Aalborg Universitet

AALBORG UNIVERSITY
DENMARK

Attention estimation by simultaneous analysis of viewer and view

Tawari, Ashish; Møgelmose, Andreas; Martin, Sujitha; Moeslund, Thomas B.; Trivedi, Mohan M.

Published in:
IEEE 17th International Conference on Intelligent Transportation Systems (ITSC), 2014

DOI (link to publication from Publisher):
10.1109/ITSC.2014.6957880

Publication date:
2014

Document Version
Accepted author manuscript, peer reviewed version

Link to publication from Aalborg University

Citation for published version (APA):
Tawari, A., Møgelmose, A., Martin, S., Moeslund, T. B., & Trivedi, M. M. (2014). Attention estimation by simultaneous analysis of viewer and view. In *IEEE 17th International Conference on Intelligent Transportation Systems (ITSC), 2014* (pp. 1381-1387). IEEE Press. https://doi.org/10.1109/ITSC.2014.6957880

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from vbn.aau.dk on: March 28, 2024

Attention Estimation By Simultaneous Analysis of Viewer and View

Ashish Tawari[1], Andreas Møgelmose[1,2], Sujitha Martin[1], Thomas B. Moeslund[2] and Mohan M. Trivedi[1]

***Abstract*— This paper introduces a system for estimating the attention of a driver wearing a first person view camera using salient objects to improve gaze estimation. A challenging data set of pedestrians crossing intersections has been captured using Google Glass worn by a driver. A challenge unique to first person view from cars is that the interior of the car can take up a large part of the image. The proposed system automatically filters out the dashboard of the car, along with other parts of the instrumentation. The remaining area is used as a region of interest for a pedestrian detector. Two cameras looking at the driver are used to determine the direction of the driver's gaze, by examining the eye corners and the center of the iris. This coarse gaze estimation is then linked to the detected pedestrians to determine which pedestrian the driver is focused on at any given time.**

I. INTRODUCTION

First person or ego-centric vision attempts to understand human behavior by acquiring information on what the person is looking at [1]. It employs videos/images from head mounted cameras. Recently, technological advances have made lightweight, wearable, egocentric cameras both practical and popular in various fields. The GoPro camera for instance can be mounted on helmets and is popular in a lot of sports such as biking, surfing, and skiing. The Microsoft SenseCam can be worn around the neck and has enough video storage to capture an entire day for the idea of "life logging". Cognitive scientists like to use first-person cameras attached to glasses (often in combination with eye trackers such as Tobii or SMI) to study visual attention in naturalistic environments. Most recently, emerging products like Google Glass have begun to make the first attempts to bring the idea of wearable, egocentric cameras into the mainstream.

Advances in the wearable-devices have enabled novel data acquisition in real-world scenarios. In the field of egocentric video, much of the recent work has focused on object detection, first-person action and activity detection, and data summary in context of "life-logging" video data. In this work, we present a unique data set collected during complex driving tasks with the aim of understanding driver-state and driver-'attention'. We use Google Glass to capture the driver's field of view, and a distributed camera setup instrumented in the vehicle to observe the driver's head and eye movements. Wearable and uncluttered cameras provide a practical advantage of ease of capture. The challenge, however, in our distributed camera setup, is to acquire data in sync to understand the driver-state, the environment and the vehicle-state simultaneously.

[1]Laboratory for Intelligent and Safe Automobiles, UC San Diego, United States [atawari, scmartin, mtrivedi]@ucsd.edu

[2]Visual Analysis of People Lab, Aalborg University, Denmark [am, tbm]@create.aau.dk

Driver gaze and head-pose are linked to the driver's current focus of attention [2], [3]. Therefore, eye and/or head tracking technology has been used extensively for visual distraction detection. The driving environment presents challenging conditions for a remote eye tracking technology to robustly and accurately estimate eye gaze. Even though precise eye gaze is desirable, coarse gaze direction is often sufficient in many applications [4]. By having a head mounted camera, we have direct access to the field of view of the driver. By analyzing salient regions in the field of view (bottom-up attention model), one can estimate the focus of attention of the driver [5]. However, in a complex task such as driving, it is hard to say precisely where or at what we are looking, since eye fixations are often governed by goal-driven mechanisms (top-down attention model).

