A Novel Approach for Driver Eye Gaze Estimation using Epipolar Geometry

Chen-Yu Lee, Ashish Tawari, and Mohan M. Trivedi

Abstract— **The objective of this research is to estimate the gaze point of a driver using epipolar geometry. The system requires two video inputs from internal and external views of a vehicle. We use SIFT image descriptor with L-2 norm distance to find putative correspondences, and then we use RANSAC algorithm to estimate robust fundamental matrix between two views. Once the underlaying scene geometry is estimated, we can further use affine transformation to model the distribution of the reliable correspondence points across two views. Finally, eye gaze point is estimated on the external reference image plane by mapping the center point in the internal view to the external view using the estimated affine transformation. Experiment results show that epipolar geometry with affine transformation can perform accurate driver's gaze point estimation for Computer Vision and Robotics Research Laboratory (CVRR) video dataset and produce highest accuracy than pure homography approach and epipolar line searching approach. In addition, our system can also handle the exceptional case as a driver looks at a mirror or a steering wheel which increases the robustness of the gaze estimation system.**

I. Introduction

A car with vision is an important step to achieve the concept of smart vehicle which takes aim at providing timely responses to drivers in order to avoid accidents in real world. Many car accidents happen due to distraction while driving [3], therefore a robust and reliable driver attention feedback is essential for an active safety system. [13], [7], [8], [6], [5] suggest that eye gaze is an important cue to detect driver distraction. We would like to use computer vision techniques to estimate eye gaze point and provide timely result of driver attention.

[12] describes a passive driver gaze tracking system which uses a global head model, specifically an Active Appearance Model (AAM), to track the whole head. From the AAM, the eye corners, eye region, and head pose are extracted and then used to estimate the gaze. [15] proposes a hybrid scheme to combine head pose and eye location information to obtain enhanced gaze estimation. [2] introduces a 3D eye tracking system where head motion is allowed without the need for markers or worn devices. [16], [1], [4] utilize an ellipse template matching scheme with sliding window based searching to find eye gaze point. However, all listed approaches require a camera faces at the driver. This camera setting involves driver's privacy and comfortableness. Also, these systems require extremely complicated facial feature extraction and facial model reconstruction, which makes it impossible for real time applications in active safety systems. Furthermore, the ellipse template might not be reliable in real cases considering different head poses and eye sizes.

In this paper, we introduce an eye gaze estimation system that only requires two video sequence inputs: the external view and the internal view of a vehicle as shown in Figure 1. Also, we use epipolar geometry to model the relationship of interest points between two views, and then apply affine transformation to estimate the final gaze point in the external view. The experiment results show that epipolar geometry with affine transformation outperforms homography and epipolar line search approaches. The proposed system is easy to implement without complicated facial model reconstruction, and it requires low computational time. We can achieve 19.91 pixels accuracy. This implementation shows the possibility to perform active safety system for vehicles in real world situation. This will be covered in section 2.

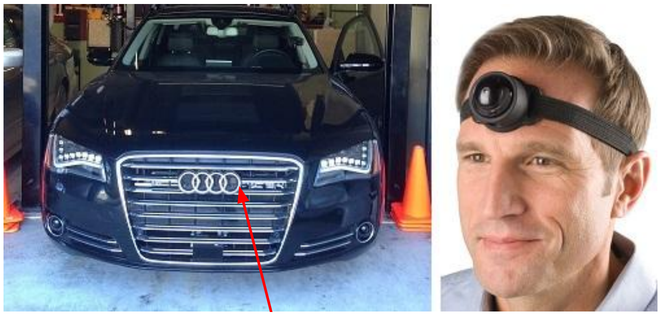

Fig. 1. Left and right images show the positions of two cameras. The external camera is located in front of a car and the internal camera is mounted on a driver's head.

In addition, we further investigate two components of a vehicle: a mirror and a steering wheel. These two components are useful to detect when a driver looks both sides or looks down when there is no strong matches for the frontal view case. We apply the same system to these parts and extract the distributions of orientations of SIFT matches for these two parts, and choose the part with maximum distribution concentration as the region where the driver looks at. The result shows that the system can detect where the driver looks at (front, sides, and down) and create the opportunity of exploiting interior parts of a vehicle to detect driver's attention. This will be covered in section 3.

II. Algorithm: the frontal view case

Figure 2 shows the pipeline of our system. In this section we will cover the three main steps in our gaze point estimation system for the frontal view case.

