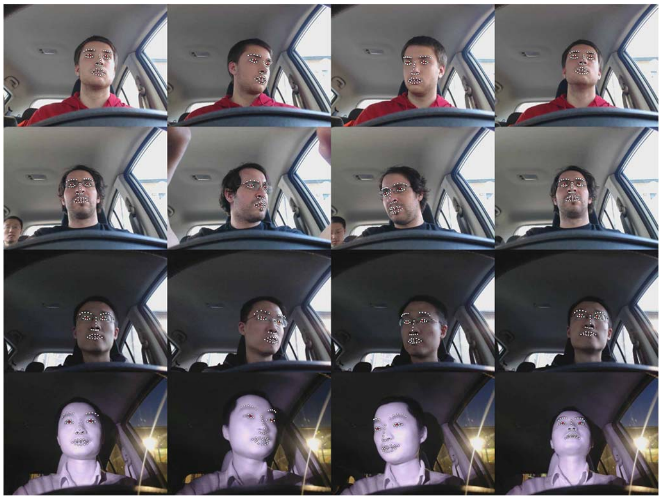


Fig. 5. SDM landmark detection under different poses, illuminations, and ethnicities.

Fig. 5 illustrates several examples of how the tracker works in real driving scenarios. The face tracking code is available at http://www.humansensing.cs.cmu.edu/intraface/.

C. Head Pose Estimation

In real driving scenarios, drivers change their head pose and facial expression while driving. Accurately estimating driver's head pose in complex situations is a challenging problem. In this section, a 3D head pose estimation system is proposed to decouple rigid and non-rigid head motion.

The head model is represented using a shape vector, $\mathbf{q} \in \mathbb{R}^{(3\cdot 49\times 1)}$, concatenating the x, y, z coordinates of all vertices. The deformable face model is constructed by computing PCA [9] on the training dataset from Cao *et al.* [10], which contains aligned 3D face shapes that have variation in both identity and expression. A new 3D shape can be reconstructed as a linear combination of eigenvectors $\mathbf{v}_i$ and the mean shape $\bar{\mathbf{q}}$

$$\mathbf{q} = \bar{\mathbf{q}} + \sum_i \beta_i \mathbf{v}_i = \bar{\mathbf{q}} + \mathbf{V}\beta. \tag{4}$$

Given 49 tracked 2D facial landmarks ($\mathbf{p}_k \in \mathbb{R}^{2\times 1}, k = 1, \ldots, 49$, excluding the pupil points) from the SDM tracker, we simultaneously fit the head shape and head pose by minimizing the difference between the 2D landmarks and the projection of the corresponding 3D points from the model. In this paper, we assume a weak-perspective camera model [43], also called scaled orthographic projection. The fitting error is defined as

$$E = \frac{1}{2}\sum_{k=1}^{K=49}\left\| s\mathbf{P}\left(\mathbf{R}\mathbf{L}_k\mathbf{q} + \mathbf{t}'_{head_p}\right) - \mathbf{p}_k\right\|_2^2 \tag{5}$$

where k is the index of the k-th facial landmark, $\mathbf{P} \in \mathbb{R}^{2\times 3}$ is a projection matrix, $\mathbf{L}_k \in \mathbb{R}^{3\times(3\cdot 49)}$ is the selection matrix that selects the vertex corresponding to the k-th facial landmark, $\mathbf{R} \in \mathbb{R}^{3\times 3}$ is a rotation matrix defined by the head pose angles, $\mathbf{t}'_{head_p} \in \mathbb{R}^{3\times 1}$ is a 3D translational vector of the driver's head relative to the camera's optical center, and s is a scale factor approximating the perspective image formation. The overall fitting error, E, which is the total fitting error of all landmarks,

[1] Bold capital letters denote a matrix $\mathbf{X}$, bold lower-case letters a column vector $\mathbf{x}$. $\mathbf{x}_i$ represents the ith column of the matrix $\mathbf{X}$. x_{ij} denotes the scalar in the ith row and jth column of the matrix $\mathbf{X}$. All non-bold letters represent scalars. $\mathbf{1}_{m\times n}, \mathbf{0}_{m\times n} \in \mathbb{R}^{m\times n}$ are matrices of ones and zeros. $\mathbf{I}_n \in \mathbb{R}^{n\times n}$ is an identity matrix. $\|\mathbf{x}\|_p = \sqrt[p]{\sum_i |x_i|^p}$ and $\|\mathbf{X}\|_F^2 = \sum_{ij} x_{ij}^2$ denote the p-norm of a vector and the Frobenius norm of a matrix, respectively.

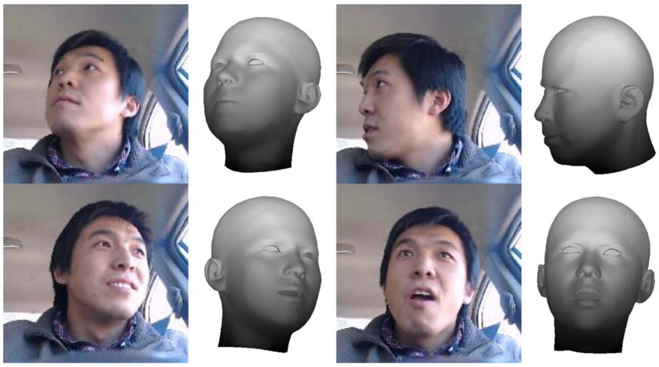


Fig. 6. Examples of 3D reconstruction and head pose estimation in the car environment.

is minimized with respect to the pose parameters $(\mathbf{R}, \mathbf{t}'_{head_p}, s)$ and shape coefficients, β, using an alternating optimization approach. We alternate between the estimation of the rigid parameters, $\mathbf{R}$ and s, and the non-rigid parameter β. These steps monotonically reduce the fitting error E, and because the function is bounded below, we converge to a critical point, see [45]. Fig. 6 shows four examples of the head pose estimation and 3D head reconstruction.

Translation Mapping: In the method described above, the head translation vector, $\mathbf{t}'_{head_p}$, is computed in pixels. To compute the EOR estimation using the geometry of the scene, we need to map pixels to centimeters, obtaining the vector $\mathbf{t}'_{head}$. This mapping is performed using a data-driven approach. For a fixed focal length, we collected a set of translation vectors in pixels and their corresponding real length in centimeters. A linear system is solved to obtain the unit mapping

$$\mathbf{t}'_{head} = \mathbf{A}\mathbf{t}'_{head_p} + \mathbf{b} \qquad (6)$$

where $\mathbf{A} \in \mathbb{R}^{3\times3}$ and $\mathbf{b} \in \mathbb{R}^{3\times1}$ are linear coefficients to solve. It is worth noting that this affine mapping was done outside the car environment and generalized well for several cameras of the same model (Logitech c920). In order to compute this mapping, we used three different individuals that were not in the testing set described in the experimental section.

Angular Compensation: As stated above, the head pose estimation algorithm is based on the weak-perspective assumption, hence we use a scaled orthographic projection. This assumption is accurate when the driver's head is near the visual axis of the camera. However, in the car scenario, due to several factors, such as the driver's height, seating position, and the fixed tilt of the camera, the driver's head may not always be close to the visual axis.