Towards this end, we propose a Looking-In-Looking-Out framework to estimate the driver's focus of attention by simultaneously observing the driver and the driver's field of view. We propose to measure coarse eye position and combine the salience of the scene to understand what object the driver is focused on at any given moment. We are not proposing a precise gaze tracking approach, but rather to determine the driver's attention by understanding coarse gaze direction and combining it with analysis of scene salience to determine important areas of interest - in our case pedestrians.

Our interest in pedestrians comes from the fact that in 2011, pedestrian deaths accounted for 14 percent of all traffic fatalities in motor vehicle traffic crashes in the United States. Almost three-fourths (73%) of pedestrian fatalities occurred in an urban setting versus a rural setting. 88% of pedestrian fatalities occurred during normal weather conditions (clear/cloudy), compared to rain, snow and foggy conditions. By knowing which pedestrians the driver has and has not seen, measures against collisions can be taken more accurately. While our main focus is on pedestrians, the framework can easily accommodate any object of interest or even a low-level saliency model to estimate the focus of attention.

The remainder of the paper is organized as follows. We give an overview of relevant related work in section II and explain the methods for determining gaze and detecting pedestrians in section III. In section IV we briefly review our captured data, and section V shows our results, before wrapping up with some concluding remarks and future work in section VI.

II. Related Work

Use of wearable cameras is not new [6]. In the last decade, gaze tracking systems such as [7], Tobii and SMI have made mobile gaze tracking in real life settings possible. More recently, the advances in hardware technology have made their usage more common in the computer vision community [8]–[12]. These systems are often used successfully in laboratory or controlled lighting conditions. Their use in complex environments is limited due to lengthy calibration, motion, illumination changes and, in case of driving, possible hindrance to the driver's front- or side-view. We discuss select work in activity recognition and gaze-behavior related research areas which are relevant in our current and larger interest in studying driver intent and behavior in real-world driving.

Ogaki et al. [13], using an inside-out camera system, combined eye-motion from inside looking camera and global motion from outside one to recognize indoor office activities. The authors suggest that joint cues from inside looking and outside looking cameras perform the best across different users. Doshi and Trivedi [3] introduced a similar system, but primarily for vehicular use. Pirsiavash and Ramanan [14] detected indoor apartment activities of daily living in first person camera view. They used object-centric action models which perform much better than low-level interest points based one to recognize activities. They show that using ground-truth object labels in the action models significantly improves recognition performance. This suggests that recognizing objects of interest is key to recognizing tasks/activities in naturalistic settings.

Gaze allocation models are usually derived from static picture viewing studies. Many of the existing works are based on the computation of image salience [15] using low-level image features such as color contrast or motion to provide a good explanation of how humans orient their attention. However, these models fail for many aspects of picture viewing and natural task performance. Borji et al. [16] observe that object-level information can better predict fixation locations than low-level saliency models. Judd et al. [17] show that incorporating top-down image semantics such as faces and cars improves saliency estimation in images.

Inspired by the above findings, we present a driver's visual attention model using inside and outside looking camera views. In particular, we propose a model to determine coarse gaze direction and combine it with an object based saliency map to determine the allocated attention of the driver. Note that our interest lies in 'higher-level' semantic information about the driver attention and not 'low-level' precise gaze measurement. Our proposed framework circumvents the precise gaze estimation problem by utilizing a saliency map to achieve robust performance. Precise eye gaze from remote cameras is difficult not only due to low resolution of the eye region, but also due large head turns, self occlusion, illumination changes and hard shadows existing in an ever changing dynamic driving environment. To deal with large head turns and self occlusion, we propose to use a distributed camera system to monitor the driver.

