Analysis of Driver Gaze and Attention to Traffic Signs

Shabnam Shabani, Steven Beauchemin, and Michael Bauer

Department of Computer Science, The University of Western Ontario, London N6A5B7, Canada

Correspondence should be addressed to Michael Bauer; bauer@uwo.ca

Received 22 March 2021; Revised 27 January 2022; Accepted 1 March 2022; Published 16 April 2022

Academic Editor: Mehdi Keyvan-Ekbatani

Copyright © 2022 Shabnam Shabani et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A driver’s actions and intent can be factors in enabling advance driver assistance systems (ADASs) to assist drivers and avoid accidents. A driver’s gaze can provide insight into the driver’s intent or awareness of situations. Knowing that a driver gazed at a traffic sign or missed a traffic sign could provide indications of whether the driver is alert to impending changes in the driving environment, such as curves and stop signs. For ADASs to determine the importance of a driver seeing or missing a sign, it is important to understand the driving environment and situation. A first step is to understand what signs drivers do see or miss while driving. This contribution presents the results of analyzing driving sequences to assess traffic signs that drivers may or may not have gazed upon. The results suggest that drivers may miss 20% of traffic signs though the percentage varies depending on the type of sign. The analysis uses image sequences of the driving environment and gazed data captured during driving. The methods used in our analysis included determining whether a driver’s gaze has fallen on the image of a traffic sign or not and subsequently determining signs missed during driving. The methods presented can be useful in other scenarios involving the analysis of driver gaze and have implications for the design of future ADASs and for understanding of driver gaze and awareness.

1. Introduction

With increasing numbers of vehicles and pedestrians, drivers are in need of the safety measures to decrease the possible risk of accidents. Advanced driver assistance systems (ADASs) have significantly progressed in many aspects and vehicles today already include technology that can assist a driver, such as advanced cruise control, automatic parking, collision avoidance, automatic braking, and lane departure and warning systems. These methods focus on assessing the driving environment and the vehicle’s operation in that environment in order to help the driver.

Further assistance could be provided by inferring a driver’s intention. Certain aspects of a driver’s intention could be determined by considering the driver’s operation of the vehicle; e.g., signalling with a turn signal is a good indicator that driver is planning to turn the vehicle (though this is not always the case). Driver’s gaze can also provide insight into the driver’s intention or awareness of their environment.

Traffic signs provide important information about the current driving environment or impending driving environment, e.g., a sharp curve is ahead or whether there is a railroad crossing ahead. Drivers that are aware of and alert to traffic signs while driving are more likely to be prepared than those that are distracted and miss seeing signs. However, do drivers need to “see” all signs? Likely not. Many signs are informational, e.g., where to park and where to exit from a highway. These signs are likely ignored except in those driving contexts where they may be of importance to the driver. Other signs are more important for ensuring safe driving, e.g., stop signs and yield signs or signs indicating traffic direction. Drivers not seeing such signs may endanger themselves or others.

For ADASs to be effective, we must understand the contexts in which a driver missing a sign may be an issue. A key first step is to have some insight into the signs drivers miss while driving. Perhaps the percentage is relatively small and something not to be concerned about? Perhaps some types of signs are missed more often than others? This may also help answer more complex questions such as whether the location of the sign relative to the driver impact whether the sign is seen or not?
The work reported in this study provides insight into answering some of these questions and contributes to our overall understanding of driver behavior. Our focus is understanding what traffic signs a driver may see, duration of gaze, and any differences among types of signs. Our analysis is based on a naturalistic dataset where we make use of traffic sign detection and recognition methods along with gaze information to identify traffic signs that fall under a driver’s gaze. If one can determine that a driver’s gaze does not fall on a sign, then it is likely that the driver missed the sign. If a driver’s gaze does fall on a sign, then there is a strong likelihood that the driver may have “seen” the traffic sign. The notion of whether a sign has been “seen” is far more complex than whether the driver’s gaze has fallen upon it as it also entails some aspect of cognition or awareness. Neuroscientists have explored this for many years, and the general notion is that the order of 250 ms is needed for an eye to fixate and react to a stimulus, with recognition taking more time and affected by the complexity of the scene and object [1, 2]. For this work, we focus strictly on gaze; whether the object is recognized by the driver is beyond the scope of this study. The research presented in this study provides some results on drivers and signs that have been missed.

