A Computational Framework for Driver’s Visual Attention Using A Fully Convolutional Architecture

Ashish Tawari¹ and Byeongkeun Kang²

Abstract—It is a challenging and important task to perceive and interact with other traffic participants in a complex driving environment. The human vision system plays one of the crucial roles to achieve this task. Particularly, visual attention mechanisms allow a human driver to cleverly attend to the salient and relevant regions of the scene to further make necessary decisions for the safe driving. Thus, it is significant to investigate human vision systems with great potential to improve assistive, and even autonomous, vehicular technologies. In this paper, we investigate driver’s gaze behavior to understand visual attention. We, first, present a Bayesian framework to model visual attention of a human driver. Further, based on the framework, we develop a fully convolutional neural network to estimate the salient region in a novel driving scene. We systematically evaluate the proposed method using on-road driving data and compare it with other state-of-the-art saliency estimation approaches. Our analyses show promising results.

I. INTRODUCTION

One of the most fascinating capabilities of human drivers is their ability to seamlessly perceive and interact with traffic participants in a complex driving environment. Human vision plays a critical role in perceiving the environment that further leads to understanding of the scene and ultimately to a suitable vehicle control behavior.

For safe driving, drivers need to allocate their attention to the most important and salient regions or objects. However, why driver sees where/what it sees is an intriguing problem that is still far from being fully understood. In fact, currently existing computational frameworks lack the ability to accurately mimic a driver’s gaze behavior and estimate saliency in different complex driving environments. Nevertheless, traffic saliency detection, which computes the important and relevant regions or objects in a given driving environment, is an important component of intelligent vehicle systems and could be useful in supporting autonomous driving, traffic sign detection, driver training, car collision warning etc.

In this paper, we investigate which region in a scene attracts driver’s attention. Here, we make the assumption that the driver’s gaze provides the region of attention, leaving aside psychological effects such as in-attentional blindness, looked-but-did-not-see etc. Our goal is to predict driver’s eye fixations in the real-world driving scene. Towards this end, we present a Bayesian framework to model visual attention of the driver. Furthermore, we develop a deep neural network to predict gaze fixation and evaluate the performance of the system using on-road driving data.

¹Author is with Honda Research Institute, 375 Ravenade Dr, Suite B, Mountain View, CA, 94043, USA. atawari@hra.com
²At the time of this work, the author was with Honda Research Institute, 375 Ravenade Dr, Suite B, Mountain View, CA, 94043, USA.

II. BACKGROUND AND RELATED WORK

Driver (in)attention studies have been actively carried out for decades in multiple disciplines from cognitive science and psychology to engineering, for their potential to save lives. Several recent approaches utilize driver monitoring systems to estimate driver’s gaze direction or gaze-zones from head pose and/or eye location cues [1], [2], [3], [4]. Their common aim is usually to detect the driver’s visual attention away from the road as it is a critical indicator of accident risk [5]. Other studies have used driver’s attentional behavior and the road scene information to estimate driver’s intent and potentially dangerous maneuvers [6], [7], [8]. Interestingly, [9], without any driver monitoring information, shows a strong correlation between visual appearance of the scene and the driver’s action including steering and braking.

In this work, our focus is to develop a framework for driver’s visual attention that can predict gaze fixations using the external scene information. An ideal system based on our framework would be able to estimate where should a driver allocate his/her visual attention in a scene while driving.

Visual attention, in general, refers to mechanisms that cleverly select important and relevant regions of a visual field to allow subsequent complex processes (e.g. object recognition) feasible in real-time. Modeling visual attention is a very active research area for the last several decades. Several theoretical and computational models have been proposed to explain eye movements (fixation/saccades) and have shown promising results in simple tasks and laboratory settings. However, they are yet not reliable to mimic human gaze behavior in complex and naturalistic settings such as driving. Here, we discuss selected studies that bring important concepts of visual attention that are applicable to the driving context. For a review on computational models of visual attention in general, we encourage readers to refer to [10], [11], [12], [13]

In literature, it is widely accepted that visual attention is guided by the combination of bottom-up and top-down mechanisms. Bottom-up cues are influenced by external stimuli and are mainly based on the characteristics of a visual scene such as image-based conspicuities [14], whereas top-down cues are goal-oriented where task, knowledge, memory, expectations, etc., guide gaze toward relevant/informative scene regions.