We evaluate the proposed framework using a novel naturalistic driving data set using multiple cameras monitoring the driver and the outside environment. We use the head mounted camera from Google Glass to capture the driver's field of view. This particular device did not provide the ability to automatically synchronize footage with other cameras at per frame level. However, the ease, quick setup time (wearing and pressing capture button) as well as clean and uncluttered face view still makes the device a good choice. To obtain frame level synchronization, we mount an outside looking camera on the ego-vehicle which in turn is synchronized to the rest of the systems. Details on our synchronization strategy is provided in section IV.A head mounted camera provides the ability to capture not only the driver's outside field of view but also inside cockpit-view. In this work, we focus on the analysis of the outside view using the head mounted camera. This view poses unique challenges as discussed later.

III. Attention Estimation: LILO Framework

To infer the driver's attention, we are interested in knowing what object, in our case which pedestrian, the driver is looking at. There are two steps involved: first, estimating where driver is looking and second, detecting objects of interest in his/her field of view.In our current analysis, we focus on horizontal gaze variation since that is the most volatile and exercised direction by the driver to gain the knowledge of the environment. As we motivated earlier, we only require coarse gaze-direction and to distinguish it from precise gaze-value, we call it gaze-surrogate.

A. Gaze-Surrogate Estimation

We automatically detect facial features - eye corners and iris center, and use cylindrical eye-model 2 to estimate coarse gaze-direction. We use a facial feature tracking approach similar to [18] for detection of eye corners. During driving, however, large out-of-plane rotation of the head severely degrades the tracking performance. Hence, we use a two camera-system as proposed by Tawari et al. [19] to continuously track the facial features. From the facial features, we also calculate head pose, to be used in the gaze-direction calculation as explained below. We encourage the reader to refer to [18] and [19] for the details about eye-corner tracking, and head pose estimation and camera hand-off procedures. Here, we detail the iris detection and gaze-direction estimation algorithms.

Iris detection: The most prominent and reliable features within the eye region are the edges of the iris. The upper and lower eyelids in real face images occlude parts of the iris contours. Only the unoccluded iris edges can be used to fit the iris contour in the image plane. We detect the iris edge using a vertical edge operator in between upper and lower eyelids. The iris contours on the image plane are simplified as circles and center of the iris is detected using the circular Hough transform. Figure 1 shows the block diagram of the iris detection algorithm.

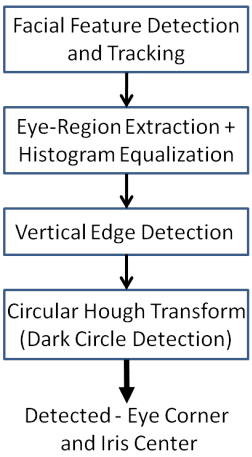


Fig. 1: Block diagram for extracting iris center

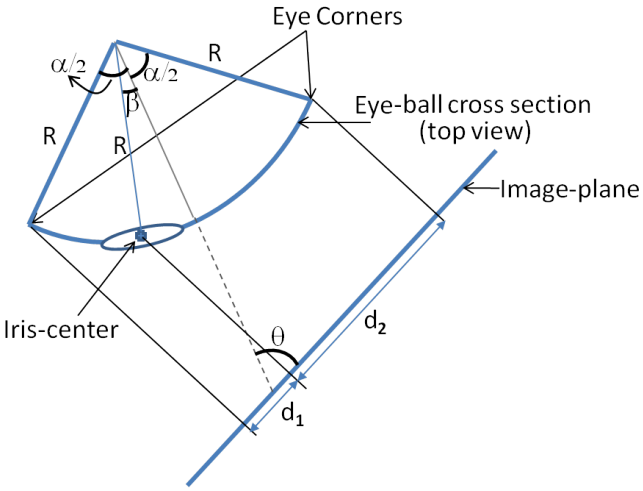


Fig. 2: Eye ball image formulation: estimating β, gaze-angle with respect to head, from $alpha$, θ, d_1 and d_2