A. Feature Extraction

To find the reliable relationship of interest points between internal view and external view, we have to use image

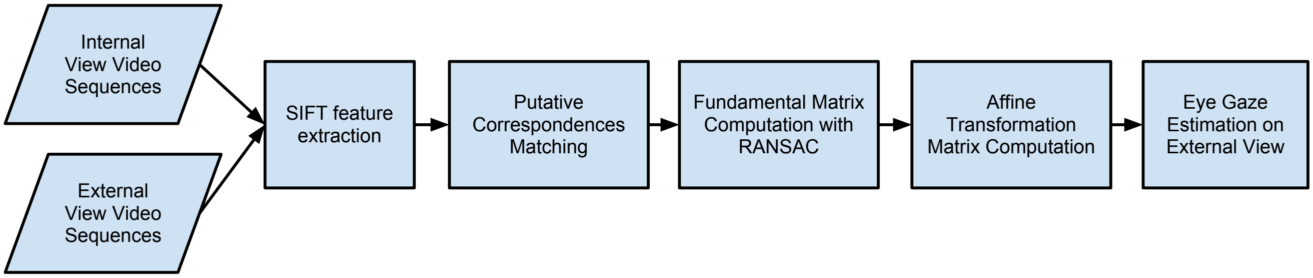
Internal View Video Sequences
External View Video Sequences
SIFT feature extraction
Putative Correspondences Matching
Fundamental Matrix Computation with RANSAC
Affine Transformation Matrix Computation
Eye Gaze Estimation on External View

Fig. 2. System pipeline of the eye gaze estimation algorithm.

descriptor to capture the texture information of images fast and correctly. David G Lowe propose local scale-invariant features transformation (SIFT) in [14], which is one of the most powerful and popular feature extraction technique in computer vision society. Figure 3 shows the result of interest points detection by Laplacian of Gaussian filter. SIFT feature extraction is then performed on these interest points.

B. Model of Stereo Images

We have to choose an appropriate method to model the relationship of interest points between two images of the same scene. Homography [11] is a powerful approach to map points and lines of a pure planar between two stereo images. However, we have a lot of objects without the planar constraint in our dataset such as cars, pedestrian, and buildings. Therefore, we have to be more carefully when choosing such model. Epipolar geometry [11] is the geometry of stereo vision without the pure planar constraint. It is an reliable mathematical model when two cameras view a 3D scene from two distinct positions based on the assumption that the cameras can be approximated by the pinhole camera model.

With an uncalibrated stereo rig, we can use the fundamental matrix [9] to model the relationship of points between two views. The fundamental matrix F is a 3×3 matrix which relates corresponding points in stereo images with homogeneous image coordinates. The basic epipolar constraint can be written as:

$$x'^T F x = 0 \tag{1}$$

where x and x' are homogeneous image coordinates of the same points in image 1 and image 2. To compute the matrix F correctly, we first use L-2 norm distance to produce putative correspondences as shown in Figure 4. We still can see a small amount of mismatch even though the majority of the matching pairs are correct. To further filter out the mismatching pairs, we use RANdom SAmple Consensus (RANSAC) [10] algorithm to iteratively compute the fundamental matrix until it converges. Figure 5 shows the correspondence matches after RANSAC algorithm. We can see that RANSAC with fundamental matrix can filter out most of the outliers and still preserve true matches.

C. Gaze Point Estimation

The final step of our system is to estimate the gaze point in the external view with the model we have. We assume the driver look at the center of the internal view, and our algorithm would find the correspondence point in the external view. With a point x in image 1, we can compute the epipolar line l in image 2 using

$$l = Fx \tag{2}$$

With the epipolar line l in image 2, we can use normalized cross correlation (NCC) to find the point with maximum response along the line l with respect to the center point in image 1. However, the center point in image 1 could have no strong texture information, thereby producing uniform responses along the epipolar line l in image 2. To overcome this problem, we model the point transform function with the affine model A based on the current correspondence matches:

$$x' = \begin{bmatrix} a & b & c \\ d & e & f \\ 0 & 0 & 1 \end{bmatrix} x = Ax \tag{3}$$

where A has 6 degree of freedom (DOF) from a to f. Now we can use affine transformation to map the center point in the internal view to the external view very accurately because the affine transformation is computed from robust correspondence matches using RANSAC and fundamental matrix instead of random points between two images. We will show the differences between homography, epipolar line search, and epipolar geometry with affine transformation in the experiment section.