Fig. 7 shows three different positions of the driver's head and their corresponding images generated by the camera, produced by a perspective projection, and a scaled orthographic projection, respectively. As we can see, the camera view and the scaled orthographic projection image match in the position b, where the driver's head is aligned with the camera visual axis. For positions a and c, camera images and weak perspective images differ due to a translation with respect to the camera visual axis. As a result, the head pose estimation algorithm will introduce an offset caused by the driver's head position. To correct this offset, it is necessary to introduce a heuristic compensation for the head rotation. Fig. 7 shows the heuristic compensation computed for a lateral translation that affects the estimation of the yaw angle. A similar compensation is computed for the pitch angle for a vertical translation. As a result, the yaw and pith angles estimated by our head pose algorithm in the car environment are given by

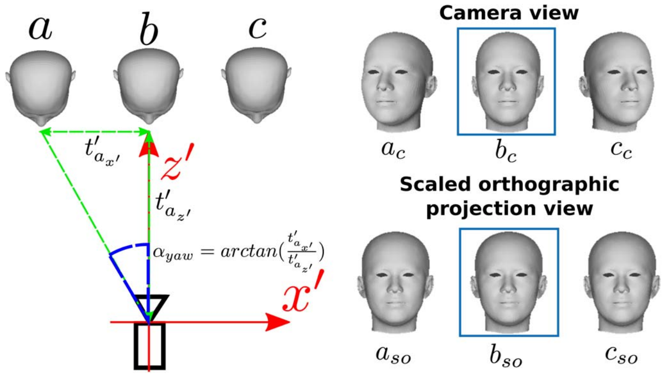


Fig. 7. Difference between camera image and weak perspective projection image caused by translation of the driver's head with respect the camera virtual axis. Top row (a_c, b_c, and c_c) shows the different images generated in the camera for positions a, b, c. Bottom row shows the weak perspective projection images using the same translations. When the face is near the visual axis (blue rectangle), the weak perspective projection image is similar to the real image captured by the camera. When the face is off the visual axis, there will be discrepancy between these two images.

$$\phi'^{head}_{yaw} = \gamma_{yaw} - \alpha_{yaw} \qquad (7)$$

$$\phi'^{head}_{pitch} = \gamma_{pitch} - \alpha_{pitch} \qquad (8)$$

where γ_{yaw} and γ_{pitch} are the original yaw and pitch angles computed by our algorithm, and α_{yaw} and α_{pitch} are the corresponding compensation angles. Note that no compensation was introduced for the roll angle, our tests showed that the roll angle was less sensitive to translations with respect to the camera virtual axis.

D. Gaze Estimation

The driver's gaze direction provides crucial information as to whether the driver is distracted or not. Gaze estimation has been a long standing problem in computer vision [23], [25]. Most existing work follows a model-based approach to gaze estimation that assumes a 3D eye model, where the eye center is the origin of the gaze ray. In this paper, we used a similar model (see Fig. 8). We make three main assumptions: First, the eyeball is spherical and thus the eye center is at a fixed point (rigid point) relative to the head model; Second, all the eye points, including the pupil, are detected using the SDM tracker described in the previous section. Note that more accurate pupil center estimates are possible using other techniques such as the Hough transform; Third, the eye is open and therefore all the eye contour points can be considered rigid. Our algorithm has two main parts: (1) Estimate the 3D position of the pupil from the rigid eye contour points, and (2) estimate the 3D gaze direction from the pupil position and the eye center.

The 3D position of the pupil is computed as follows:

1) Triangulate the eye contour points in 2D and determine which triangle mesh contains the pupil. See Fig. 8.

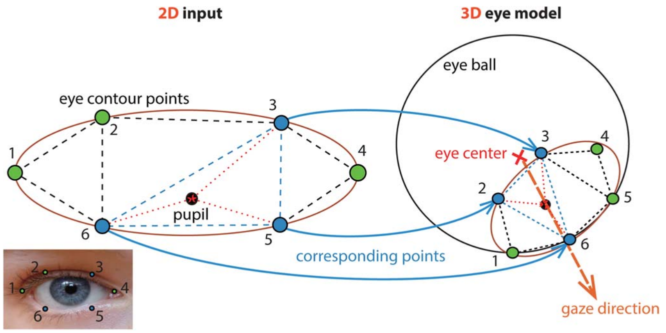


Fig. 8. 3D Gaze estimation. We first triangulate the eye contour points to get the eye mesh. The tracked 2D pupil (black dot) landmark is assigned to the closest triangle (blue mesh). Using the correspondences between the 2D and the 3D mesh, the 3D pupil point is computed in the barycentric coordinate system of the triangle. The 3D gaze direction can then be determined using the eye center point and the pupil point.

2) Compute the barycentric coordinates of the pupil inside the triangle mesh that contains the pupil.
3) Apply the barycentric coordinates to the corresponding eye contour points in 3D to get the 3D position of pupil.

After we obtain the 3D position of the pupil, the gaze direction can be simply estimated as the ray that goes through the 3D eye center and the 3D pupil. We can thus obtain the gaze angles.

Finally, to compute the gaze angles with respect to the camera coordinate system, $\Phi'^{gaze} = (\phi'^{gaze}_{yaw}, \phi'^{gaze}_{pitch})$, it is necessary to rotate the estimated gaze angles using the compensated head pose rotation matrix.

E. Eyes Off the Road Detection

The EOR estimation is based on a 3D ray tracing method that uses the geometry of the scene as described in Fig. 9. Our EOR estimation algorithm computes the point where the driver's 3D gaze line, $\mathbf{v}_{gaze}$ in Fig. 9, intersects the car windshield plane Π. If the intersection point lies outside of the defined on-the-road area, an alarm is triggered. In our approach, we only used the gaze from the driver's left eye since it suffers from less occlusion (only short head movements to check the driver mirror) than the right eye.

To compute the 3D gaze vector, we need the 3D position of the eye and the gaze direction (gaze yaw and pitch angles). Let O' and O be the origins of the camera coordinate system, (x', y', z'), and the world coordinate system, (x, y, z), respectively. Both systems are measured in centimeters. The world coordinate system is the camera coordinate system rotated by the camera tilt γ_{tilt}, it follows that $O = O'$. The relation between the point P, in the world coordinate system, and the point P' in the camera coordinate system is expressed by $P = \mathbf{R}_{c/w}P'$, where $\mathbf{R}_{c/w}$ is the rotation matrix from the camera coordinate system to the world coordinate system. This rotation matrix is defined by the camera tilt, γ_{tilt}, see Fig. 9.

The gaze direction algorithm described in Section III-D, provides the driver's gaze direction, $\Phi'^{gaze} = (\phi'^{gaze}_{yaw}, \phi'^{gaze}_{pitch})$, with respect to the camera coordinate system. We can build the 3D gaze vector, $\mathbf{u}'_{gaze}$, as

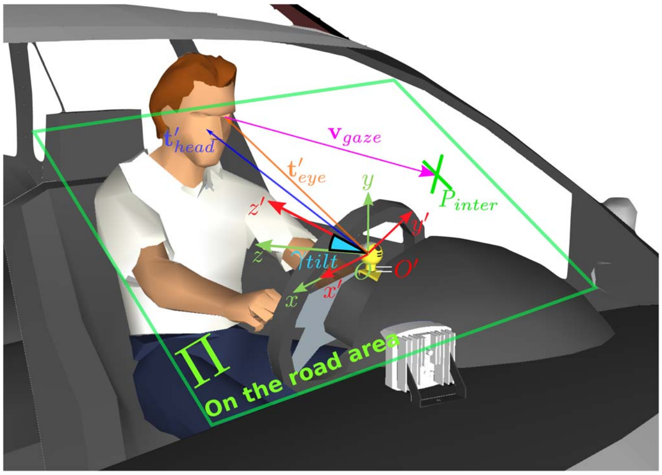


Fig. 9. Geometric analysis for EOR estimation.

$$\mathbf{u}'_{gaze} = \begin{bmatrix} \cos\left(\phi'^{gaze}_{pitch}\right) \cdot \sin\left(\phi'^{gaze}_{yaw}\right) \\ \sin\left(\phi'^{gaze}_{pitch}\right) \\ -\cos\left(\phi'^{gaze}_{pitch}\right) \cdot \cos\left(\phi'^{gaze}_{yaw}\right) \end{bmatrix}. \tag{9}$$

Using the 3D head coordinates, $\mathbf{q}$ in Eq. (4), our head pose algorithm estimates the 3D position of the driver's head and eyes with respect to the camera coordinate system, vectors $\mathbf{t}'_{head}$ and $\mathbf{t}'_{eye}$ respectively. Hence, the 3D gaze line can be expressed using the parametric 3D line form as

$$\mathbf{v}_{gaze} = \mathbf{R}_{c/w}\left(\mathbf{t}'_{eye} + \lambda \mathbf{u}'_{gaze}\right). \tag{10}$$

Finally, the intersection point, P_{inter}, is given by the intersection of the gaze vector, $\mathbf{v}_{gaze}$, with the windshield plane Π. The equation of the windshield plane Π in the world coordinate system is estimated using least squares plane fitting.