Data for this research come from the RoadLab project [3]. RoadLab is an initiative to collect data and develop methods for the development of ADASs. The RoadLab data for this work were collected by an in-vehicle laboratory instrumented with an on-board diagnostic system using the CANbus protocol, cameras for collecting video sequences of the driving environment in front of the vehicle, cameras for collecting the ocular behavior of the driver such as driver gaze, and other information, such as GPS. The work reported here is based on the video sequences and ocular data from ten drivers.

Our primary contributions are results on the numbers and percentages of traffic signs missed and by sign type in an urban environment. We also introduce a method for tracking gaze across several images and so are able to estimate how long a driver gazed at a particular sign; given the neuroscience research on eye fixation and recognition, this is important in considerations of whether a driver actually recognized a sign. A secondary contribution is the overall approach used for the analysis as it is an approach that can be used for understanding characteristics of traffic signs that may impact driver recognition and as an approach for similar studies on driver gaze. The work presented here is an extension of some previous work by Shabani [4].

The remainder of the study is structured as follows. Section 2 provides an overview of related work on driver attention and gaze. This is followed by Section 3 that outlines the methodology used in the study and describes the driving data and methods used, including traffic sign detection and classification, determination of the point of gaze and how it is mapped to individual images, and determination that a sign has fallen under the gaze of a driver given that a single sign may span multiple images. The results are then presented and discussed. The study concludes with a summary of key results and considerations for subsequent work.

2. Related Work

Previous researchers have recognized that driver gaze and head position are key indicators of someone safely driving a vehicle. Much of this work has been concerned with driver attention, or more precisely, driver inattention or distraction [5–7]. To determine a driver’s attention or distraction, it is crucial to extract information about the head position and gaze [8, 9]. Through the information about the position of the head, we can estimate the driver’s field of view and current focus of attention. It is obvious that visual gaze direction and head position are connected intrinsically. We can take advantage of this feature when the eyes are not visible, and the direction of eyes can be estimated by the head pose. Langton et al. [9] used the combination of gaze direction and head pose to get the gaze information.

Fletcher and Zelinsky [10] proposed an approach to estimate a driver’s area focus from within the vehicle using the integration of eye-gaze tracking and road scene events. Ishikawa et al. [11] also considered the direction of the driver’s gaze mapped on to the road scene, though specific features of the road were not determined. Jabon et al. [12] aggregated a large number of facial features in pre-accident intervals as key features to predict accidents. Takemura et al. [13] found relationships between head-gaze direction and driving tasks, such as awareness of pedestrians walking or the presence of traffic signs.

One way to monitor driver behavior is to use a combination of eye-gaze tracking and data from an instrumented vehicle. Some research has looked at gaze direction detection with different sensors and for a variety of applications. Trivedi et al. [14] used a combination of driver head pose and vehicle information in an analysis of driver behavior and areas of the road observed. Vincente et al. [6] use a vision-based system to detect eyes off the road. They make use of facial feature tracking, head post, gaze estimation, and a 3D face model. Palazzi et al. [15] looked at developing models to predict the driver’s focus of attention while driving. They used a naturalistic driving dataset of driving scenes which included eye-tracking annotations. The dataset contains images from roof-mounted cameras and gaze information from glasses worn by drivers.

Simon et al. [16] proposed a model of visual saliency linking the size of an object and a size-independent saliency. The work focused on detecting traffic signs and then estimating the visual saliency. Fletcher et al. [17] investigated whether speed limit signs had been missed by the drivers or not. In [18], Chan et al. studied traffic sign detection and recognition where they considered eye movement as a feature for detection and classification of the traffic signs. Traffic sign detection and recognition based on the gaze of the driver were explored by Zabihi et al. [19], where a template matching system was used to detect and classify types of signs in images. They then utilized the point of gaze of the driver to determine whether the sign was within the visual field of attention of the driver. The focus was on the methods for detection and recognition of signs and determining the visual field of attention and the work only considered a subset of frames with traffic signs; no analysis
estimating numbers of signs missed was performed. Martin and Tawari [20] introduce a machine vision framework to estimate the object of fixation given the fixation in driving scenarios. Their work focuses on being able to assign a probability to each object in the scene on their likelihood of being the object of the fixation leveraging gaze behavior of the driver. Fixation was determined manually from data collected via wearable eye tracking glasses.