III. Algorithm: a mirror and a steering wheel

In this section, we exploit two components (left mirror and steering wheel) of a vehicle to determine where the driver looks at.

A. Component Templates

It is hard to find a overlap region for internal and external views when a driver looks extremely left or down, but we still want to detect these cases for a driver distraction system. An intuitive way is to crop a left mirror region and a steering

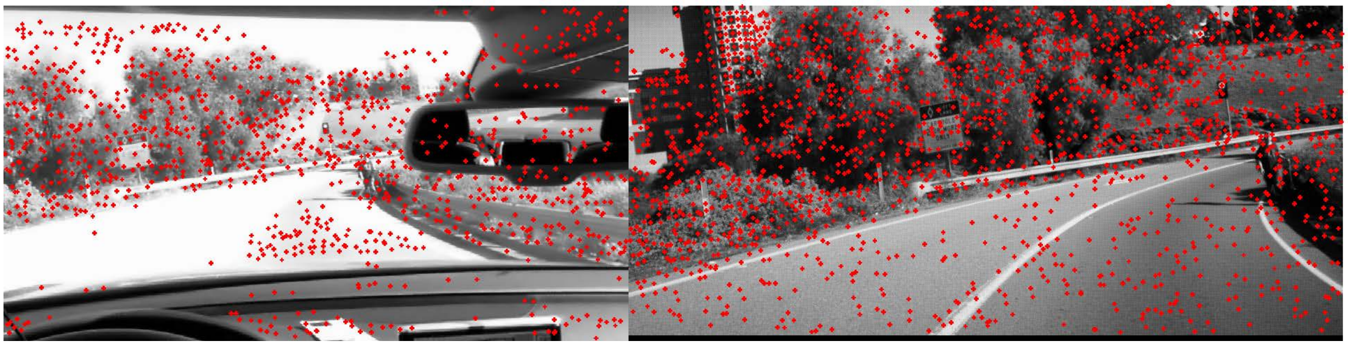

Fig. 3. SIFT feature extraction for internal view (left) and external view (right) of a vehicle.

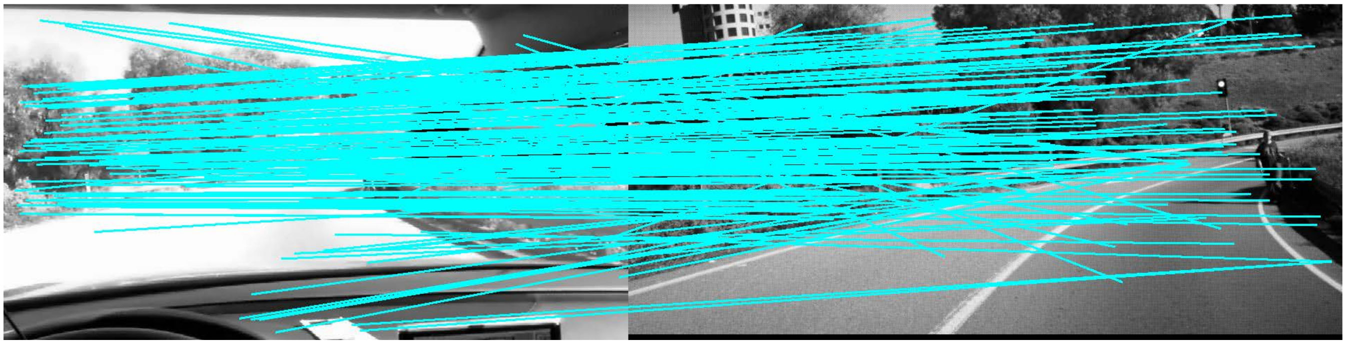

Fig. 4. Putative correspondence using L-2 norm distance.