F. Sunglasses Detector

Our system works reliably with drivers of different ethnicities wearing different types of glasses. However, if the driver is wearing sunglasses, it is not possible to robustly detect the pupil. Thus, to produce a reliable EOR estimation in this situation, the vector $\mathbf{v}_{gaze}$ will be computed using the head pose angles.

The sunglasses detection pipeline is shown in Fig. 10. First, our system extracts SIFT descriptors from the area of the eyes and eyebrows, $\mathbf{h_1}, \ldots, \mathbf{h_n}$, and concatenates them to build the feature vector $\mathbf{\Psi}$. Second, a linear Support Vector Machine (SVM) classifier is used to estimate if the driver is wearing sunglasses. The SVM classifier has been trained using 7500 images the databases CMU Multi-PIE face database [20] and the PubFig database [28]. The classifier obtained 98% accuracy in the test set, which was composed of 600 images evenly distributed between positive and negative classes.

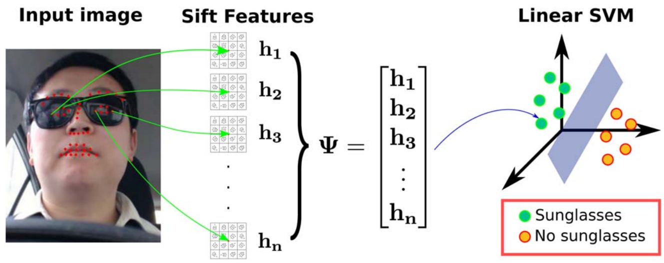


Fig. 10. Sunglasses classifier pipeline.

IV. Experiments

This section evaluates the accuracy of our system in different tasks. First, we compare our head pose estimation to other state-of-the-art approaches. Second, we report the performance of our EOR detection system in videos recorded in the car environment. Finally, we evaluate the robustness of the head pose estimation algorithm to extreme facial deformations.

A. Head Pose Estimation

To evaluate our 3D-based head pose estimation algorithm, we used the Boston University (BU) dataset provided by La Cascia *et al.* [11]. This dataset contains 45 video sequences from 5 different people with 200 frames in each video. As described in the previous sections, facial feature detection is performed in each input frame. Using the 2D tracked landmarks, we estimated the 3D head orientation and translation. The distance units were pre-normalized to ensure that translation metrics were in the same scale.

We compared our head pose estimation system with deformable 3D head model (3D-Deform) against four other methods in the literature. Table I shows the Mean Absolute Error (MAE) of the head pose angles for different algorithms. La Cascia *et al.* [11] reported a method that used a manually initialized cylindrical model and recursive least squares optimization. Sung *et al.* [40] proposed a method that was based on active appearance model [12] and cylinder head models. Valenti *et al.* [42] used a cylindrical model and a template update scheme to estimate model parameters on-the-fly. As Table I shows, our method based on a deformable 3D face model is more accurate than these three methods. Finally, Saragih *et al.* [38] proposed an algorithm based on a 3D version Constrained Local Model (CLM) which estimates the pose parameters by maximizing the local patch response using linear SVMs with logistic regressors. This method obtained better MAE than our method, however this algorithm requires a recalibration procedure using the ground truth when a large drift occurs, which is infeasible in the real car environment.

To demonstrate the robustness of our head pose estimation system against expressions, we conducted experiments in the car environment with different expressions settings. See Section IV-C for details. Fig. 11 illustrates how our head pose estimation algorithm works in the car environment. We can see how the yaw angle of the driver's head pose varies as the driver moves his head.

TABLE I
Head Pose Estimation Results on the BU Dataset, Measured in MAE (Mean Absolute Error)

Method	Yaw	Pitch	Roll	Mean
La Cascia et al. [11]	3.3	6.6	9.8	6.4
Sung et al. [40]	5.4	5.6	3.1	4.7
Valenti et al. [42]	6.6	6.4	4.2	5.6
Saragih et al. [38]	5.2	4.5	2.6	4.1
3D-Deform (Ours)	4.3	6.2	3.2	4.6

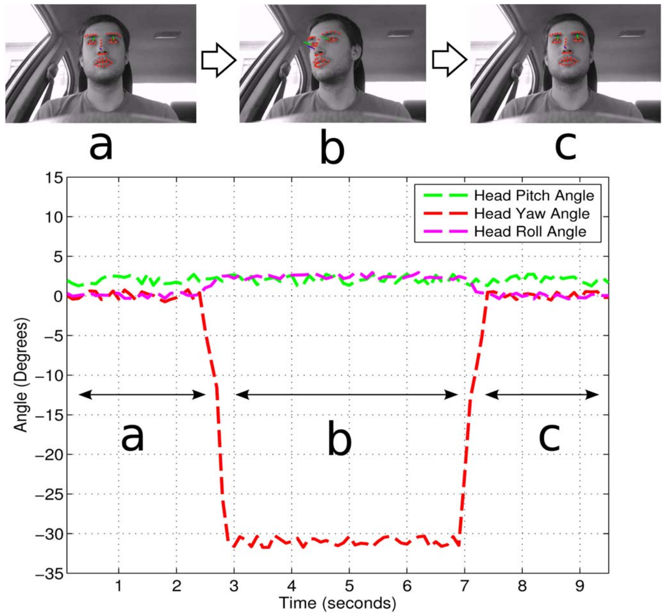


Fig. 11. Head pose estimation, yaw, pitch, and roll angles. Section *a*: Driver looks ahead. Section *b*: Driver looks at the driver mirror. Section *c*: Driver recovers initial position.

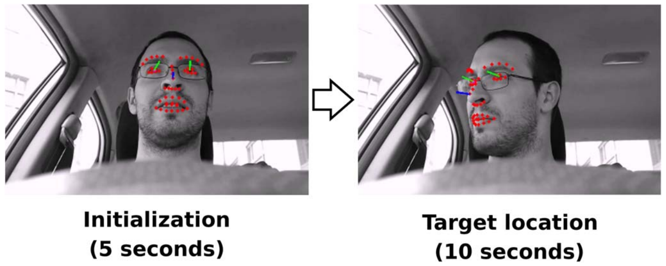


Fig. 12. Evaluation protocol for the EOR detection.

B. Eyes Off/On the Road Estimation

This section reports experimental results of our EOR system in a real car scenario. First, we provide details of the evaluation protocol and dataset that was used to evaluate the system performance. Then, we present the performance analysis.

