2018 IEEE Intelligent Vehicles Symposium (IV)
Changshu, Suzhou, China, June 26-30, 2018

Object of Fixation Estimation by Joint Analysis of Gaze and Object Dynamics

Sujitha Martin and Ashish Tawari

Abstract— **Determining object of fixation is an important factor in many application of intelligent vehicles including driver's situational awareness estimation. The objective of this work is to infer object of fixation given fixation is occurring. We propose a system architecture that identifies object tracks in the scene, derives object characteristics independent of and jointly with gaze behavior, and utilizes a spatio-temporal sensitive machine learning framework to estimate the likelihood of an object being the object of fixation. Performance evaluation is conducted on a dataset of on-road driving, centered around urban intersections, with manual annotations of object of fixation. Our proposed system can achieve up to 83% average precision accuracy when compared to baseline of 78%. Furthermore, comparing the effects of different combinations of object characteristics on precision and recall accuracy show promising insights on factors affecting reliable estimation of object of fixation.**

I. Introduction

Driving is a visually demanding task, where drivers predominantly gather driving task relevant information by maintaining visual gaze around objects or regions of interest for a certain time period; commonly referred to as fixation. Deriving when the fixation occurs, at what the fixation is on, or how long is the fixation, are critical sources of information and can provide a greater understanding of what information the driver uses in decision making for safe driving [1]. For example, repeated fixation on the same object could indicate relative importance of the object with respect to the scene. Such knowledge, when learned properly from real-world data, can be used to analyze and assign importance to different elements of the scene, which in turn can be used in Advanced Driver Assistance Systems (ADAS) or even in autonomous vehicles for better decision making [2], [3]. However, analyzing object of fixation for large-scale real-world on-road driving data can be very tedious and time consuming if done manually. The objective of this work, therefore, is to automate the process of fixation analysis.

In particular, this work focuses on developing a machine vision framework to determine the object of fixation given the fixation. A common misconception is that once it is known where the driver's gaze is in the camera frame of reference (e.g. gaze from an eye tracking device) at a given time step, it equates to knowing at what the driver is looking (e.g. pedestrian, traffic sign). Such correlations are hard to uphold when we take into consideration a driver's $5°$ foveal field of view and the varying depth, concentration, type etc. of objects that could be encompassed in the field of view. Hence, not just the nearness of object to the gaze, but also the visual saliency, motion, depth, trajectory over time, etc. of objects are important features [4], as exemplified in Fig. 1. In this work, a subset of these object characteristics are jointly observed with gaze behavior to estimate the likelihood of an object being the object of fixation.

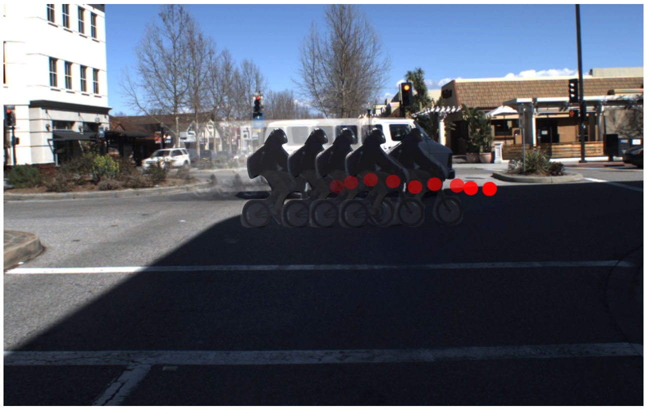


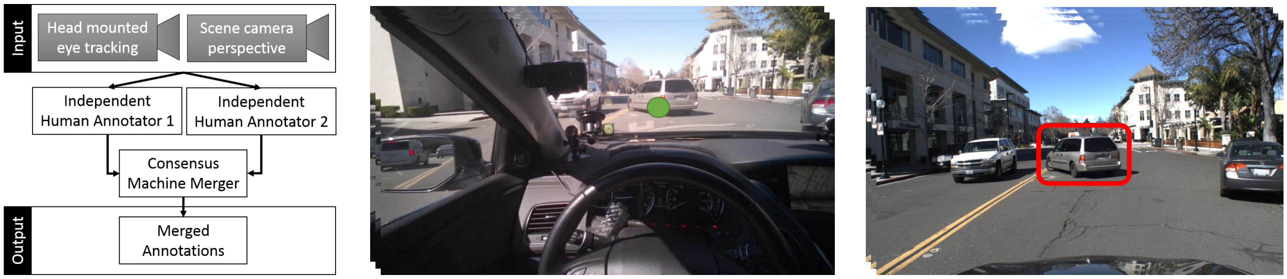
Fig. 1. Illustration of a dynamically changing driving environment where the dynamics of gaze and object characteristics has the potential to automatically infer object of fixation.