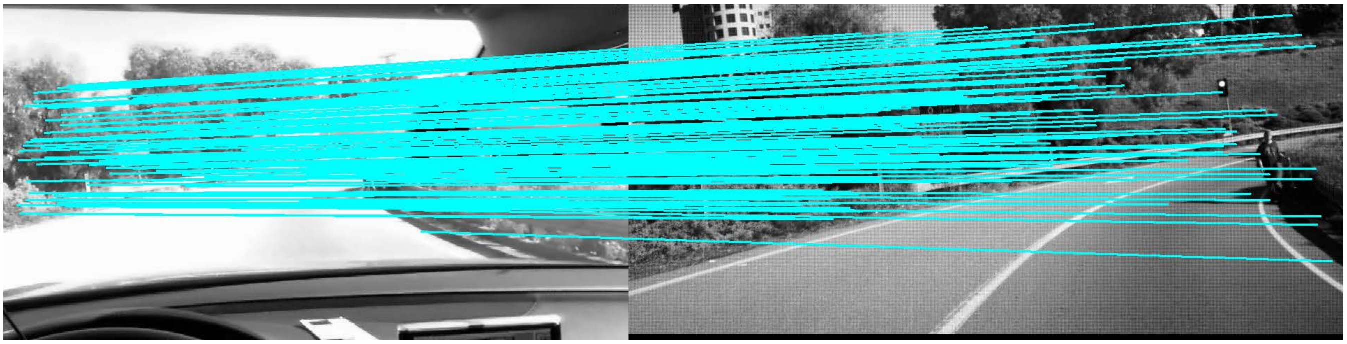

Fig. 5. Feature point correspondence after RANSAC with fundamental matrix.

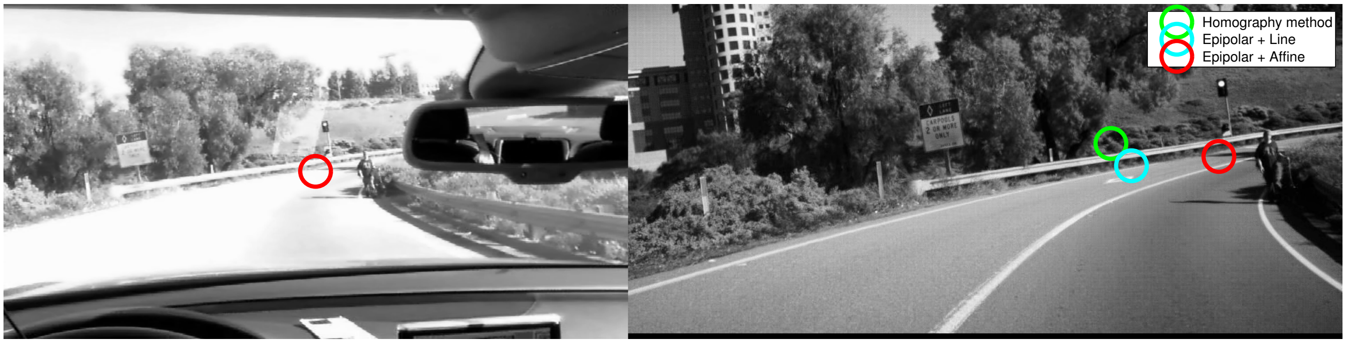

Fig. 6. System output for frontal view case. Epipolar geometry with affine transformation can achieve better eye gaze point estimation than other two methods.

wheel region at a certain time sequence, and run the same system for these two parts to get SIFT feature point matches.

B. *Distribution of Orientations of SIFT Matches*

To examinate SIFT feature point matches, we construct three histograms of orientations of these match lines as shown in Figure 7, 8, and 9. These histograms give statistical information of how well the SIFT points are matched. If a driver looks at front, then frontal view would give highly concentrate distribution with only few peaks. If a driver looks the left mirror, then frontal view would give several peaks with low peak values.

C. Looking Region Determination

We also generate three corresponding histograms to represent ideal cases for these three parts (frontal, wheel, and left mirror) as shown in bottom right of Figure 7, 8, and 9. If a driver looks at a centain part then it should generate similar histogram result as the ideal case. Goodness-of-fit allows us to determine whether the observed histogram corresponds to the ideal case. Chi-squared (χ^2) distance between two histograms h^i and h^j with dimension index $k \in [1,d]$

$$D(h^i,h^j) = \frac{1}{2}\sum_{k=1}^{d}\frac{(h^i_k - h^j_k)^2}{h^i_k + h^j_k} \tag{4}$$

is a good choice for comparing discrete probability distributions. It performs better than Euclidean distance because it gives higher weights for those bins with low bin values, so it can emphasis more on those bad SIFT matches. Here we use $d \in [-90,90]$ so that the histograms have 180 bins indivisually. We compute chi-squared distances for the three parts and determine where the driver looks at with the lowest chi-squared distance.

IV. Experimental Analysis

We conduct experiment on the dataset provided by Computer Vision and Robotics Research Laboratory (CVRR) at University of California, San Diego. In this experiment we use an Audi A8 sedan with Eight-speed Tiptronic transmission and quattro all-wheel drive. Also we use a video camera in front of the vehicle and a logitech webcam mounted on the driver's head. Here we record 263 frames in real road situation with a variety of noise such as cars, pedestrains, and buildings. Interval video is 960×1280 and external video is 472×1024.