A considerable amount of research has been conducted to study bottom-up saliency based attention. Bottom-up saliency intuitively characterizes some parts or events in the visual field that stand out from their neighboring background.
For example, in the driving context, objects that pop out against the background due to high relative contrast, such as retroreflective traffic signs or events such as the flashing of indicators of a car, onset of tail brake light, etc., are salient.

Top-down attention, on the other hand, is task-driven or goal-oriented. In a classic study of top-down attention by Yarbus [15], subjects were asked to watch the same scene (a room with a family and an unexpected visitor entering the room) under different tasks such as determining the ages of the people or simply freely observing the scene etc. The experimental results showed considerable differences in the eye movements and fixations. This makes the modeling of top-down attention a conceptually hard problem since different tasks require different algorithms. Doshi and Trivedi [16] intuitively modeled top-down maps for goal-oriented behavior in the lane-keeping and lane-changing tasks. For the lane-keeping task, the front of the vehicle and for the left/right lane-changing task, the respective side of the vehicle are set to be the more salient regions.

Driving is a complex dynamic environment where different top-down factors evolving over time play a very active role in governing the gaze behavior. Factors such as planning of a maneuver (e.g. turning left/right, taking the next exit, etc.), knowledge of traffic laws, expectation of finding other road participants in a given location, etc., compete with bottom-up events as mentioned above, and can greatly influence gaze behavior. Hence, there is a need to incorporate both the bottom-up and the top-down factors while modeling the visual attention. Zhang et al. [17] presented a Bayesian framework to incorporate bottom-up and top-down influences. In this framework, bottom-up saliency emerges naturally as the self-information of visual feature which is similar to other works in Oliva et al. [18], Torralba [19] and Bruce and Tsotsos [20]. The overall model incorporates the top-down influences in terms of the knowledge of the target’s appearance and locations-priors (expectation of the presence of targets in a particular location).

Inspired by these studies, we present a Bayesian framework to incorporate task dependent top-down and bottom-up factors in modeling driver’s visual attention. Our contributions are: first, we model visual saliency using a fully convolutional neural network to predict driver’s gaze fixations; second, we perform thorough evaluations and conduct comparative studies using on-road driving data; and finally, we illustrate the top-down influence of different “tasks” as inferred from the vehicle state information.

III. VISUAL ATTENTION DURING DRIVING:
A BAYESIAN FRAMEWORK

We represent saliency $s_z = p(O = 1|F = f_z, L = l_z)$, where $z$ is a point in the visual field of the driver. A Point here is loosely defined and in this work, it refers to a pixel in the frame of the scene camera. $f_z$ and $l_z$ represent the visual features and the location $(x, y)$ of the point $z$, respectively. $O$ is a binary variable where $O = 1$ represents the presence of objects/regions, henceforth called targets, relevant to the driving scenario. So, the higher the probability of the relevant targets at a point $z$ is, the more salient the point $z$ becomes.

Driving environment is a highly dynamic environment that consists of different tasks at different point in time such as car following, lane-keeping, turning, lane-changing, taking an exit, etc. The same driving scene with different tasks in mind can influence gaze behavior as described in earlier by the illustrative Yarbus’s experiment. We model such influences due to the different tasks explicitly in our framework. Let $T$ be a discrete random variable drawn from the space of all tasks, $T \in \{T_1, T_2, \ldots, T_n\}$. Then,

$$s_z = \sum_{T_i} p(O = 1, T = T_i|F = f_z, L = l_z) = \sum_{T_i} \frac{p(O = 1|f_z, l_z, T_i) p(T_i)}{s_z(T_i)}$$

(1)

Looking closer to the first component of the right-hand side (abbreviated as $S_z(T_i)$ due to the space constraint) of Equation 1, using Bayes rule:

$$S_z(T_i) = p(O = 1|f_z, l_z, T_i) = \frac{p(f_z, l_z|O = 1, T_i) p(O = 1|T_i)}{p(f_z, l_z|T_i)}$$