1) Evaluation Protocol and Dataset: In order to evaluate the EOR performance of our system, we selected four on-the-road and fourteen off-the-road locations in the car interior and windshield. Fig. 13 shows the size of the on-the-road area and the selected locations. Red dots are considered

TABLE II
SYSTEM EOR OVERALL ACCURACY FOR DIFFERENT SCENARIOS

Target area	**Day time accuracy**				**Night time accuracy**		
	Total %	**No-glasses %**	**Glasses %**	**Sunglasses %**	**Total %**	**No-Glasses %**	**Glasses %**
On-the-road area	97.28	97.51	98.03	95.94	98.49	98.09	98.84
Off-the-road area	94.49	98.63	93.45	91.91	96.25	97.98	94.81

TABLE III
DAY TIME SYSTEM EOR ACCURACY FOR THE 18 TARGET LOCATIONS

Target location	**Day time accuracy results**			
	Total %	**No-glasses %**	**Glasses %**	**Sunglasses %**
1. Driver window	97.95	100	98.22	95.51
2. Driver mirror	99.31	100	100	97.59
3. Driver visor	80.68	100	80.34	61.87
4. Passenger visor	87.20	84.88	80.23	100
5. Passenger window	100	100	100	100
6. Passenger mirror	99.97	100	99.94	100
7. Glove box	100	100	100	100
8. Navigation System	97.83	100	99.72	92.85
9. Steering wheel	91.05	96.59	83.33	97.09
10. Windshield top left	93.85	99.33	95.29	86.21
11. Windshield top center	90.15	100	82.94	91.14
12. Windshield top right	99.42	100	100	98.00
13. Windshield left	98.30	98.75	98.44	97.67
14. Windshield center	99.68	99.75	99.44	100
15. Windshield right	99.24	100	100	97.34
16. Windshield bottom left	94.13	92.27	94.24	95.84
17. Windshield bottom	97.00	99.25	100	90.28
18. Windshield bottom right	86.16	100	88.26	69.18

off-the-road gaze locations and green dots are considered on-the-road gaze locations. We recorded several videos under different conditions where the driver is looking to these locations, and compute the percentage of times that the system correctly predicted on-the-road and off-the-road.

Drivers were asked to follow a scripted series of actions specifying the direction of their gaze. Initially, drivers were looking straight to the windshield. At an audio signal, drivers were instructed to look naturally to one of the target locations shown in Fig. 13 for a period of ten seconds. For users with sunglasses, drivers were encouraged to orient their head towards the target location. Fig. 12 illustrates this process. Experiments were conducted using a total of twelve different individuals covering a wide spectrum of facial appearances and illuminations. The population was composed of eight Asians and four Caucasians. For day time experiments, four individuals were wearing no-glasses, six were wearing glasses, and four sunglasses. For the night experiments, six subjects had glasses and five no-glasses. As a performance measure, we used the percentage of frames that were correctly predicted in terms of eyes on/off-the-road during the ten seconds period where the subject was looking at a particular location. A total of $\sim$135 000 frames of day and night experiments were used to evaluate the on/off-the-road accuracy estimation (25 sessions × 18 locations × 10 seconds × 30 FPS). Notice that the frames generated during the initialization stage of five second are not used to compute the EOR accuracy. For locations inside the off-the-road area, we reported the percentage of frames predicted as EOR. On the contrary, for locations placed on-the-road area, we report the percentage of on-the-road predictions.

The EOR system runs in MATLAB ($\sim$25 FPS) with an image resolution of 1280 × 720 pixels, using an average of 38.4 ms to process every frame on a Intel i5 processor with 2.6 GHz using 350 MB of RAM.

2) Experimental Results: Table II shows the overall performance of the system under different scenarios. The accuracy of the system for the on-the-road area is above 95%, hence, the system exhibits a low false alarm rate (below 5%) for all scenarios. This is a very desirable feature for EOR detection systems, because drivers will not be disturbed by unnecessary sound alerts fired by the system. Similarly, the accuracy of the system in the off-the-road area is above 90% for all scenarios. Moreover, there is no significant difference between the night and day time results. The IR illuminator effectively illuminates the driver's face, allowing the system to accurately track the driver's facial landmarks, see Fig. 16. Furthermore, SIFT features are robust to illumination changes, allowing the system to track the driver's face in scenarios with reduced illumination.

Table III shows the day time accuracy results obtained for each one of the 18 target areas for no-glasses, glasses, and sunglasses. Overall, the system achieved high accuracy for 15 out of the 18 evaluated target areas shown in Fig. 13. The system achieved an accuracy below 90% for the following positions: Driver visor, passenger visor, and windshield bottom right with 80.68%, 87.20%, and 86.16%, respectively. The drop of the system performance for these positions is caused by facial feature tracking problems due to extreme driver head poses. Furthermore, this problem is exacerbated for drivers wearing glasses, due to occlusions introduced by the glasses frame.

For the driver visor position, the system achieved a perfect score for drivers that were not wearing glasses. On the other hand, for drivers wearing glasses, the system performance

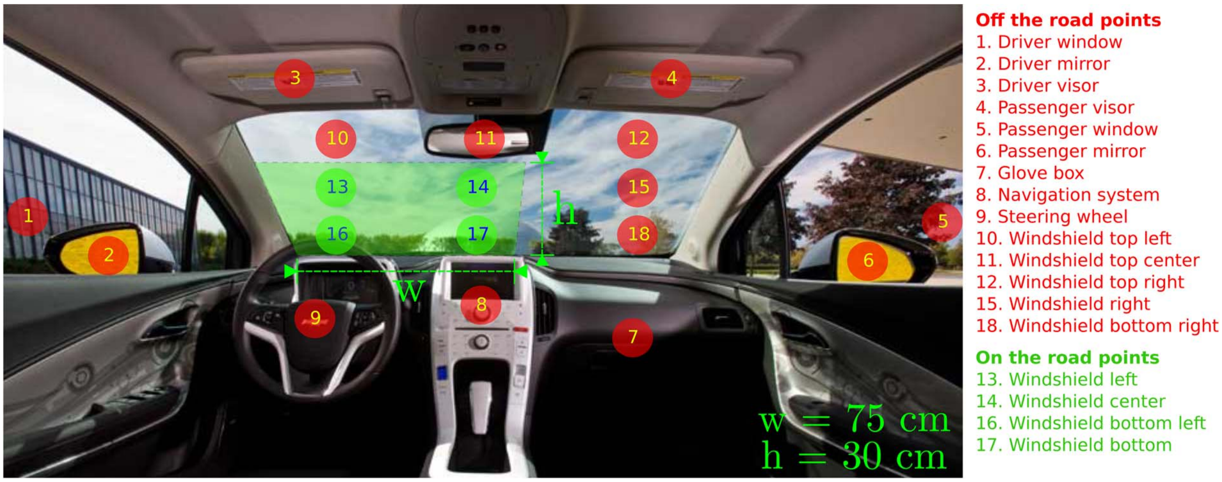
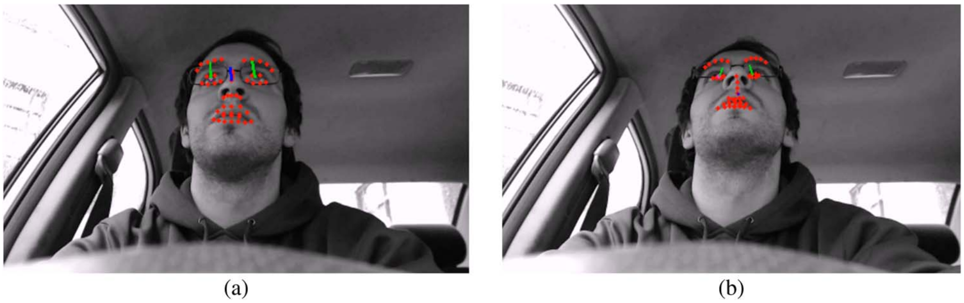
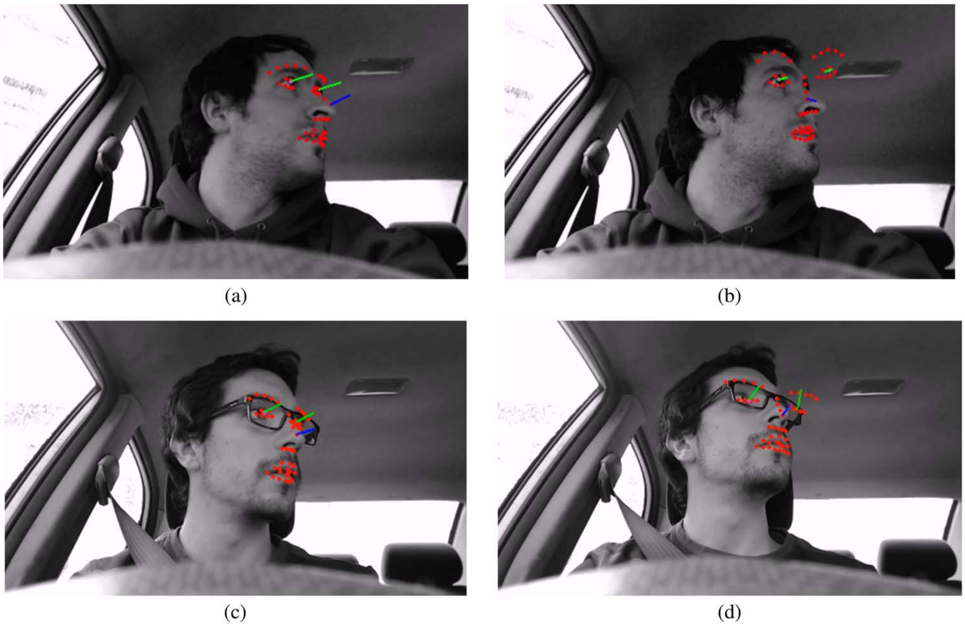