The research interest on analyzing eye movements of drivers dates back to 1970s [5]. Several of these existing studies focus on understanding visual search strategies employed by drivers for vehicle control, situational awareness (by analyzing essential visual elements), navigation etc. [6] and understanding driver distraction [7]. Motivations behind such studies include development of a better driver training program [5], [8], rules and regulation to mitigate driver distraction and even better design of the infrastructure [9]. More recently, with modern vehicles fitted with complex infotainment systems, eye movement studies are employed to design and verify in-vehicle infotainment system in terms of their influence on eye movement and visual demand, and any distractions that may arise [10], [11].

Many studies are often done in laboratory simulator setup due to the ease of conducting and controlling the experiments. On the other hand, on-road driving data with gaze information is desirable for obvious reasons but is rarer. Difficulties lie in both data collection and data analysis. During on-road driving data collection, care must be taken for proper calibration and slippage of the eye tracking device (both remote or head mounted) due to head and/or vehicle movements; eye tracking technologies are getting better and more accurate. Secondly, many of the eye movement studies employ very tedious and manual frame by frame processing

The authors are with Honda Research Institute, 375 Ravendale Dr., Mountain View, CA, USA {smartin, atawari}@honda-ri.com

978-1-5386-4451-5/18/$31.00 ©2018 IEEE


Input
Head mounted eye tracking
Scene camera perspective
Independent Human Annotator 1
Independent Human Annotator 2
Consensus Machine Merger
Output
Merged Annotations

(a) Flow diagram (b) Input to human annotator (c) Output from human annotator

Fig. 2. (a) Flow diagram of acquiring annotations of object of fixations. Two human annotations were independently tasked to annotate object of fixation given head mounted camera perspective with gaze and car mounted camera perspective. A merged set of annotations is derived automatically from mutual consent between the two human annotators. (b & c) Sample instance showing input view (i.e. driver perspective with instantaneous gaze and fixed car perspective) and output annotations (i.e. bounding boxes around object of fixation for the whole duration of fixation).

of visual data to compute relevant measurements. There is much need for effective and efficient data analysis and to this end, we present a framework to jointly analyze gaze and scene information and, in particular, to use this joint analysis to estimate object of fixation in on-road driving scenarios.

To the best of our knowledge, this is the first work of its kind to automatically estimate the object of fixation on on-road driving data, where objects are dynamic in nature. One of the main challenges involved in this study of automatically analyzing the object of fixation is in the creation of the dataset. One notable and most comparable dataset to date is the DE(eye)VE dataset [1]; however, their annotations are on pixel level, whereas our dataset annotations are on object-level. Another major challenge is the dynamics of the scene (e.g ego-vehicle can be stationary or moving, varying number and types of object in the scene). The contributions of this work, therefore, are as follows:

- A unique dataset of on-road driving centered around urban intersections with manual annotations of object of fixation.
- Presenting a spatio-temporal framework to infer object of fixation by leveraging object characteristics independent of and jointly with gaze behavior.
- An ablation study over different object characteristics and presenting their effects on the system performance.

While the approach proposed in this work can be adapted for on-line analysis, our target as it stands is for off-line processing to analyze large quantities of on-road driving data.

II. On-road Driving Dataset

A vehicular testbed capable of capturing on-road driving scene, holistically and synchronously, is used to collect the dataset used in this work. Of interest in this study are two sensing instruments: one is a car mounted camera looking out in the forward driving direction and another is a wearable eye tracking device. The latter instrument produces both a driver perspective view and where the driver is looking within the frame of reference. Fig. 2b & c show a sample synchronized instance from the dataset, where the head mounted driver perspective with gaze is shown on the left and the car mounted fixed perspective is shown on the right.