(2)

We further make a simplifying assumption that the features and the locations of the point $z$ are conditionally independent. This can be interpreted as the features’ distribution does not change with the location across scene regardless of whether or not it appears on the target during any given task. Because of the space constraint, we also abbreviate $O = 1$ as $O$. Equation 2 can then be decomposed into meaningful components as below:

$$S = \frac{p(f_z, l_z|O, T_i) p(O|T_i)}{p(f_z, l_z|T_i)} \approx \frac{p(f_z|O, T_i) p(l_z|O, T_i) p(O|T_i)}{p(f_z|T_i) p(l_z|T_i)} = \frac{1}{p(f_z|T_i)} \left( \frac{p(f_z|O, T_i) p(l_z|O, T_i) p(O|T_i)}{p(l_z|T_i)} \right)$$

(3)

The first component of Equation 3 is referred as bottom-up saliency as it does not depend on the target. Also, notice that the less probable the feature of the point $z$ is, the more salient the point $z$ becomes. In other words, rare features are salient. The second component of the Equation 3 depends on target and related knowledge and hence is referred as top-down saliency. The first part of this component encourages features that are found in targets. In other words, features that are important are salient. The second part of this component encodes the knowledge of targets’ expected location, referred as a location-prior. From a driving perspective, this entails
that the driver develops a prior expectation of relevant targets in a particular location of the scene while executing a particular task (e.g. checking side mirror or looking over shoulder while changing lane).

Accurately learning the high dimensional feature distributions as in $p(f_z|T_i)$ and $p(f_z|O,T_i)$ is difficult. Hence, we rearrange first two terms in Equation 3 using Bayes rule as follows:

$$ S = \frac{1}{p(f_z|T_i)} p(f_z|O,T_i) p(O|l_z,T_i) $$

$$ = \frac{p(f_z, O|T_i)}{p(O|T_i)p(f_z|T_i)} p(O|l_z,T_i) $$

$$ = p(O|f_z, T_i) p(O|l_z,T_i) p(O|T_i)^{-1} \quad (4) $$

The last term of Equation 4, $p(O|T_i)$ is the prior probability of the target class given a particular task, and in this work, is assumed to be uniform and hence a constant.

Thus, the overall saliency from Equation 1 becomes

$$ s_z \propto \sum_{T_i} p(O|f_z, T_i)p(O|l_z,T_i)p(T_i) $$

$$ s_z = \frac{1}{Z} \sum_{T_i} p(O|f_z, T_i)p(O|l_z,T_i)p(T_i) \quad (5) $$

where $Z$ is a normalizing factor. We learn $p(O|f_z, T_i)$ and $p(O|l_z,T_i)$ from the driving data. In particular, we model $p(O|f_z, T_i)$ using fully convolutional neural network. Further details including network architecture are discussed in the next section. For $p(O|l_z,T_i)$, we learn the location-prior for each task. It modulates the salient regions using the weights that are estimated based on the learned prior distribution. The definition of the “task” and the learned prior distribution are described in the Section V-B.

IV. NETWORK ARCHITECTURE AND TRAINING DETAILS

It’s very clear from the literature that the choice of architecture is a very important factor when it comes to utilizing a neural network framework. Hence, we give a careful look to the requirements of the problem at hand.

Modeling $p(O|f_z, T_i)$ is to learn the weights for the feature vector in a given “task” $T_i$ to discriminate between the target classes (salient vs not-salient). Since point $z$ refers to a pixel location, it becomes a pixel-wise classification problem. For the driving data, however, a longer fixation at a point is interpreted as receiving more attention to the point by the driver, and hence more salient. Thus, instead of classification, we model saliency as a pixel-wise regression problem. It can be argued that classification score can provide the strength of saliency, but such requirement is not enforced explicitly in the classification problem formulation.

Another important consideration is that the local conspicuity aspects of saliency inherently require the analysis of surrounding background. In other words, local features are not analyzed independently but instead in connection with the surround features. From network architecture’s perspective, this can be achieved by skip connections [21]. Skip connection allows the early feature response to directly interact with the later feature response which often works with the downsampled version (e.g. due to intermediate max-pooling layer as in our case) of the earlier maps. Hence, the later feature response covers a bigger area around a pixel in the original input frame for the same size of the receptive field (as in our case).