A. Frontal Case

To evaluate the performance of the system, we implement three approaches and compare the eye gaze estimation accuracies and computation complexity in Table I. We also show the result of three different approaches in Figure 6. Epipolar geometry with affine transformation clearly outperforms pure homography and epipolar line search methods in terms of accuracies. The reason why homography model does not work well is because the object in real world is not pure planar. Also, epipolar line search is worse then epipolar + affine is because the image could have no strong texture to perform normalized cross correlation.

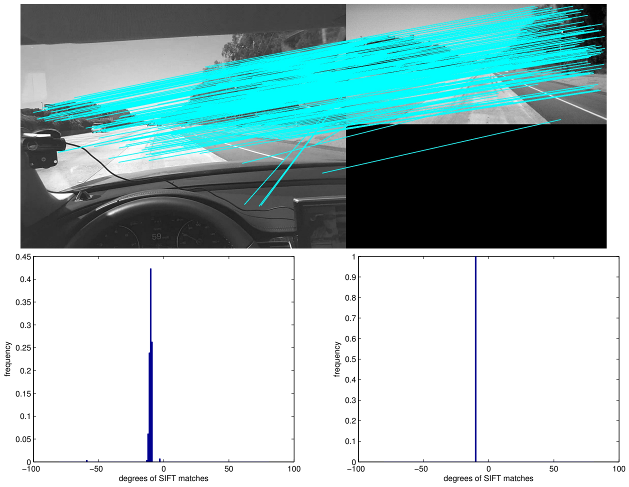

Fig. 7. Top image shows SIFT matches for frontal view. Bottom left image shows the distribution of orientations of SIFT matches. Bottom right image shows the ideal distribution if the driver looks at front.

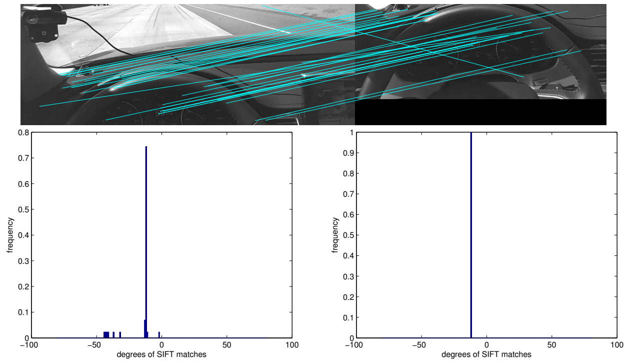

Fig. 8. Top image shows SIFT matches for wheel view. Bottom left image shows the distribution of orientations of SIFT matches. Bottom right image shows the ideal distribution if the driver looks at wheel.

B. Left mirror and Steering wheel Cases

Table II shows the presicion and recall rates of three vehicle parts detection. We also compute F-scores $= 2 * precision * recall/(precision + recall)$ as shown in Table III for three parts to determine which part has the best performance. Here frontal view has highest F-score because it usually has a large amount of overlap between external view and internal view. However, left mirror part has lowest F-score because when a driver looks at left, it usually has a lot of non-mirror region that will interfere the SIFT matching results. This result shows that we can detect not only the frontal case, but also the left mirror region and the steering wheel region if there is no strong matches for the frontal view.

V. Concluding Remarks

In this paper we demonstrate three different approaches to perform eye gaze estimation for drivers in real world situation. The result shows that epipolar geometry is a reliable method to model the relationship between the internal view and external view of a vehicle. With affine transformation computed using robust correspondence, we

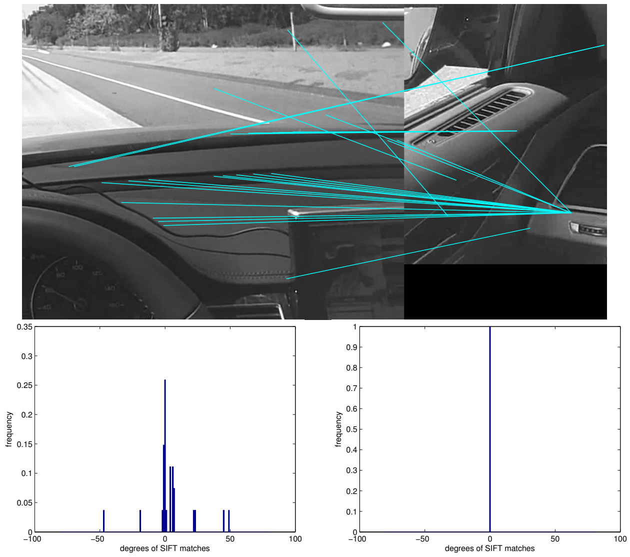

Fig. 9. Top image shows SIFT matches for left mirror view. Bottom left image shows the distribution of orientations of SIFT matches. Bottom right image shows the ideal distribution if the driver looks at left mirror.