Fig. 13. Target areas and dimensions of the on-the-road area. The height $(\mathbf{h})$ and width $(\mathbf{w})$ of the on-the-road area are defined in the x and y axis of the world coordinate system defined in Fig. 9.


(a)

(b)

Fig. 14. Illustration of tracking failure due to extreme pose and glasses. Green lines starting at driver's eyes show the estimated gaze direction. The blue line starting at the driver's nose shows the estimated head pose orientation. a) Driver looking at the windshield top left. b) Driver looking at the driver visor position.


(a)

(b)

(c)

(d)

Fig. 15. Examples of drivers looking at the passenger visor location. a) and c) show the driver's face successfully tracked, accurate head pose and gaze estimation. b) and d) illustrate facial landmarks tracking failures caused by an extreme pose and the glasses frame, respectively.

dropped to 80.34% due to errors in the facial landmark tracking caused by the glasses frames. Errors in facial landmark tracking resulted in unreliable head pose and gaze estimation. Fig. 14 illustrates a tracking failure caused by the glasses frames as the driver moves his head to look at the driver visor using an extreme head pose. Note that in Fig. 14(b), the eye landmarks are placed on the glasses frame and thus the pupils are poorly tracked. For drivers wearing sunglasses, the system achieved

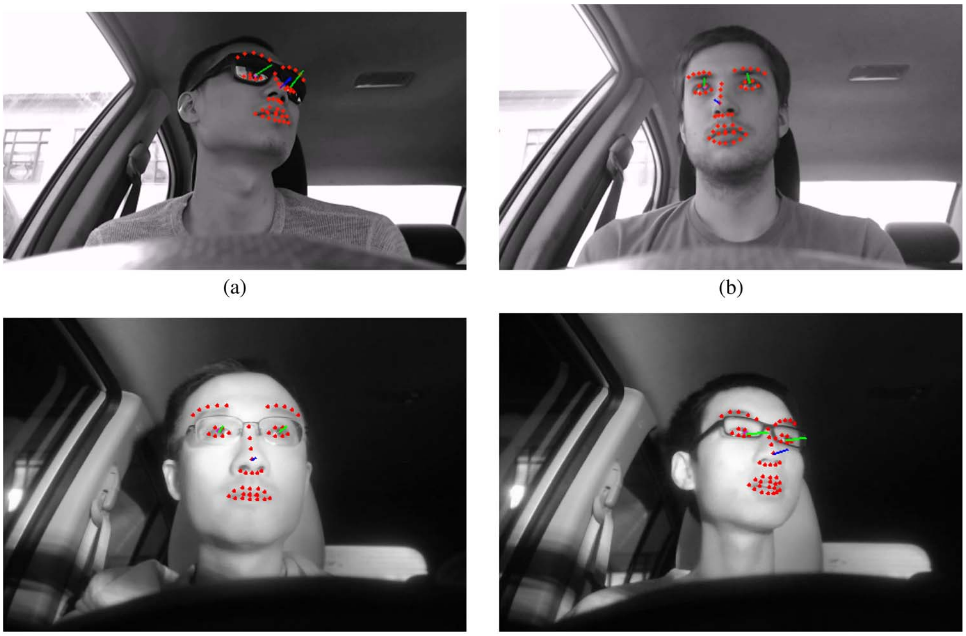

(a) (b) (c) (d)

Fig. 16. Sample images captured during day and night time. Green lines starting from the eyes show the gaze of the driver. The blue line starting from the nose show the head pose direction. a) shows the EOR system working when the driver is wearing sunglasses. Recall that in this case the EOR estimation is computed using the head pose angles (blue line); b) shows the system working during day time, the driver is not wearing glasses; c) and d) show the system operating during night time with drivers with glasses.

an accuracy of 61.87% due to severe tracking errors. Consequently, our EOR estimation based on head pose alone was not reliable enough to achieve high accuracy for the driver visor position. Similarly to the driver visor position, the EOR accuracy for windshield bottom right location is affected by non-reliable facial feature tracking for glasses and sunglasses subjects, obtaining an accuracy value of 86.16%.

In the case of the passenger visor location, the system achieved an overall day time accuracy of 87.20%. For drivers wearing no-glasses and glasses, our EOR estimation system achieved similar performance, 84.88% and 80.23%, respectively. Similarly to the driver visor and the windshield bottom right positions, errors in the facial landmarks led to mispredictions in the driver's head pose and gaze direction, see Fig. 15. However, for drivers wearing sunglasses, the system did not exhibit facial feature tracking problems [see Fig. 16(a)], obtaining a perfect EOR estimation accuracy.

Table IV shows the night time system performance. Again, our system achieved above 90% accuracy for 15 out of the 18 evaluated positions. The three locations below 90% accuracy were: Passenger visor, windshield top left, and windshield bottom right; Achieving 86.25%, 89.87%, and 89.70%, respectively. Notice that our system also achieved an accuracy below 90% the positions passenger visor and windshield bottom right in the day time experiments. During the night time experiments similar tracking problems were present for drivers wearing no-glasses and glasses as they appeared in the day time. Hence, the night time system behavior was consistent with the day time results, due to the efficient illumination provided by the IR illumination. However, according to Table II, the accuracy of the system for both on-the-road and off-the-road areas is slightly higher (∼1% for day and night time) during night time.