Also, saliency datasets [22], [23], [24] reveal a strong center bias of human eye fixation in free viewing image and video frames. In fact, simply using a Gaussian blob centered in the middle of the image frame as the saliency...
Fig. 2: Location-priors learned for the different “tasks” as inferred from the yaw rate. Top row and bottom row show the effects of negative yaw rate (turning-left) and positive yaw rate (turning-right), respectively. Also, as the magnitude of yaw rate increases, location-prior shifts away from the center.

map produces excellent results [25], [26], [17]. Even from the driving data perspective, the driver tends to pay attention to the front for most of the time. Therefore, it’s important to make sure that our model is not learning trivial center-bias solution.

Based on the above requirements, we choose a Fully-Convolutional Network (FCN) for our problem. In general, Convolutional Neural Networks (CNNs) have shown state-of-the-art performance in many computer vision problems [27], [28], [29], [30], [31]. For image-level classification such as image recognition, Simonyan and Zisserman presented a new architecture with 16 layers which achieved state-of-the-art performance [32] [28]. To apply the architecture designed for image-level classification to semantic segmentation (a pixel-level classification), Long et al. proposed fully convolutional networks by converting fully connected layers to convolutional layers. Fully convolutional networks take the input of an arbitrary size and produce a correspondingly-sized output [29], [30]. Also, fully convolutional networks (with no fully connected layer) treat each image pixel, irrespective of its location, identically. As long as the receptive field of the convolutional layers is not too big to cause edge effects (e.g., an extreme case would be when the size of the receptive field becomes same as the size of its input layer), the model does not have any way to exploit location information and so the center bias effect.

Since we approach the saliency estimation task as a pixel-wise regression problem, we adapt the fully convolutional network for the regression problem. Particularly, we employ FCN-8s architecture by [30] that has multiple skip connections with minor modifications such as changing the score layers to reflect a single channel saliency score and the loss layer for the regression. See Figure 1 for more details. For loss function, we used L2 loss $L$ as follows:

$$L = \frac{1}{2N} \sum_{n=1}^{N} ||\hat{y}_n - y_n||_2^2 \quad (6)$$

where $N$ is the total number of data, $\hat{y}$ is the estimated saliency, and $y$ is the targeted saliency.

We followed the training strategies as in [29], [30] such as fixed deconvolutional layer with bilinear up-sampled filter weights. We initialized our model using the weights of the fully convolutional network (FCN-8s) trained using the PASCAL VOC segmentation dataset [33], except for the saliency-score layer of the model, which is initialized randomly. Then, we trained for our saliency estimation task using DR(eye)VE training dataset [34]. During the training, we used the full image size $1920 \times 1080$ and the learning rate $10^{-12}$, with momentum and weight decay values as 0.99 and 0.0005, respectively.

V. EXPERIMENTAL EVALUATIONS

A. Dataset

In our experiments, we used DR(eye)VE dataset which consists of 74 sequences of 5 minutes each [34]. The dataset provides videos from a roof-mounted camera and a head mounted camera, captured gaze locations from a wearable eye tracking device, and other information from Global Positioning System (GPS) related to the vehicle status (speed, course, latitude, longitude, etc.). The captured gaze location is further processed using a spatio-temporal Gaussian $G(\sigma_s, \sigma_t)$, with $\sigma_s = 200$ pixels and $\sigma_t = k/2$, where $k = 25$ frames, to acquire the smoothed ground truth saliency map.

The dataset was collected from eight drivers, in different areas (downtown, countryside, and highway), under different weather conditions (sunny, cloudy, and rainy), and at
different times of the day (morning, evening, and night). We followed the dataset separation for training and testing (the first 37 sequences for the training and the last 37 sequences for the testing). In all the experiments, we exclude the frames with errors which are provided by the authors of the dataset. For training, we also exclude any frame when the vehicle is stationary. This is because generally when the vehicle is not moving, the driver is not expected to be attentive of the driving related events. Though, the driver, of course, would become attentive just before the motion for its preparation and this is usually captured in the successive frames when the speed becomes just greater than zero.