Our work is most similar to the work of Fletcher et al. [17], Chan et al. [18], and Zabihi et al. [19]. Chan et al.’s work focused primarily on traffic sign detection and recognition though used eye movement data as additional features. Fletcher et al. only considered speed limit signs. Zabihi’s work was not only primarily focused on detection and recognition of signs but also determined the visual field of attention and focused on a subset of example frames. While our work uses traffic sign detection and recognition and point-of-gaze computation, the focus is on reporting results on signs missed, types of signs, and duration of gaze. It should be noted that Zahibi is from our lab, and we make use of the point-of-gaze computation method that he used and that was developed in our lab [3, 19].

3. Approach, Data, and Methods

Our overall methodology entails five steps.

1. **Sign Detection.** We analyze images in the sequences to find and identify traffic signs. This requires us to analyze each image to detect regions within the image that may be traffic signs and then determine if those regions are traffic signs.

2. **Sign Classification.** Once a portion of an image has been identified as a sign, we determine the type of sign; this is a classification problem.

3. **Determine the Attentional Field of the Driver.** The attentional visual field of the driver is obtained by analyzing the combination of data from a front-view stereo imaging system and a noncontact 3D gaze tracker; some of the algorithms for this aspect of the work have been carried out in previous research on the RoadLab data [3, 37]. The gaze of the driver on an image, specifically the point of gaze on the image, was determined for all driving frame sequences in our dataset.

4. **Determine if Point of Gaze Is on a Sign.** For each frame, using the point of gaze and the regions corresponding to traffic signs, the intersection of the point of gaze with regions is determined. If there is an intersection with a region, then the traffic sign corresponding to that region is deemed to have been “gazed upon” by the driver. This is recorded, and for each image, in the sequence, we have a corresponding data entry on signs in the image (as detected) and whether gazed upon or not. Essentially, each image is annotated with the regions corresponding to the signs detected in that image and the point of gaze.

5. **Analyze Gaze.** For each driver, we analyze the sequence of data on the signs detected and those that have intersected the driver’s gaze. We determine the number of signs, how long they were visible in the driving sequence, and whether or not the driver actually “saw” the sign (as determined by gaze). This provides information on just what the drivers did and did not see during their drives.

The data from RoadLab, described in Section 3.1, were used in our research. Since we must detect, recognize, and classify traffic signs, we also need datasets containing images of traffic signs. We discuss the traffic sign datasets used to train out models as well.

3.1. Datasets

3.1.1. Road Lab Dataset. RoadLab is an initiative to collect data and develop methods for the development of ADASs. RoadLab data were collected using a vehicle instrumented (see Figure 1) with an on-board diagnostic system which used the CANbus protocol [3] and provided vehicle parameters, such as brake pedal pressure and steering wheel angle. The RoadLab system encompasses different instruments, including stereo cameras, LCD screens, and GPS units. It also has cameras that are used for eye tracking to record the gaze of the driver. Video sequences of the driving environment in front of the vehicle and other optical behavior such as driver gaze as well as GPS data were collected.

Our image sequences come from a study of individual drivers between 20 and 47 years old in the city of London and Ontario, Canada using RoadLab. Each driver used the RoadLab vehicle to drive over the same route. Two other observers were also present in the vehicle to both supervise the performance of the equipment and guide the driver to correctly navigate the route.

The data coming from RoadLab were recorded in real time, under real environmental conditions. Each drive took 50–60 minutes for each driver; the times differ based on the drivers or other road events, such as traffic. Dataset images are recorded with a resolution of 320 by 240 through a front stereo rig mounted on the front roof of the vehicle. For 60 minutes of driving, approximately 100,000 frames were recorded. Figure 2 illustrates some samples of images from our dataset.