Expert drivers drove this instrumented vehicle on pre-defined routes and freestyle routes in downtown Mountain View, CA, USA area. The routes contained many types of intersections, including pedestrian crossings, stop-controlled intersections, signal-controlled intersections, roundabouts, etc. Our dataset for this work is created by extracting data from up to 25 meters before and after passing the intersection, resulting in a total of 134 intersection events.

A. Manual Annotations

Two human annotators were asked to annotate object of fixation in the dataset, independently. Each annotator had access to synchronized view of the driver perspective and fixed car perspective with instantaneous gaze overlaid on the driver perspective. The annotators were asked to carefully take into consideration the dynamic actors within the scene, rules of the road, traffic signs, road boundaries, intended/pending maneuvers, etc. Then, only when highly confident, annotators annotated object of fixation. The annotations are in the form of bounding boxes around the object of fixation, for the entire duration of the fixation. This form of annotation serves two purposes, one is to mark the beginning and end of a fixation and another is to label the object of fixation within the marked fixation window.

As these annotations are completed by two independent annotators, differences in annotations can occur in at least two ways: one is the markers identifying the beginning and end of a fixation, and second is the object of fixation given both annotators agree there is on-going fixation. Due to potential subjectivity in this task, a consensus based machine merger is developed to determine when and where annotations from the two human annotators are consistent. The merger works on a frame level where it checks the output bounding box from each annotator and acknowledges an object of fixation if there is sufficient overlap (i.e. intersection over union of the bounding boxes is over a threshold). When acknowledged, the intersecting area is taken as the new ground truth bounding box for object of fixation. Fig. 2 illustrates the work flow of the annotation process as well as the output bounding box of the object of fixation.

III. Object of Fixation Likelihood Estimation

The objective of this work is to infer object of fixation in a given fixation. The challenge is there are varying number of objects of interest (e.g. cars, pedestrians, traffic signs) on the

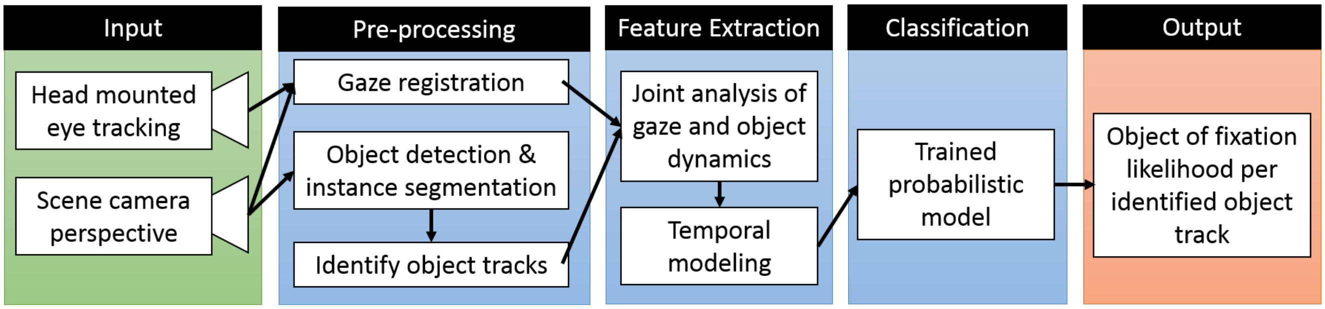

Input
Head mounted eye tracking
Scene camera perspective
Pre-processing
Gaze registration
Object detection & instance segmentation
Identify object tracks
Feature Extraction
Joint analysis of gaze and object dynamics
Temporal modeling
Classification
Trained probabilistic model
Output
Object of fixation likelihood per identified object track

Fig. 3. This diagram illustrates the system architecture, which is to identify object tracks in the scene within a window of fixation, derive object characteristics independent of and jointly with gaze behavior, and train a probabilistic model to estimate the likelihood of an object being the object of fixation.

road and there are varying number of object characteristics (e.g. eccentricity, trajectory, importance) as to why a certain object is relatively more likely to be an object of fixation over another. Our approach, given a time window of fixation, is to identify object tracks in the scene, characterize the joint dynamics of gaze and each individual object, and train a machine learning framework that learns to assign a probability to each individual object on their likelihood of being the object of fixation. Fig. 3 outlines the major building blocks of our system architecture, for which detailed descriptions are given below.