Gaze-direction: Once the iris center is detected in the image plane, the gaze-direction β with respect to the head, see figure 2, is estimated as a function of α, the angle subtended by an eye in the horizontal direction, head-pose (yaw) angle θ, and the ratio of the distances of iris center from the detected corner of the eyes in the image plane. Equation 1-2 shows the calculation steps.

$$\frac{d_1}{d_2} = \frac{\cos(\theta - \alpha/2) - \cos(\theta - \beta)}{\cos(\theta - \beta) + \cos(180 - \theta - \alpha/2)} \quad (1)$$

$$\beta = \theta - \arccos\left(\frac{2}{d_1/d_2 + 1}\sin(\theta)\sin(\alpha/2) + \cos(\alpha/2 + \theta)\right) \quad (2)$$

Since the raw eye-tracking data is noisy (due to blinking and tracking errors), we smooth angle β with a median filter.



Fig. 3: Three dashboard images showing examples of the very unconstrained position and orientation the dashboard can have in the field of view.

B. Salient Object Detection

focus of attention detection in this paper is on pedestrians, thus requiring a pedestrian detector. Using first person view presents a number of interesting challenges compared to a stationary car-mounted camera. The major challenge is to determine the region of interest in which to look for pedestrians. With a stationary camera, it is either mounted so there are no obstructions in its view of the road, or it is mounted so any obstructions can easily be masked out manually.

This is not the case for first person view, where the perspective constantly changes and there is no way of setting up a constant mask. This section introduces an algorithm to automatically mask out the dashboard and other unwanted areas.

The pedestrian detection module in this system is based on the classic HOG-SVM detection presented by Dalal and Triggs in [20]. It is trained on the Inria person dataset from the same paper. The pedestrian detection itself is simply a module in the full system, and it could be swapped with other approaches without issues.

The most important part of the interior mask is the dashboard mask. The dashboard can take up just the bottom of the image, the majority of the image, or not be present at all (fig. 3) and the algorithm must handle all of those situations. We detect the distinct line between the windshield and dashboard and build from that:

1) Smooth out the input image with a Gaussian blur to even out noise.
2) Detect edges using the Canny edge detector [21].
3) Determine the major lines in the image using the generalized Hough transform [22].
4) Filter the lines by angle to include only near-horizontal lines.
5) Build a confidence map of the dashboard.

Fig. 4 shows sample output of step 4. Green lines are those that are horizontal enough to be considered in the dashboard map, red lines are ignored due to their extreme angles.

For each detected line, a polygon is drawn, which masks out all of the image below the line. These masks are

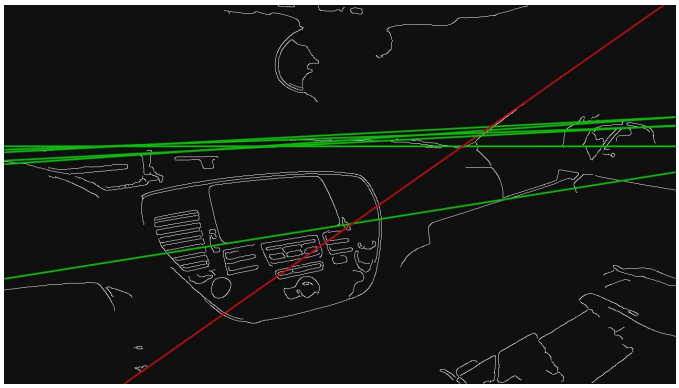


Fig. 4: Detected lines in the image. Red lines are discarded due to too much of a skew to constiture the dashboard edge.