TABLE IV
NIGHT TIME SYSTEM EOR ACCURACY FOR THE 18 TARGET LOCATIONS

Target location	Night time accuracy results		
	Total %	No-glasses %	Glasses %
1. Driver window	99.78	99.53	100
2. Driver mirror	100	100	100
3. Driver visor	92.59	99.13	87.15
4. Passenger visor	86.25	87.17	85.49
5. Passenger window	97.72	95.14	99.88
6. Passenger mirror	99.84	99.66	100
7. Glove box	98.42	96.54	100
8. Navigation System	100	100	100
9. Steering wheel	100	100	100
10. Windshield top left	89.87	99.86	81.56
11. Windshield top center	97.96	97.40	98.44
12. Windshield top right	98.90	97.67	99.94
13. Windshield left	98.15	95.94	100
14. Windshield center	97.90	98.40	97.50
15. Windshield right	96.37	100	93.35
16. Windshield bottom left	99.27	98.40	100
17. Windshield bottom	98.64	99.60	97.84
18. Windshield bottom right	89.70	99.60	81.45

Fig. 16 shows image samples captured during day and night time experiments.

C. Robustness of Head Pose Across Expression Changes

This section describes an experiment to evaluate the robustness of the head pose estimation against extreme facial expression.

1) Evaluation Protocol and Dataset: We evaluated the head pose estimation algorithm under three different exaggerated facial expressions: mouth open, smile, and dynamic facial expression deformation. In mouth open and smile, the subject kept this expression for the entirety of the ten seconds recordings.

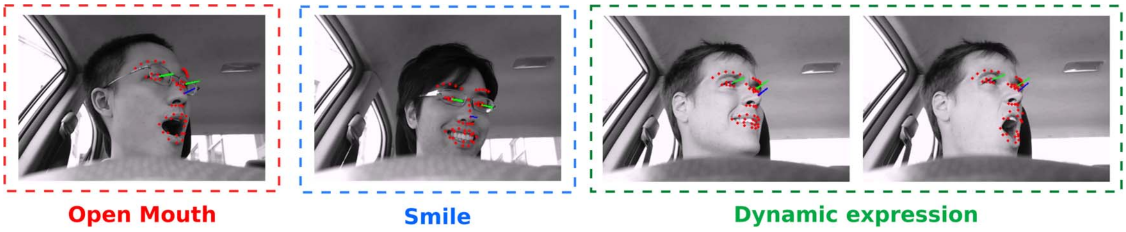


Fig. 17. Examples of the three facial expressions under evaluation.

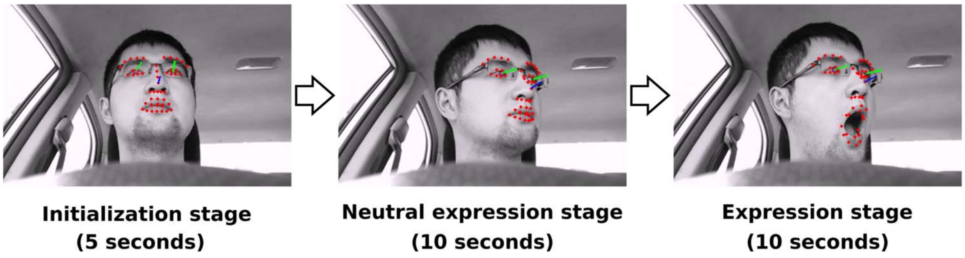


Fig. 18. Evaluation protocol to measure the facial expression impact on head pose estimation.

Whereas, in the dynamic facial expression the subject randomly and continuously moves the mouth (e.g., talks, open mouth, smile). Fig. 17 shows examples of these deformations.

The subjects were asked to look in the windshield for five seconds as an initialization step. Later, subjects were instructed to look to several particular positions within the car. We selected a subset of the target locations shown in Fig. 13. Namely, we used the subset of nine positions: $S = \{2, 6, 8, 10, 11, 12, 14, 16, 18\}$. The remaining locations were redundant for this particular experiment. The subject first looked to a particular location for ten seconds, while maintaining a neutral expression. The subject then performed one of the three expressions, and we recorded her for another ten seconds. In these 20 seconds, the subjects did not change their head pose. In total, for every subject looking at a location we had a 25 seconds video (including the initialization stage). Fig. 18 illustrates this data collection process. As in the previous experiment, we did not use the frames of the initialization stage to compute the performance measure.

For every expression, we recorded five different subjects. The population ethnicity distribution was comprised of six Asians and four Caucasians. Moreover, seven out of ten individuals wore glasses. In total, we had ~81 000 frames (3 expressions × 5 subjects × 20 seconds × 30 FPS × 9 positions). We evaluated the deviation in the head pose angles when the expression occurred. That is, we computed the difference in the mean head pose angles between the ten seconds of the neutral expression stage, and the ten seconds of the expression stage. We computed the Mean Absolute Error (MAE), the standard deviation for the neutral ($\overline{\sigma}_{NoExp}$), and the expression ($\overline{\sigma}_{Exp}$) stages.

2) *Experiment Results:* Tables V–VII summarize the results in this experiment.

Table V shows how the head pose estimation is robust against wide open mouth expressions. According to the MAE measures, roll estimation is minimally affected by this expression, while yaw and pitch angles are more sensitive. This is a common effect across all three facial expressions under evaluation. Yaw estimation suffered the highest MAE for the top right and windshield top center positions, with 5.51 and 4.96 degrees of deviation. In the pitch angle, the maximum deviation occurred for the driver mirror and windshield top left, with a MAE of 4.30 and 4.22, respectively. Errors in yaw and pitch estimation were induced by problems in the tracking of the facial landmarks. Incorrect landmark estimation produced a corrupt 3D model of the driver head, hence errors in head pose estimation. However, the variance in the head pose estimation during the no expression ($\overline{\sigma}_{NoExp}$) and expression ($\overline{\sigma}_{Exp}$) stages did not exhibit any significant difference. Recall that the open mouth is a static expression, that is, the driver kept this facial expression during the ten seconds of the expression stage.

Table VI shows the results obtained for the smile expression. Similar to the case of the mouth wide open expression, there is no significant difference in the variance of the estimated head pose angles. In this case, the smile expression is easier to track and the tracker got lost less frequently. This resulted in MAE estimates with less error and variance, except for a high MAE in the yaw estimation for the navigation system location.

Table VII shows the results obtained for drivers performing dynamic expressions. As we can see, the amount of variation in the head pose estimation has increased remarkably for all the target locations during the expression stage ($\overline{\sigma}_{Exp}$). This is caused by large changes in individuals' facial expressions during this test. However, we can see that the maximum absolute error for yaw, pitch, and roll angles was similar to the maximum absolute error of the previously studied facial expressions. The larger variance in the estimation of the head pose angles is caused by tracking instabilities while users changed abruptly from one expression to another.