A. *Identifying Objects Tracks*

First, we analyze the scene to detect objects relevant in the driving context. For this, we use off-the-shelf Mask RCNN [12] on the video stream associated with the car mounted camera in our dataset. It is applied independently on each video frame over all intersection events, where the outputs are 2D bounding boxes of objects, their class and pixelwise instance-level segmentation.

Next, to associate the same object appearing over multiple frames in a given fixation window to a unique identification, tracking by detection is employed. Tracking is initialized at the start of each fixation by automatically assigning all detected objects to unique tracks. Each detected box in the following frames are matched uniquely to existing tracks based on overlap and when any best matched overlaps are below a certain threshold, new tracks are created. Furthermore, at every frame, tracks with no recent updates are removed and tracking stops at the end of each fixation. The following section describes how to extract characteristics over these object tracks with respect to gaze behavior.

B. *Joint analysis of Gaze and Object Dynamics*

First step towards joint dynamic analysis of gaze and object is gaze registration. As noted in Section II, gaze data is provided with respect to driver perspective from the head mounted gaze tracking device. Whereas, object level analysis is occurring in the car mounted fixed camera perspective. The following steps are taken to register gaze from driver perspective to fixed ego-vehicle perspective: dense key points are extracted from both perspectives, points between perspectives are matched based on descriptor similarity, fundamental matrix is applied to find inliers within matching pairs, and finally, homography transformation matrix is estimated with RANSAC [13] to project gaze onto the fixed camera perspective.

The projected gaze is now in the same frame of reference as the identified object tracks. For each object track, then, the following object characteristics are extracted:

- *Gaze distance to object center*: A vector of euclidean distance is computed between gaze and object center for every instance in the object track.
- *Gaze distance to object bounding box*: A vector of euclidean distance is computed between gaze and bounding box of object, for every instance in the object track. Distance is greater than zero when gaze is outside the box, but is set to zero when within the box.
- *Gaze distance to object contour*: A vector of euclidean distance is computed between gaze and contour of object, as obtained from pixel-wise segmentation, for every instance in the object track. Distance is greater than zero when gaze is outside the contour, but is set to zero when within the contour.
- *Object Importance*: Among objects in the path of the ego vehicle, a closer object at a given time is likely to be more relevant. For this, object pixel height normalized by the pixel height at a given fixed distance, respective of the object class, is used.

A combination of the above characteristics capture varying information about the joint dynamics of gaze and object, and object relevance.

C. *Temporal Modeling and Classification*

Given a time window of fixation, the tracked duration of different objects in the scene can vary and therefore the length of respective object's characteristics; reasons for varying length include the duration of fixation itself but also due to occlusion, noise in detection or tracking, etc. First, we transform a varying length input to a fixed length by computing statistics over a given tracked object duration (i.e. mean, standard deviation, min, max, range, and lower-, median- and upper quartile.)

Second, a binary Support Vector Machine (SVM) is trained per combination of object characteristics on two classes: object of fixation and background object; here any object that is not the object of fixation is considered a background object. At testing time, the output is a class membership probability [14] of an object belonging to the class of object of fixation. More details on how the dataset is divided into training and testing is given in the following section.

IV. Performance Evaluation

The system architecture is to identify object tracks in the scene within a window of fixation, derive object characteristics independent of and jointly with gaze behavior, and train a probabilistic model to estimate the likelihood of an object being the object of fixation. In this section, we present an ablation study over different object characteristics and present their effects on the system performance. All performance results are obtained by 5-fold cross-validation, where each fold comprises of unique intersection segments (separated by space or time), i.e. fixation windows from any one intersection segment are all either in training or testing.