### B. Results

We earlier argued that during driving, tasks such as lane changing, turning left/right, exiting highways, etc., influence top-down attention. We suggested to learn the probability distributions $p(O|f_z, T_i)$ and $p(O|z, T_i)$ conditioned upon these tasks. Hence, we learn these distributions from the portion of the dataset when the driver is engaged in such tasks. The dataset, however, does not have such task information currently. We, therefore, define these “tasks” based on vehicle dynamics. Particularly, we divide the dataset based on yaw rates. Yaw rate is indicative of events such as turning (right/left), exiting, curve-following, and provides, though simplistically, a reasonable and an automatic way to infer task contexts. In the dataset, since yaw rate is not recorded directly from the Controller Area Network (CAN) interface of the vehicle, we compute this from the course measurement provided by the GPS. The course is the same as the heading for ground vehicles such as cars and the yaw rate is estimated as the rate of the change of the heading.

We divide the dataset into discrete intervals of the yaw rate with the bin size $5^\circ$/sec. Then, the location-prior, $p(O|z, T_i)$, is calculated as the average of all the attentional maps within a bin in the training set. Figure 2 shows yaw rate effects on the estimation of the location-prior. Notice that as the magnitude of yaw rate increases, the location-prior becomes more and more skewed towards the edges (away from the center). Also, positive yaw rates (turning-right events) shift the location-prior towards the right of the center and the opposite for negative yaw rates (turning-left events).

Learning $p(O|f_z, T_i)$ requires training the neural network. However, as the magnitude of yaw rate increases, the size of the dataset for training within a bin dramatically decreases. This makes the learning of the deep neural network difficult. So, we approximated $p(O|f_z, T_i)$ to $p(O|z)$ by taking all the data for this component.

For quantitative analysis, we computed linear correlation coefficient (CC) which is also known as Pearson's linear coefficient between estimated saliency map and ground truth saliency map. For each saliency map $s$, we first normalize it as follow:

$$ s'_z = (s_z - \overline{s}) / \sigma(s) $$

where $\overline{s}$ represents the mean of saliency map $s$, and $\sigma(s)$ is the standard deviation of $s$, and $z$ is the pixel in the scene camera frame. Then, CC is computed as follow:

$$ CC = \frac{\sum_z (s'_z - \overline{s'})(\hat{s}'_z - \overline{s'})}{\sqrt{\left(\sum_z (s'_z - \overline{s'})^2\right) \left(\sum_z (\hat{s}'_z - \overline{s'})^2\right)}} $$

where $s'$ represents the normalized ground truth saliency map, and $\hat{s}'$ is the normalized estimated saliency map.

For a baseline, we compute the performance with the center-bias-filter, see Figure 3. As discussed earlier, center bias, i.e. fixation at the center of the image frame, can provide excellent results. In fact, such a baseline beats many benchmark methods in the MIT [39] dataset. This is a special case in our formulation and corresponds to the location-prior component (i.e. ignoring visual feature information in Equation 3). Thus, we compare our approach with this simple yet effective baseline. We also compare our framework with a recent deep neural network based approach reported using the above dataset [38] as well as other traditional well-known image saliency approaches by [37], [35], [36]. In all our results, we have removed frames for the analysis that are marked as either erroneous or having blink or saccade events as provided by the ground-truth annotations in the dataset.
TABLE I: Test results obtained by the baseline, traditional bottom-up saliency methods and the proposed approach. Result in the parenthesis is obtained by incorporating the learned location-priors.

<table>
<thead>
<tr>
<th>Method</th>
<th>Baseline-Center</th>
<th>Itti [35]</th>
<th>Image Signature [36]</th>
<th>GBVS [37]</th>
<th>DR(EYE)VE [38]</th>
<th>Proposed Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.47 ± 0.24</td>
<td>0.16 ± 0.10</td>
<td>0.14 ± 0.12</td>
<td>0.20 ± 0.10</td>
<td>0.55 ± 0.28</td>
<td>0.55 ± 0.28 (0.55 ± 0.28)</td>
</tr>
</tbody>
</table>

Fig. 4: Effects of the location-prior on the CC score of the test sequences with yaw rate $> 15^\circ$/sec.