TABLE V
IMPACT OF THE OPEN MOUTH EXPRESSION ON HEAD POSE ESTIMATION

Position	Open mouth								
	Yaw			Pitch			Roll		
	MAE	$\overline{\sigma}_{NoExp}$	$\overline{\sigma}_{Exp}$	MAE	$\overline{\sigma}_{NoExp}$	$\overline{\sigma}_{Exp}$	MAE	$\overline{\sigma}_{NoExp}$	$\overline{\sigma}_{Exp}$
2. Driver mirror	3.12	0.90	0.36	4.30	1.38	0.39	1.78	0.71	0.31
6. Passenger mirror	1.52	1.22	0.44	3.79	1.50	0.68	1.60	0.75	0.42
8. Navigation system	2.88	1.05	0.64	2.96	1.20	0.59	1.77	0.54	0.35
10. Windshield top left	1.43	0.56	0.52	4.22	1.46	0.56	0.48	0.41	0.36
11. Windshield top center	5.51	1.98	1.77	2.82	1.81	0.72	1.16	0.80	0.34
12. Windshield top right	4.96	1.72	0.56	3.48	1.46	0.76	2.31	0.73	0.41
14. Windshield center	2.31	0.73	0.83	2.63	0.76	0.68	0.57	0.26	0.27
16. Windshield bottom left	1.57	0.59	0.36	3.01	1.20	0.76	0.78	0.40	0.27
18. Windshield bottom right	2.10	0.89	0.61	3.30	1.17	0.46	1.23	0.55	0.31

TABLE VI
IMPACT OF THE SMILE EXPRESSION ON HEAD POSE ESTIMATION

Position	Smile								
	Yaw			Pitch			Roll		
	MAE	$\overline{\sigma}_{NoExp}$	$\overline{\sigma}_{Exp}$	MAE	$\overline{\sigma}_{NoExp}$	$\overline{\sigma}_{Exp}$	MAE	$\overline{\sigma}_{NoExp}$	$\overline{\sigma}_{Exp}$
2. Driver mirror	0.90	0.64	0.90	2.28	1.06	0.55	1.68	0.66	0.53
6. Passenger mirror	0.74	1.00	0.56	1.08	1.11	0.79	1.23	0.64	0.59
8. Navigation system	5.06	1.30	0.72	3.56	1.05	0.48	1.42	0.70	0.31
10. Windshield top left	1.09	0.42	0.45	2.04	0.58	0.45	0.59	0.26	0.24
11. Windshield top center	1.69	1.00	0.71	2.03	1.10	0.80	1.82	0.77	0.35
12. Windshield top right	2.38	0.69	0.88	1.70	0.87	0.53	1.25	0.37	0.66
14. Windshield center	1.19	0.66	0.56	1.83	0.62	0.46	1.26	0.32	0.28
16. Windshield bottom left	0.77	0.48	0.38	1.15	0.75	0.59	0.68	0.26	0.20
18. Windshield bottom right	2.42	0.88	0.49	2.13	1.03	0.55	1.38	0.61	0.44

TABLE VII
IMPACT OF THE DYNAMIC EXPRESSION ON HEAD POSE ESTIMATION

Position	Dynamic expression								
	Yaw			Pitch			Roll		
	MAE	$\overline{\sigma}_{NoExp}$	$\overline{\sigma}_{Exp}$	MAE	$\overline{\sigma}_{NoExp}$	$\overline{\sigma}_{Exp}$	MAE	$\overline{\sigma}_{NoExp}$	$\overline{\sigma}_{Exp}$
2. Driver mirror	2.28	0.48	1.67	1.34	0.47	1.24	0.68	0.33	0.73
6. Passenger mirror	1.42	0.94	2.23	2.35	0.93	2.25	1.87	0.54	1.32
8. Navigation system	1.94	0.85	1.85	2.30	1.00	0.98	1.84	0.65	0.45
10. Windshield top left	0.91	0.67	1.12	2.14	0.75	1.28	1.07	0.61	0.72
11. Windshield top center	3.04	2.56	2.99	1.79	1.86	1.61	1.19	0.82	1.09
12. Windshield top right	2.09	1.51	2.72	1.51	1.07	1.76	0.69	0.72	1.22
14. Windshield center	1.75	0.87	1.33	0.97	0.75	1.30	0.70	0.25	0.47
16. Windshield bottom left	1.28	0.62	1.21	1.32	0.83	1.27	0.53	0.53	0.45
18. Windshield bottom right	1.34	0.95	1.13	1.58	0.96	1.49	1.22	0.54	0.64

V. CONCLUSION

This paper describes a real-time EOR system using the video from a monocular camera installed on steering wheel column. Three are the main novelties of the proposed system: (1) Robust face landmark tracker based on the Supervised Descent Method, (2) accurate estimation of 3D driver pose, position, and gaze direction robust to non-rigid facial deformations, (3) 3D analysis of car/driver geometry for EOR prediction. The proposed system is able to detect EOR at day and night, and under a wide range of driver's characteristics (e.g., glasses/sunglasses/no glasses, ethnicities, ages, . . .). The system does not require specific calibration or manual initialization. More importantly, no major re-calibration is necessary if the camera position is changed or if we re-define a new on-the-road area. This is due to the explicit use of 3D geometric reasoning. Hence, the installation of the system in different car models does not require any additional theoretical development.

The system achieved an accuracy above 90 % for all of the scenarios evaluated, including night time operation. In addition, the false alarm rate in the on-the-road area is below 5 %. Our experiments showed that our head pose estimation algorithm is robust to extreme facial deformations. While our system provided encouraging results, we expect that improving the facial feature detection in challenging situations (e.g., profile faces, faces with glasses with thick frames) will boost the performance of our system. Currently, we are also working on improving the pupil detection using Hough transform-based techniques to further improve the gaze estimation.

ACKNOWLEDGMENT

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We would like to thank the students of the Human Sensing Laboratory for their time and patience for recording the videos between 2011 and 2013.

REFERENCES

[1] [Online]. Available: http://www.distraction.gov/content/get-the-facts/facts-and-statistics.html

[2] [Online]. Available: http://www.seeingmachines.com

[3] [Online]. Available: http://www.smarteye.se

[4] [Online]. Available: http://www.smivision.com

[5] C. Ahlstrom, K. Kircher, and A. Kircher, "A gaze-based driver distraction warning system and its effect on visual behavior," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 2, pp. 965–973, Jun. 2013.

[6] A. Nabo, "Driver attention—Dealing with drowsiness and distraction," Smart Eye, Gothenburg, Sweden, Tech. Rep., 2009.

[7] J. P. Batista, "A real-time driver visual attention monitoring system," in *Pattern Recognition and Image Analysis*, vol. 3522, Berlin, Germany: Springer-Verlag, 2005, pp. 200–208.

[8] L. M. Bergasa, J. Nuevo, M. A. Sotelo, R. Barea, and M. E. Lopez, "Real-time system for monitoring driver vigilance," *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 1, pp. 63–77, Mar. 2006.

[9] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. 26th Annu. Conf. Comput. Graph. Interact. Tech.*, 1999, pp. 187–194.

[10] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: A 3D facial expression database for visual computing," *IEEE Trans. Vis. Comput. Graphics*, vol. 20, no. 3, pp. 413–425, Mar. 2014.

[11] M. L. Cascia, S. Sclaroff, and V. Athitsos, "Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3D models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 6, pp. 322–336, Apr. 2000.

[12] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[13] F. De la Torre and M. H. Nguyen, "Parameterized kernel principal component analysis: Theory and applications to supervised and unsupervised image alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2008, pp. 1–8.

[14] 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.

[15] N. Edenborough *et al.*, "Driver state monitor from DELPHI," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2005, pp. 1206–1207.

[16] G. M. Fitch *et al.*, "The impact of hand-held and hands-free cell phone use on driving performance and safety-critical event risk," Nat. Highway Traffic Safety Admin., Washington, DC, USA, Tech. Rep. DOT HS 811 757, 2013.

[17] L. Fletcher, N. Apostoloff, L. Petersson, and A. Zelinsky, "Vision in and out of vehicles," *IEEE Intell. Syst.*, vol. 18, no. 3, pp. 12–17, May/Jun. 2003.

[18] L. Fletcher, G. Loy, N. Barnes, and A. Zelinsky, "Correlating driver gaze with the road scene for driver assistance systems," *Robot. Auton. Syst.*, vol. 52, no. 1, pp. 71–84, Jul. 2005.