During training, positive and negative object samples are extracted from every fixation window. Here, positive object sample refers to any object track within the fixation window whose average overlap (defined as intersection over union) with the annotated object of fixation is above a certain threshold, otherwise the track is considered a negative sample.

At testing time, each detected object track within a fixation window is assigned a probability of its likelihood of being the object of fixation. If there are any object tracks whose probability of being the object of fixation is above a certain threshold and whose average overlap with ground truth annotated object of fixation is above a certain threshold, then the fixation window associated with the object track is considered a true positive. If there are any object tracks whose probability is above a certain threshold but whose average overlap with ground truth annotated object of fixation is below a certain threshold, then every one of those object tracks is considered a false positive. Note that there are two thresholds - one for the fixation probability and another for the overlap value. A Precision- Recall curve is generated by varying the fixation probability threshold for a fixed overlap threshold (0.5 in our experiments).

A cross-examination via Precision-Recall curve and average precision is conducted for the following combination of object characteristics:

1) **Combination 1:** Gaze distance to object center with and without object importance
2) **Combination 2:** Gaze distance to object bounding box with and without object importance
3) **Combination 3:** Gaze distance to object contour with and without object importance

Figure 4 shows the precision-recall curve for the above mentioned combination of object characteristics. The average precision for these curves are $3.2\%/61.6\%$, $77.7\%/78.9\%$ and $80.8\%/83.0$ for Combination1, Combination2 and Combination3, with and without object importance, respectively. As expected Combination1, which is simply the dynamics of the gaze distance to the object center, performs poorly, whereas distance to object box (Combination 2) has significantly higher performance because it incorporates object size in addition to distance to object center. Combination 3 has higher performance to Combination 2 because it provides better separation between objects in cluttered scenes. The incremental advantages of these combinations is showcased in Fig. 5. In the figure, we illustrate one instance where distance to object contour is a better measure than distance to object box (Fig. 5, row 1) and another instance where distance to object contour plus object importance is a better measure than distance metric alone (Fig. 5, row 2).

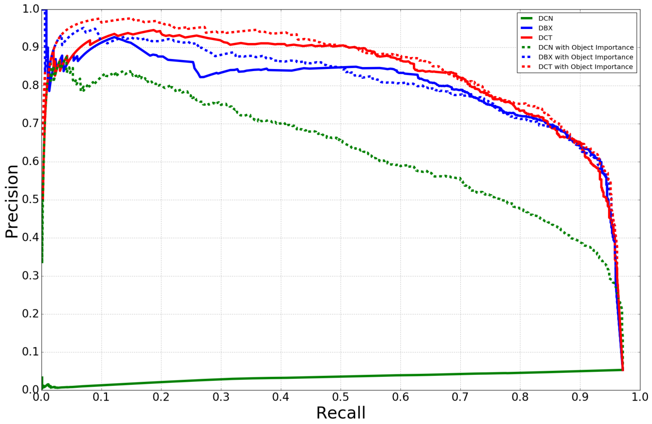


Fig. 4. Precision recall curve for cross-examining the system performance with respect to object characteristics. In the legend, DCN stands for distance to object center, DBX stands for distance to object box and DCT stands for distance to object contour.

Next we discuss challenging cases where proposed system has room for improvement. First, there are some obvious cases such as failure in proper object detection (Figure 6 row 1) - missed detection or false detection. Some object could not be detected for the whole fixation window. Hence, there is no bounding box or contour found of the true object of fixation. A false detection of object close to gaze position during the fixation window results in false object bounding box/contour and hence causes error. These cases can be resolved by improving detection and tracking module which can reject false detection and track the missed object from previous frames (e.g. by looking beyond the fixation window time frame). Another cause of the error is in erroneous gaze registration (Figure 6 row 2). When driver looks too far to the left/right, the overlap between gaze camera and car mounted fixed camera is very little causing very little image correspondence points leading to erroneous registration. Increasing car mounted camera(s) to increase field of view can possibly solve this issues.