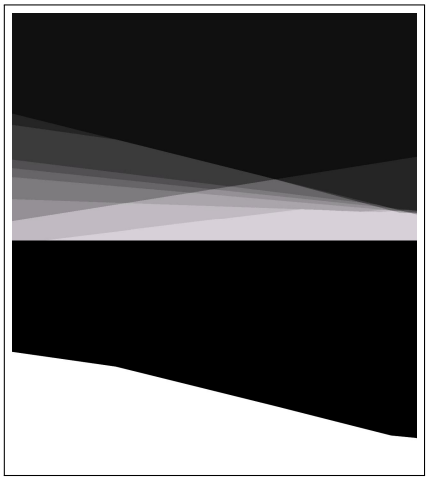


Fig. 5: Confidence map (top) and its resulting mask (bottom).

combined and result in a single-frame dashboard map. To counter noisy line detections, a cumulative confidence map is introduced.

The cumulative confidence map is created by adding 1 to all pixels in the map covered by the current single-frame map and subtracting 1 from all pixels not covered by the current single-frame map. Areas that are detected in several subsequent frames will grow to a high confidence, but after a while of no detections, the confidence will fall and eventually the mask disappears. Examples of confidence maps and masks are in fig. 5.

The use of the cumulative map is governed by two parameters, κ and λ. κ is the mask threshold. Any pixel in the confidence map with a value higher than κ is considered part of the dashboard map. In this implementation $\kappa = 2$. This parameter controls how confident the system must be in a given pixel to include it in the mask. λ is the upper limit of confidence values.For a very high λ value, the confidence can grow very high, thus resulting in a long delay before the pixel goes below κ.λ defines how long the memory of the system is. In this implementation $\lambda = 10$.

Apart from filtering out the dashboard, we detect and filter

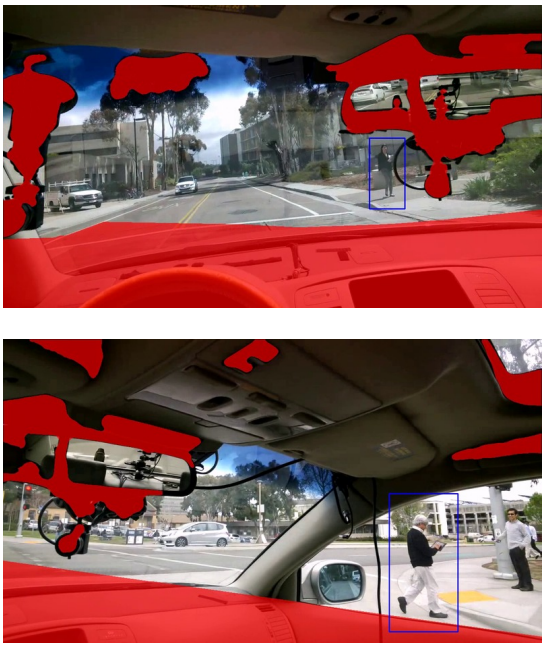


Fig. 6: Examples of pedestrian detection scenarios where the exclusion mask has been overlaid in red.

out black blobs large enough that they can only be part of the interior. We also discard pedestrian bounding boxes larger than 40% of the frame height.

C. Attended Object Determination

This step combines gaze-direction (β) and the salient object detected to determine which object the driver is attended to. This requires mapping from gaze-direction with respect to the head, β to pixel position in the external looking camera image. Equation 3 shows the mapping function to determine the x-position (i.e. the yaw-direction) in the image plane.

$$P_x(\beta; C_x, M_x, \phi) = C_x - M_x * \frac{\sin(\beta)}{\cos(\phi + \beta)} \qquad (3)$$

where ϕ is the angle between the external camera image-plane and eye-image plane, C_x is the pixel position when looking straight ($\beta = 0$), and M_x is a multiplication-factor, determining the change in pixel position with change in gaze direction. A calibration step with the user's cooperation (by asking them took in particular directions) can be performed to determine the parameters. Since the device is not firmly fixed to the head and can move during usage, we ideally need to perform calibration again. However, for our purposeswe found that as long as the camera is not rotated (along the vertical-axis allowed by the device for adjusting the display), it did not degrade the performance during normal usage. A Gaussian kernel around this location is combined with the detected object based image saliency to infer the allocated attention location. This leads to the attended object as the closest object detected around the gaze-location.