Table I shows the performance comparison of the proposed method with other approaches. Overall, our proposed network achieves 0.55 CC score. The traditional methods, on the other hand, show no correlation (CC $< 0.3$). Even the baseline result, which corresponds to a simple top-down cue in our formulation, performs much better. This finding is somewhat expected as the traditional methods purely rely on bottom-up cues based on local conspicuities. It further strengthens and highlights the fact that as experimental settings become more complex and less unnatural like the real-world driving, bottom-up cues alone based methods perform very poorly. Our proposed approach significantly outperforms the baseline as well as the traditional approaches. It achieves the state-of-the-art results reported so far in this dataset [38]. Our method, however, uses a single frame to predict the fixation region, as opposed to a sequence of frames by [38]. Hence, it is also computationally much more efficient.

Additionally, it can be noticed that the location-prior information seems to not improve overall performance. Giving a closer look, we found out that the majority of the test dataset consists of going-straight (i.e. lane following) events as oppose to turning or exiting events where we expect the location-prior to improve performance. Hence, we looked into the cases where the yaw rate is greater than $15^\circ$/sec. Among these cases, we found that those cases with the velocity greater than 10km/h shows 10 percentage improvement over using visual feature only, see Figure 4. These are, in fact, the cases where the driver is actively involved in maneuvers such as taking a turn (left/right) or taking an exit.

Fig. 5: Saliency score vs. Velocity: Each point presents the average correlation coefficient of the frames with the velocity greater than given velocity.

We also found out that as the velocity increases, the performance improves to the CC score of $\sim 0.70$ for the velocity greater than 100km/h, see Figure 5. This fact can be explained as, because naturally the driver is more focused and less distracted by other unrelated events while driving at high speed. Thus, the driver tends to constantly follow road features like lane markings which are very well captured by our learned network. Also, if we exclude frames when the vehicle is stationary, the performance further improves by $\sim 5\%$ (the second data point on the curves in the Figure 5). This can be attributed to the fact that when the vehicle is not moving, drivers can look around freely to non-driving events. Hence, excluding these frames with spurious gaze ground truth improves the performance.

VI. DISCUSSIONS AND FUTURE DIRECTIONS

In this work, we presented a Bayesian framework to model visual attention of a human driver. We further trained a fully convolutional neural network to learn features that are important to replicate driver’s gaze behavior. Giving a closer look at the network’s output, we found out that our learned model can capture very well those road features that attract the driver’s attention (Figure 6). Specifically, the vanishing point of the lane markings affects driver’s gaze behavior and our network is able to learn those meaningful representations. In fact, it’s been suggested that following vanishing point is
Fig. 6: Qualitative results of our approach along with the state-of-the-art methods GBVS [37], ITTI [35] and Image Signature [36] for the prediction of driver’s eye fixation during different “tasks”. Last column shows the ground truth fixation map (GT).

indeed an optimal strategy to mimic driver’s gaze behavior [40]. Also, from the on-road dataset, looking to the road is the most frequent event [38]. While such behavior is natural, they are somewhat trivial and of little interest (even the simple baseline shows decent results which have nothing to do with the driving scene events). However, there exist other events, though comparatively infrequent, that requires driver to actively seek surrounding information and react either because driver has a goal in his/her mind (e.g. next turn, stop sign, traffic lights) or due to other road users (e.g. crossing pedestrian, front-vehicle braking, etc.). There is a need to collect more of such reactive events for both training and evaluation purposes.

From the gaze data, it is clear that the current driving “task” is an important factor. For example, whether the driver is planning to take the imminent exit or not will influence his/her gaze behavior (row 6 from the top in Figure 6). Using only visual features, such factors cannot be incorporated to mimic the gaze behavior. We modeled such task-oriented expectation using location-prior. In general, any information independent of visual features can be incorporated as prior information and learned from the data.
Finally, in future, we will investigate the joint modeling of visual and other information from CAN, GPS, map/navigation route, etc. in a multi-modal neural network framework. We also plan to collect more comprehensive on-road data to highlight non-trivial reactive events.

References