A challenging cause for error relates to the driving context and object importance (Figure 6 row 3, column 1). An expert human annotator can evaluate driving situation in a very comprehensive manner and understand the object relevance in the current driving context. For example, a front car making a parallel parking (Figure 6 row 3, column 1) causes an ego vehicle to stop for it. An ambiguous gaze position around nearby objects can be resolved with this contextual

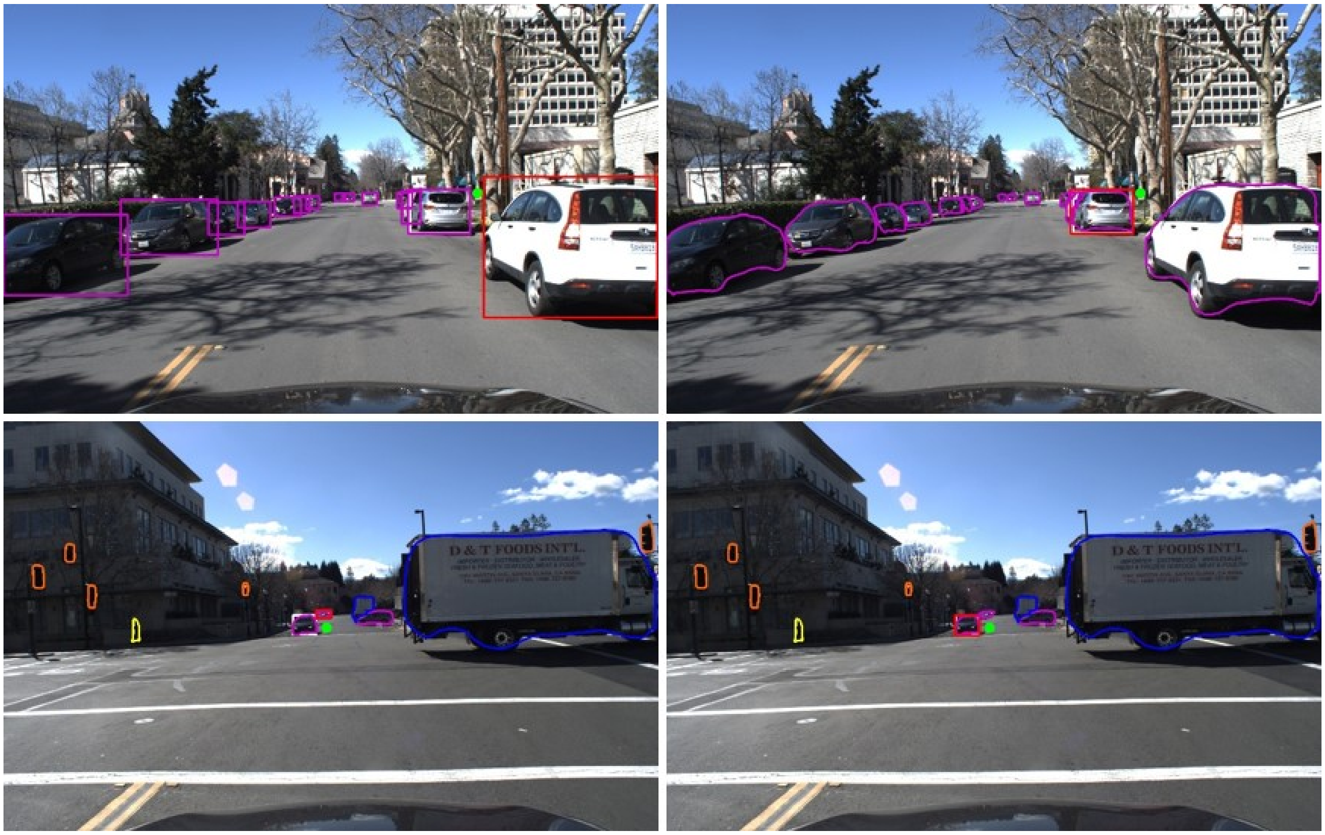


Fig. 5. This figure illustrates some incremental advantages of using object characteristics to estimate object of fixation. In each image, white bounding box represents the annotated object of fixation, red bounding box represents the estimated object of fixation, green circle represents instantaneous gaze and contours represent segmented boundaries of detected objects; different colors for contours are used to illustrate which class they belong to (e.g. car is purple, pedestrian is yellow, traffic light is orange). Row 1 illustrates two instances where distance to object contour is a better measure than distance to object box. Row 2 illustrates two instances where distance to object contour plus object importance is a better measure than distance metric alone.

information. A better measure of object importance (e.g. based on location and motion feature) can help in such cases.