Fig. 7: An example of annotated sequence from the time synchronized video.

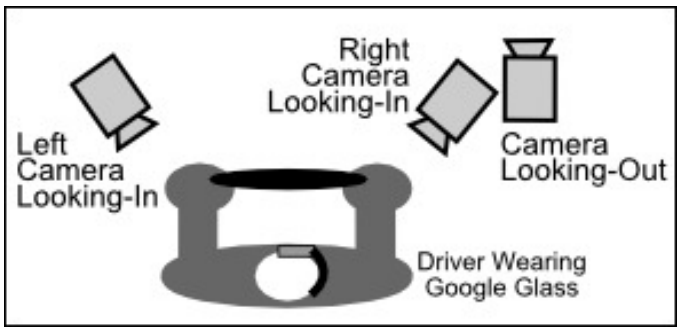


Fig. 8: Top view diagram of the test bed setup.

IV. Data set

Data is collected from naturalistic on-road driving using a vehicular test bed, which is equipped with one Google Glass and three GigE cameras as shown in fig. 8. The Google Glass is worn by the driver to give a first-person perspective. Data is captured from Google Glass at 50 frames per second at a resolution of 1280 x 720 pixels, and stored internally on the device. Of the three GigE cameras, one is mounted to the left of the driver near the A-pillar on the windshield looking at the driver, and two are mounted to the right of driver near the rear-view mirror on the windshield - one looking at the driver and one looking outside. A multi-perspective two camera approach is adopted to look at the driver because it increases the operational range when the driver makes spatially large head movements [19]. Data from the GigE cameras is captured at a resolution of 960 x 1280 pixels and is stored on a laptop with time stamps of millisecond precision. This allows for time synchronized videos.

In order to synchronize the first-person video with the videos looking at the driver, synchronization points are annotated using the first-person view and the outside front view. The criteria used in choosing these synchronization points include naturally occurring changes in traffic lights and artificially introduced momentary but periodic bursts of light (e.g. LED lights mounted to be visible in both first person view and outside front view). Then, assuming constant frame rate in the first-person video, linear interpolation is used to synchronize the first-person video with videos looking at the driver.

Using this test bed, multiple drivers were asked to drive on local streets. Approximately 40 minutes of data was collected in total, where the drivers passed through many stop signs, traffic signals and pedestrian crossings. In this paper, we are interested in events where the vehicular test bed is near or at these intersections, because these times are especially rich with visual interaction between driver and pedestrians. To evaluate our proposed attention system, two sets of ground truth labels are created via manual annotation on interesting event segments in the driving sequences. First, we manually annotated 410 frames and 1413 pedestrians, as seen in the first person perspective camera, with bounding boxes when either their face is visible or a significant portion of their body is visible. Second, we manually annotated 300 frames of where the driver is looking in the first person view - in particular, we annotated possible pedestrian candidate(s) as shown in fig. 7. This is accomplished by carefully looking at the driver's head and eye movements in the time synchronized videos with significant utilization of temporal and spatial context. For example, by looking at the driver's gaze over a time period, we are able to zero-in on particular pedestrians within a larger group. Annotating what the driver is looking at is especially challenging, and we have attempted to address this by obtaining consensus from multiple experts.

V. Experimental Evaluation

In this section we discuss the results of an experimental evaluation over several hundred frames of manually labeled data. There are two main contributions to evaluate: the impact of the dashboard masking and the attention estimation performance. In this section, both will be tested separately and then in combination.

Dashboard masking cuts the number of false positives in half with a low impact on the detection rate, as shown in table I. This paper is not about pedestrian detection as such, but the detection rates have been included to demonstrate that the masking does not impact them negatively in a significant way. The test set (1413 annotated pedestrians over 410 frames) is very challenging with articulated pedestrians and heavy occlusions, and while the detection numbers are low from an absolute point of view, the attention estimation still works well, as we shall see below.