[19] L. Fletcher and A. Zelinsky, "Driver inattention detection based on eye gaze—Road event correlation," *Int. J. Robot. Res.*, vol. 28, no. 6, pp. 774–801, Jun. 2009.

[20] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010.

[21] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 478–500, Mar. 2010.

[22] J. L. Harbluk, Y. I. Noy, P. L. Trbovich, and M. Eizenman, "An on-road assessment of cognitive distraction: Impacts on drivers' visual behavior and braking performance," *Accid. Anal. Prev.*, vol. 39, no. 2, pp. 372–379, Mar. 2007.

[23] J. Heinzmann and A. Zelinsky, "3D facial pose and gaze point estimation using a robust real-time tracking paradigm," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recog.*, 1998, pp. 142–147.

[24] H. Ishiguro *et al.*, "Development of facial-direction detection sensor," in *Proc. 13th ITS World Congr.*, 2006, pp. 1–8.

[25] T. Ishikawa, S. Baker, I. Matthews, and T. Kanade, "Passive driver gaze tracking with active appearance models," in *Proc. 11th World Congr. Intell. Transp. Syst.*, 2004, pp. 1–12.

[26] Q. Ji and X. Yang, "Real time visual cues extraction for monitoring driver vigilance," in *Computer Vision Systems*, Berlin, Germany: Springer-Verlag, 2001, pp. 107–124.

[27] Q. Ji and X. Yang, "Real-time eye, gaze, and face pose tracking for monitoring driver vigilance," *Real-Time Imag.*, vol. 8, no. 5, pp. 357–377, Oct. 2002.

[28] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *Proc. IEEE ICCV*, Oct. 2009, pp. 365–372.

[29] J. Lee *et al.*, "Detection of driver distraction using vision-based algorithms," in *Proc. 23rd Enhanced Safety Veh. Conf.*, Seoul, Korea, 2013, 11-0322.

[30] S. J. Lee, J. Jo, H. G. Jung, K. R. Park, and J. Kim, "Real-time gaze estimator based on driver's head orientation for forward collision warning system," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, pp. 254–267, Mar. 2011.

[31] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[32] C. Morimoto, D. Koons, A. Amir, and M. Flickner, "Pupil detection and tracking using multiple light sources," *Image Vis. Comput.*, vol. 18, no. 4, pp. 331–335, Mar. 2000.

[33] E. Murphy-Chutorian, A. Doshi, and M. M. Trivedi, "Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation," in *Proc. IEEE Intell. Transp. Syst. Conf.*, 2007, pp. 709–714.

[34] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, Apr. 2009.

[35] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation and augmented reality tracking: An integrated system and evaluation for monitoring driver awareness," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 2, pp. 300–311, Jun. 2010.

[36] E. M. Rantanen and J. H. Goldberg, "The effect of mental workload on the visual field size and shape," *Ergonomics*, vol. 42, no. 6, pp. 816–834, Jun. 1999.

[37] M. Rezaei and R. Kletter, "Look at the driver, look at the road: No distraction! No accident!" in *Proc. IEEE CVPR*, 2014, pp. 129–136.

[38] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," *Int. J. Comput. Vis.*, vol. 91, no. 2, pp. 200–215, Jan. 2011.

[39] P. Smith, M. Shah, and N. da Vitoria Lobo, "Determining driver visual attention with one camera," *IEEE Trans. Intell. Transp. Syst.*, vol. 4, no. 4, pp. 205–218, Dec. 2003.

[40] J. Sung, T. Kanade, and D. Kim, "Pose robust face tracking by combining active appearance models and cylinder head models," *Int. J. Comput. Vis.*, vol. 80, no. 2, pp. 260–274, Nov. 2008.

[41] J. Tison, N. Chaudhary, and L. Cosgrove, "National phone survey on distracted driving attitudes and behaviors," Nat. Highway Traffic Safety Admin., Washington, DC, USA, Tech. Rep., 2011.

[42] R. Valenti, Z. Yucel, and T. Gevers, "Robustifying eye center localization by head pose cues," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 612–618.

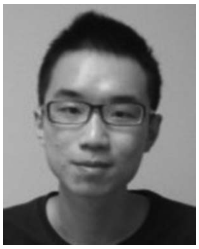
[43] D. Vlasic, M. Brand, H. Pfister, and J. Popović, "Face transfer with multilinear models," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 426–433, Jul. 2005.

[44] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 532–539.

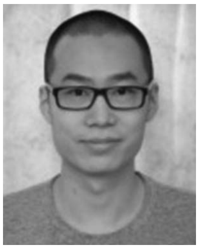
[45] F. Yang, E. Shechtman, J. Wang, L. Bourdev, and D. Metaxas, "Face morphing using 3D-aware appearance optimization," in *Proc. Graph. Interace Conf. Can. Inf. Process. Soc.*, 2012, pp. 93–99.

Francisco Vicente received a B.Sc. degree in telecommunications and a M.Sc. degree in telematics from Universidad Politecnica de Cartagena, Cartagena, Spain, in 2007 and 2008, respectively. He is currently working toward a M.S. degree in robotics at Carnegie Mellon University, Pittsburgh, PA, USA.

Since 2011, he has been a Research Associate with the Robotics Institute, Carnegie Mellon University. His research interests are in the fields of computer vision and machine learning.



Zehua Huang received a B.Sc. degree in computer science from Beihang University, Beijing, China, in 2012. He is currently working toward a M.S. degree at the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA. His research interests include 3-D face reconstruction and 3-D-based face analysis.



Xuehan Xiong received a B.Sc. degree in computer science from University of Arizona, Tucson, AZ, USA, in 2009 and a M.S. degree in robotics from Carnegie Mellon University, Pittsburgh, PA, USA, in 2011. He is currently working toward a Ph.D. degree in robotics at Carnegie Mellon University. His research interests include computer vision, machine learning, and optimization.

Fernando De la Torre received the B.Sc. degree in telecommunications and the M.Sc. and Ph.D. degrees in electronic engineering from Ramon Llull University, Barcelona, Spain, in 1994, 1996, and 2002, respectively.

He is an Associate Research Professor at the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA. His research interests are in the fields of computer vision and machine learning. He is currently directing the Human Sensing Laboratory (http://humansensing.cs.cmu.edu), Carnegie Mellon University. He has over 150 publications in referred journals and conferences.

Dr. De la Torre is an Associate Editor of IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. He has organized or co-organized several workshops and has given tutorials at international conferences on the use and extensions of component analysis methods.

Wende Zhang received the M.Sc. and Ph.D. degrees in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 2002 and 2006, respectively.

He is a Senior Researcher with the Electrical and Controls Systems Research Laboratory, Research and Development Center, General Motors, Detroit, MI, USA. He has been the Team Leader of the Next Generation Perception System (global GM R&D and Engineering Team) since 2010. He has published more than 30 conference and journal papers. His research is focused on computer vision and image processing for automotive applications.

Dan Levi received the B.Sc. degree (with honors) in mathematics and computer science from Tel Aviv University, Tel Aviv, Israel, in 2000 and the M.Sc. and Ph.D. degrees in applied mathematics and computer science from Weizmann Institute, Rehovot, Israel, in 2004 and 2009, respectively.

In the Weizmann Institute, he conducted research in human and computer vision under the instruction of Prof. Shimon Ullman. Since 2007 he has been conducting industrial computer vision research and development at several companies, including Elbit Systems, Israel. Since 2009 he has been a Senior Researcher at the Smart Sensing and Vision Group, Advanced Technical Center Israel, General Motors Research, Herzliya, Israel. His research interests are in computer vision, machine learning, and object recognition.