Another challenging cause of failure relates to the gaze context (Figure 6 row 3, column 2). be resolved by observing gaze behavior. For example, gaze fixation going back and forth between two object locations; even though both fixations are close to these objects, it is much easier to relate which fixation is most likely associated with which object with respect to each other but in isolation it would be difficult to pin point the correct object of fixation. In our current framework, we do not incorporate such information (each fixation is treated independently). We will explore contextual information in our future work.

V. Concluding Remarks

Object of fixation estimation is a difficult problem for many reasons including the number of objects (e.g. cars, pedestrians, signs) with varying concentration in the scene and varying number of object characteristics (e.g. saliency, trajectory, importance) which influence the likelihood of a certain object being the object of fixation over another, etc. The objective of this work is to introduce a system architecture that learns to assign a probability to each individual object in the scene on their likelihood of being the object of fixation by leveraging gaze behavior and object characteristics. Experimental evaluation of this system architecture with a few combination of object characteristics gives best performance when characterizing object with gaze distance to object contour plus object importance; this combination achieved an average precision of 83%. To better leverage the rich driving context, future studies will be in the direction of increasing the dictionary of object characteristics (e.g. motion, visual saliency, trajectory) and improved modeling of temporal dynamics of gaze behavior and object characteristics.

References

[1] S. Martin, S. Vora, K. Yuen, and M. M. Trivedi, “Dynamics of driver's gaze: Explorations in behavior modeling maneuver prediction,” *IEEE Transactions on Intelligent Vehicles*, 2018.

[2] A. Palazzi, F. Solera, S. Calderara, S. Alletto, and R. Cucchiara, “Learning where to attend like a human driver,” 2017.

[3] A. Tawari and B. Kang, “A computational framework for driver's visual attention using a fully convolutional architecture,” 2017.

[4] E. Ohn-Bar and M. M. Trivedi, “Are all objects equal? deep spatio-temporal importance prediction in driving videos,” *Pattern Recognition*, 2017.

[5] R. R. Mourant and T. H. Rockwell, “Strategies of visual search by novice and experienced drivers,” *Human factors*, 1972.

[6] A. Doshi and M. M. Trivedi, “Head and eye gaze dynamics during visual attention shifts in complex environments,” *Journal of vision*, 2012.