3.1.2. Traffic Sign Data. In order to train and test a detector to find traffic signs, we need access to a large number of samples and preferably a dataset with images under different environmental situations, such as weather conditions, different illumination levels, and occlusion and notations. A number of research groups have made datasets available for the community, e.g., [38, 39]. These datasets provide images for both detection and recognition. However, traffic signs are different between Europe and North America. There is no open source dataset for North America and none for Canada, in particular. Our system is based on Canadian traffic signs, so a new dataset for our system was needed.
We extracted images from the driving sequences coming from RoadLab. We cropped the traffic signs and resized them. These images were categorized into 30 different classes to be used to train a multiclass classifier. We also add images from this set through resizing, rotation, and distortion to increase the number of samples in different configurations. Examples of the extracted signs are illustrated in Figure 3.

As will be explained in the following sections, we needed two different datasets, one for detection and one for classification. For the detection part, we trained a linear binary classifier; we used a positive dataset containing 950 samples of signs and a negative dataset of 2500 samples of other objects or the background. For the classification stage, we identified 30 different classes of signs. Our classifier is a multiclass support vector machine. We use 50 samples for each class as our positive dataset and use 500 negative samples as background. The total number of positive samples for the classification stage is 1500 and the total number of negative samples is 500.

3.2. Traffic Sign Detection and Classification. In this section, we describe our methods used for traffic sign detection and recognition. There has been substantial work on both traffic sign detection [24–32] and on classification [33–38]. In processing of the image sequences, we needed our detection method to identify regions, as bounding boxes, within an image that were signs in order to determine if, in an image, the driver’s gaze fell on that region. We also needed to classify the region as a particular type of sign in the Canadian driving context. As a result, we developed our own methods for detection and classification, though we drew heavily on previous work in these areas.
3.2.1. Detection Stage. Maximally stable extremal region (MSER) [39] is a method of blob detection for detecting regions in images based on brightness. This algorithm extracts from an image a number of co-variant regions called MSERs; an MSER is a stable connected component of some gray level of the image. Several different regions can be detected by this approach and it can be applied to all kinds of images, regardless of the texture of image.

The goal of the detection stage is to find the regions that are most likely to include traffic signs. We utilized the MSER algorithm but with some changes. Since we are interested in traffic signs, we are interested in segments of the images as candidate regions which include the colors of the sign (such as green, yellow, red, and orange). We look to extract these regions of color and then create images with different gray scales for analysis by the MSER algorithm. Preprocessing is carried out to enhance the color of the images containing the signs. Other work [40] used similar approach for region of interest extraction for sign detection. That approach extracted the regions containing signs with red and blue colors. Other colors, such as green, yellow, and orange, were not considered. We use a region detector to find candidate regions using several different colors (not only blue and red but also yellow, green, and white).

There are two types of MSER regions: dark connected components on a brighter background (MSER +) and bright ones surrounded by a darker background (MSER−). We used the latter since we were extracting specific colors. In order to find and extract regions containing specific color pixels, colored pixels are converted in a way corresponding to one of the main color channels (red, blue, and green). By extracting these pixels, an image can be converted to a grayscale image of bright and dark pixels based on a specific color and can be processed using the MSER region detector. Bright pixels are considered as MSER regions.

Our detection method consists of three steps:

(1) Preprocessing stages to extract desired pixels of colors
(2) Extract MSER regions and connected components
(3) Determine bounding boxes for each region and then merge regions

(1). Step 1 (preprocessing): in this step, different pixels of colors are extracted in order to identify the MSER regions. Three different functions based on three primary colors are used to acquire the MSER regions and connected components. In each function, the desired MSER regions are found for each specific color. They all use the median filter and contrast normalization to improve the results of later processing. Normalization transforms an n-dimensional grayscale image I: \(X \subseteq \mathbb{R}^n\) \(\rightarrow\) \(\{\text{Min}, \ldots, \text{Max}\}\) into a new image \(I_N: X \subseteq \mathbb{R}^n\) \(\rightarrow\) \(\{\text{newMin}, \ldots, \text{newMax}\}\).