The attention estimation has been tested on the same sequence with manually annotated pedestrians. Ground truth

TABLE I: Dashboard masking cuts the false positive rate in half, without impacting the detection performance too much.

	False positives per frame (FPPF)	Detection rate
Non-filtered (baseline)	**2.94**	0.27
With dashboard filter	**1.45**	0.21

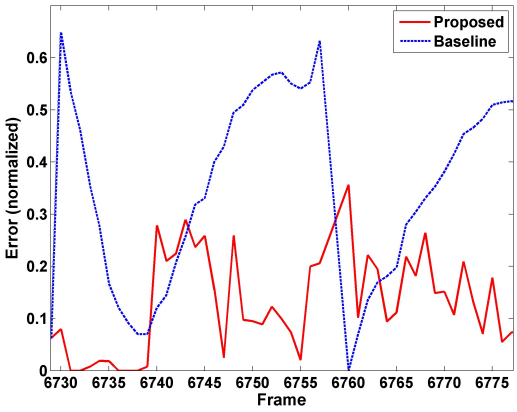


Fig. 9: Normalized error of surrogate gaze estimate on a continuous segment.

for the attentional location also was determined manually. Table II shows the accuracy of the proposed system given a perfect pedestrian detector, as well as the combined system. As points of comparison, we also include results of a simple attention estimator using only head pose - the center-bias based solution. This places the focus of attention on the central field of view of the driver's head.

The gaze-surrogate estimation significantly outperforms the baseline. The extra information gained by monitoring the eye gaze on top of the head pose gives rise to much better accuracy. The full system gives the correct subject of attention in nearly half the cases and with a relatively low median error. The attention estimation works better with a perfect pedestrian detector, enhancing the accuracy by 172% from 46.0% to 79.4%. Since it relies on detected pedestrian bounding boxes, it will inevitably give the wrong output if the correct pedestrian is not detected - in that case the attention will simply be associated with the nearest detected bounding box. Implementing a perfect pedestrian detector is outside the scope of this paper, but the entire system would work with a different and better detector. It is very likely that tracking of pedestrians could improve the detection system by compensating for missed detections, but it is also worth to note that due to the, at times, rather extreme ego-motion of the driver's head, this is not a trivial task.

Fig. 10 shows the pixel error of the gaze-surrogate detection over a full test sequence and the system is almost universally better than the baseline, except in the few situations where the subject of attention is right in the middle of the field-of-view, where the baseline system is better by sheer coincidence.

VI. Concluding Remarks

We have introduced a new approach to analyzing the attention state of a human subject, given cameras focused on the subject and their environment. In particular, we are motivated by and focus on the task of analyzing the focus of attention of a human driver. We presented a Looking-In and Looking-out framework combining gaze surrogate and object based saliency to determine the focus of attention. We evaluated our system in a naturalistic real-world driving data set with no scripted experiments. This made the data set very challenging, but realistic. We showed that by combining driver state (using face analysis), we significantly improve the performance over a baseline system based on image saliency with center bias alone. The proposed framework circumvents the precise gaze estimation problem (a very challenging task in real-world environment like driving) and hence, provide a robust approach for driver focus of attention estimation.

The challenges associated with ego-centric vision are unique (with large ego-motion) and compounded by the driving environment. It presents a difficult 'in-the-wild' scenario for object detection such as pedestrians, cars etc. We propose methods to prune false detection by incorporating a region-of-interest. There is still room for improvement. In the future, we will work to provide a comprehensive and rich data set from driver's field of view camera. The novel and unique vehicle test bed will also be very useful in other areas of interest e.g. driver's activity recognition.

References

[1] T. Kanade, "First-person, inside-out vision," in *IEEE Workshop on Egocentric Vision, CVPR*, 2009.

[2] E. Murphy-Chutorian and M. Trivedi, "Head Pose Estimation in Computer Vision: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 607–626, April 2009.

[3] A. Doshi and M. M. Trivedi, "Attention estimation by simultaneous observation of viewer and view," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 21–27.