TABLE I

EXPERIMENT RESULTS FOR GAZE ESTIMATION OF THREE DIFFERENT APPROACHES. SECOND COLUMN SHOWS THE AVERAGE DISTANCES BETWEEN ESTIMATED GAZE POINT AND GROUND TRUTH POINT.

Method	Distances (pixels)
Homography	295.62
Epipolar line search	60.29
Epipolar + Affine	**19.91**

can further improve the accuracy with mean position error to 19.91 pixel. Our system can also detect important parts of a vehicle and provide statistical information about where a driver looks at. We can further improve the templates of vehicle parts by using online learning for future work.

REFERENCES

[1] T. Bar, J. F. Reuter, and J. M. Zollner. Driver head pose and gaze estimation based on multi-template icp 3-d point cloud alignment. *IEEE Conference on Intelligent Transportation Systems*, 2012.
[2] D. Beymer and M. Flickner. Eye gaze tracking using an active stereo head. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2003.
[3] Y. Dong, Z. Hu, K. Uchimura, and N. Murayama. Driver inattention monitoring system for intelligent vehicles: A review. *IEEE Transaction on Intelligent Transportation Systems*, 2011.
[4] A. Doshi and M. M. Trivedi. Drivers view and vehicle surround estimation using omnidirectional video stream. *IEEE Intelligent Vehicles Symposium*, 2003.
[5] A. Doshi and M. M. Trivedi. Head and gaze dynamics in visual attention and context learning. *IEEE CVPR Joint Workshop for Visual and Contextual Learning and Visual Scene Understanding*, 2009.
[6] A. Doshi and M. M. Trivedi. Investigating the relationships between gaze patterns, dynamic vehicle. *IEEE Intelligent Vehicles Symposium*, 2009.
[7] A. Doshi and M. M. Trivedi. On the roles of eye gaze and head dynamics in predicting drivers intent to change lanes. *IEEE Transaction on Intelligent Transportation Systems*, 2009.
[8] A. Doshi and M. M. Trivedi. Tactical driver behavior prediction and intent inference: A review. *IEEE Conference on Intelligent Transportation Systems*, 2011.

TABLE II

CONFUSION MATRIX OF LOOKING REGION DETECTION.

	Front	Steering wheel	Left mirror	recall
Front	164/200	13/200	23/200	0.82
Steering wheel	16/35	19/35	0	0.54
Left mirror	20/28	0	8/28	0.29
precision	0.82	0.41	0.74	

TABLE III

F-SCORES FOR THREE PARTS OF A VEHICLE.

Parts	F-score
Front	0.82
Steering wheel	0.47
Left mirror	0.42

[9] O. D. Faugeras. What can be seen in three dimensions with an uncalibrated stereo rig. *Proceedings of European Conference on Computer Vision*, 1992.
[10] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. of the ACM*, 1981.
[11] R. Hartley and A. Zisserman. Multiple view geometry in computer vision. *Cambridge University Press*, 2003.
[12] T. Ishikawa, S. Baker, I. Matthews, and T. Kanade. Passive driver gaze tracking with active appearance models. *CMU-RI-TR-04-08*, 2004.
[13] Y. Liang, M. L. Reyes, and J. D. Lee. Real-time detection of driver cognitive distraction using support vector machines. *IEEE Transaction on Intelligent Transportation Systems*, 2007.
[14] D. G. Lowe. Object recognition from local scale-invariant features. *Proceedings of the International Conference on Computer Vision*, 1999.
[15] R. Valenti, N. Sebe, , and T. Gevers. Combining head pose and eye location information for gaze estimation. *IEEE Transaction on Image Processing*, 2012.
[16] D. H. Yoo and M. J. Chung. A novel non-intrusive eye gaze estimation using cross-ratio under large head motion. *Computer Vision and Image Understanding*, 2005.