The next step is to find the regions possibly containing the traffic signs. An image is separated into the R, G, and B different channels, and a median filter is applied to individual channels to remove noise. We extract red pixels from the image and convert them to grayscale [40]; the extracted red pixels become white in the gray scale and the other pixels are black. Figure 4 illustrates an example of extracting red pixels. The MSER algorithm is used to process these images and the algorithm then identifies white pixels as regions of interest. We process the image for yellow, green, and white pixels in a similar manner, though to extract yellow pixels we invert our image to turn the yellow pixels blue in the inverted image, and then, we subtract the blue channel of the image from the grayscale image.

(2). Step 2 (extracting MSER regions and connected components): after preprocessing, we applied MSER as a region detector to search and find regions in images for each particular colored pixels. Regions with high levels of contrast in a grayscale scene are detected by MSER. The regions with extracted color pixels have the high level of contrast in a grayscale image. We used the MSER feature detector from the Matlab toolbox. This function has two important inputs: the image to analyze and a threshold delta which is used to compute the intensity of the threshold levels. The input range of this parameter is between 0.8 and 4. We empirically chose a threshold with value 3. We chose this threshold as it can be used for the color orange as well as red and so the red function can cover red and orange color pixels. For the other types of colors, such as green and yellow, we tried different thresholds and finally chose the same as the red threshold. This seemed to generate the best results for all colors of interest. Through this function, we can find the MSER regions and MSER connected components.

We repeat this step for all the colors, and the connected components derived from this step for each color are stored in a matrix. The matrices are merged together in order to be evaluated in region properties measurement stage.

Finally, we used an SVM to determine if certain regions do not correspond to signs. The SVM uses histogram of oriented gradients’ (HOG) feature detector [38]. We trained the SVM using an iterative process on our training images (950 images with signs and 2500 images with no signs) to train the SVM in an iterative process. When the SVM predicts a negative sample as a positive sample (false positive), we add this sample to the training data of the SVM for next iteration as a negative sample. We repeat this process for 10 iterations. The use of the SVM increased the accuracy of identifying likely regions.

(3). Step 3 (finding bounding boxes): After finding the regions, we find the bounding boxes. We store all constructed bounding boxes for each found region in an image. In order to determine whether the bounding boxes encompass the border of the sign, we expanded the values of the edge coordinates of the bounding box. We extract all four coordinates of bounding boxes and expand them according an empirically predetermined small value (0.02); this value was determined empirically; smaller values were too restrictive and much larger values created large boxes which caused challenges in determining overlapping regions. This value is used by the application to expand neighboring bounding boxes and is employed for all the found bounding boxes. Expanding neighboring bounding boxes is a crucial step that is carried out in preparation of the final merge of individual bounding boxes. This expansion is necessary to detect the varied areas of signs.
Some images contain multiple bounding boxes overlapping each other. We need them to be merged together to form a single bounding box around an individual region. This was addressed by setting an overlap ratio between all the bounding box pairs. This quantifies the distance between all pairs of sign regions so that it is possible to find the groups of neighboring sign regions by looking for nonzero overlap ratios. An overlap ratio is a value between 0 and 1 between the pairs of bounding boxes. The best value is 1 indicating the perfect overlap. We used "bboxOverlapRatio" function of MATLAB toolbox to compute the pairwise overlap ratios for all the expanded bounding boxes. The output of this function is a matrix by $M$ rows and $N$ columns. Each $(i, j)$ element in this matrix corresponds to the overlap ratio between the bounding box of row $i$ and the bounding box of row $j$.

The bounding boxes are merged according to this overlap ratio. Even after this merger, there may still be more overlapping bounding boxes. In order to address this issue, we consider one more constraint. The ratio of the height to the width or the width to the height of the bounding boxes is also considered. If the ratio of the height to the width is less than 1.3 or if the ratio of the width to the height is less than 1.2, then we merge the bounding boxes to determine the bounding box over a sign; again, these values were determined empirically.

Although the generated results through the combination of color information and the MSER algorithm were satisfactory, there were still some regions wrongly considered as a sign. To address this, we added one more step to the detection part.