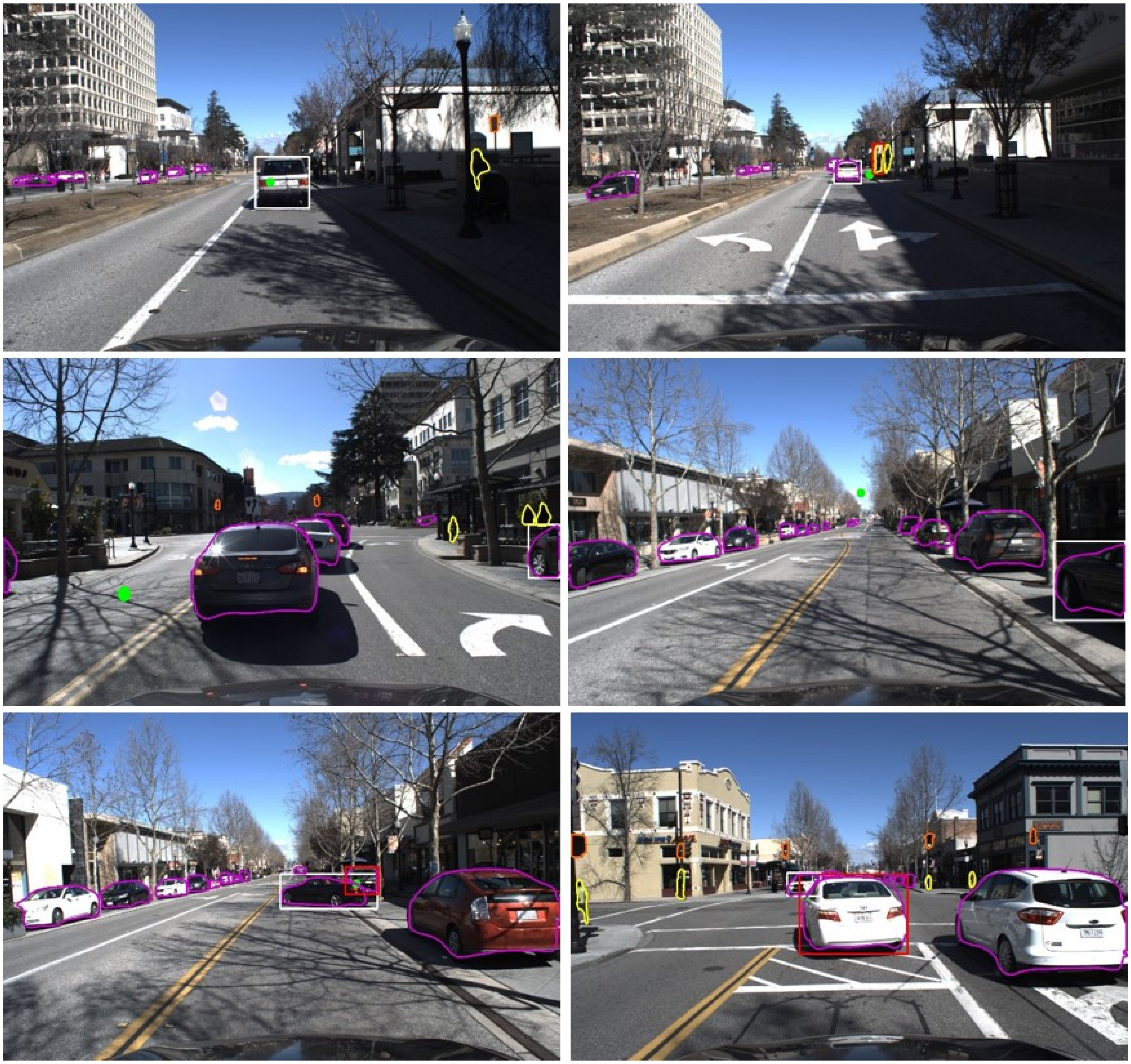


Fig. 6. This figure illustrates some challenging cases where our proposed system failed. In each image, white bounding box represents the annotated object of fixation, red bounding box represents the estimated object of fixation, green circle represents instantaneous gaze and contours represent segmented boundaries of detected objects; different colors for contours are used to illustrate which class they belong to (e.g. car is purple, pedestrian is yellow, traffic light is orange). Row 1 highlights some cases of missed or false detection of object of fixation. Row 2 depicts instances where gaze registration can lead to estimating object of fixation incorrectly. Row 3 illustrates context rich dynamics that human annotators leverage in order to determine object of fixation, whereas the current system has room to improve by leveraging similar information.

[7] W. Horrey and C. Wickens, "In-vehicle glance duration: distributions, tails, and model of crash risk," *Transportation Research Record: Journal of the Transportation Research Board*, 2007.

[8] G. Underwood, P. Chapman, N. Brocklehurst, J. Underwood, and D. Crundall, "Visual attention while driving: sequences of eye fixations made by experienced and novice drivers," *Ergonomics*, 2003.

[9] K. Abdelgawad, J. Gausemeier, R. Dumitrescu, M. Grafe, J. Stöcklein, and J. Berssenbrügge, "Networked driving simulation: Applications, state of the art, and design considerations," *Designs*, 2017.

[10] M. AblaBmeier, T. Poitschke, F. W. K. Bengler, and G. Rigoll, "Eye gaze studies comparing head-up and head-down displays in vehicles," 2007.

[11] C. Purucker, F. Naujoks, A. Prill, and A. Neukum, "Evaluating distraction of in-vehicle information systems while driving by predicting total eyes-off-road times with keystroke level modeling," *Applied ergonomics*, 2017.

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[13] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," in *Readings in computer vision*. Elsevier, 1987.

[14] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, 1999.