[4] A. Tawari and M. M. Trivedi, "Dynamic Analysis of Multiple Face Videos for Robust and Continuous Estimation of Driver Gaze Zone," *IEEE Intelligent Vehicle Symposium*, 2014.

[5] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 1, pp. 185–207, 2013.

[6] B. Schiele, N. Olivier, T. Jebara, and A. Pentland, "An Interactive Computer Vision System DyPERS: Dynamic Personal Enhanced Reality System," in *International Conference on Vision Systems*, 1999.

[7] M. S. Devyver, A. Tsukada, and T. Kanade, "A Wearable Device for First Person Vision(FICCDAT workshop)," in *3rd International Symposium on Quality of Life Technology*, July 2011.

[8] E. H. Spriggs, F. De la Torre Frade, and M. Hebert, "Temporal Segmentation and Activity Classification from First-person Sensing," in *IEEE Workshop on Egocentric Vision, CVPR 2009*, June 2009.

[9] K. Kitani, T. Okabe, Y. Sato, and A. Sugimoto, "Fast unsupervised ego-action learning for first-person sports videos," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2011, pp. 3241–3248.

[10] X. Ren and C. Gu, "Figure-ground segmentation improves handled object recognition in egocentric video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2010, pp. 3137–3144.

TABLE II: Performance of the attention estimator. Pedestrian accuracy shows how many pedestrian bounding boxes are correctly determined to be the attention point for the driver. Mean and median error are measures of how far the gaze-surrogate point is from the correct pedestrian bounding box.

Estimator	Mean gaze error (in pixels)	Median gaze error (in pixels)	Attended pedestrian accuracy (%)	
			Manually annotated pedestrians	Full system
Center-bias based (baseline)	148.3	127.0	55.9	37.0
Proposed	54.1	32.2	79.4	46.0



Fig. 10: Visualization of the LILO attention result in a sequence where gaze switches from one salient location to other. The red box around the pedestrian illustrates the salient region, the Gaussian kernel with yellow center shows the gaze location in the driver's field of view. The solid box is the pedestrian that is the subject of the driver's attention as detected by the full system.

[11] A. Fathi, Y. Li, and J. Rehg, "Learning to Recognize Daily Actions Using Gaze," in *Computer Vision ECCV*, ser. Lecture Notes in Computer Science, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Springer Berlin Heidelberg, 2012, vol. 7572, pp. 314–327.

[12] Z. Lu and K. Grauman, "Story-Driven Summarization for Egocentric Video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),*, June 2013, pp. 2714–2721.

[13] K. Ogaki, K. Kitani, Y. Sugano, and Y. Sato, "Coupling eye-motion and ego-motion features for first-person activity recognition," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, June 2012, pp. 1–7.

[14] H. Pirsiavash and D. Ramanan, "Detecting Activities of Daily Living in First-person Camera Views," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012.

[15] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 11, pp. 1254–1259, Nov 1998.

[16] A. Borji, D. Sihite, and L. Itti, "Probabilistic learning of task-specific visual attention," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012, pp. 470–477.

[17] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to Predict Where Humans Look," in *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[18] X. Xiong and F. De la Torre Frade, "Supervised Descent Method and its Applications to Face Alignment," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, May 2013.

[19] A. Tawari, S. Martin, and M. M. Trivedi, "Continuous Head Movement Estimator (CoHMET) for Driver Assistance: Issues, Algorithms and On-Road Evaluations," *IEEE Trans. Intelligent Transportation Systems*, 2014.

[20] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *CVPR*, 2005.

[21] J. Canny, "A Computational Approach to Edge Detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PAMI-8, no. 6, pp. 679–698, Nov 1986.

[22] R. O. Duda and P. E. Hart, "Use of the Hough Transformation to Detect Lines and Curves in Pictures," *Commun. ACM*, vol. 15, no. 1, pp. 11–15, Jan. 1972. [Online]. Available: http://doi.acm.org/10.1145/361237.361242