### 3.2.2. Classification Stage

For classification of traffic signs, we again make use of an SVM using HOG features. For training the classifier, we considered 30 different classes containing 50 samples of each class. These samples are extracted from our recorded images. The input images and their labels are input, and each class is processed. Then, the HOG features of images in each class are extracted to be used as features to train the SVM classifier. The function used to extract bounding boxes over detected signs is called to provide the regions believed to be detected road signs. The size of the bounding boxes is resized to $64 \times 64$ to fit the HOG window size. The HOG features are extracted from each bounding box. So, as to not to miss any small-scale detail, the cell size is selected as $5 \times 5$. Examples of different recognized traffic signs are illustrated in Figure 5.

### 3.2.3. Detection and Classification Results

Our detection and classification approach consisted of two steps. In the first step, regions were identified, and an SVM was used to classify the region as a sign or not. This SVM (see Section 3.2.1) was trained on a variety of signs (positive examples) and images of other objects (negative examples). This determined whether a region was likely a sign or not. In the second step, a region identified, as a sign, was then classified into one of the 30 different classes of signs.

We validated the methods against the images for the driving sequence for one driver (driver number 14). The calculation of the number of true positive and false negative was done by hand. For step one, identifying a region as a sign or not, we obtained the accuracy rate 80.98%. The total number of frames considered was 87131, and there were a total of 20505 images of signs in those frames. Our detection algorithm identified 16606 of these as signs (see Table 1).

The result of our classification (recognition) method, step two, is summarized in Table 2. Of the 16606 signs identified in the detection stage, 13568 were correctly classified into one of the 30 classes. This means that 3038 regions that were correctly identified as signs were misclassified resulting in an accuracy of 81.71% of correctly classifying objects identified as signs.

### 3.3. Finding the Attentional Area of the Driver

In the following, we describe how we determined the attentional field of view of the driver. This approach was developed in our laboratory [37], and we summarize it here for completeness.

The 3D point of gaze (PoG) of the driver is obtained by relating the 3D line-of-gaze (LoG) of the driver to the depth map derived from the front camera system. The coordinates of the 3D PoG on each frame of reference is derived by the method proposed in our laboratory in [41]. We consider a plane vertically placed at the 3D point of gaze along the 3D line of gaze and represents the region of interest. When this cone is mapped to the imaging plane of stereoscopic image, it becomes a 2D ellipse. The gaze area has the range of 6.5-degree extension for each pitch and yaw of direction [41]. So, we assume $\theta = 6.5$ degrees; hence, a radius of 13. Given the eye position ($e$) and the 3D point of gaze obtained through the eye tracker system ($g$), the radius of 3D gaze is computed by

$$r = \tan(\theta)d(e, g),$$

(1)

where $e = (e_x, e_y, e_z)$ represents the coordinates of eye position and $g = (g_x, g_y, g_z)$ is the coordinates of points of gaze.

The Euclidian distance between the eye positions and the 3D points of gaze is obtained by

$$d(e, g) = \sqrt{(e_x - g_x)^2 + (e_y - g_y)^2 + (e_z - g_z)^2}.$$  

(2)

Also, a 3D circle can be parametrically defined as

$$S(\varphi) = (X(\varphi), Y(\varphi), Z(\varphi))^T = g + r(\cos(\varphi)u + r(\sin(\varphi)v)).$$

(3)

Here, $u = (u_x, u_y, u_z)$ and $v = (v_x, v_y, v_z)$ are the coordinates of two perpendicular vectors in the plane and $\varphi$ is the angles with different values from 0 to $2\pi$. Both eye tracker and forward stereoscopic systems have their own frame of reference. We need the transformation of the 3D circle, which is equal to the attentional area of the driver, to the aspect of the stereo system of the driver. The computation of parameters related to the transformation was previously computed in our laboratory by a cross-
Show Detected Sign

validation process [37]. This process was applied between the eye tracker system and stereoscopic system [37] and computed in

$$S'(\varphi) = R^T (S(\varphi) - T),$$

(4)

where $R$ and $T$ are the rotation and translation matrices, respectively.

Table 1: Summary of traffic sign detection results.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of regions in images that are signs</td>
<td>20505</td>
</tr>
<tr>
<td>Number of detected signs</td>
<td>16606</td>
</tr>
<tr>
<td>Number of missed signs</td>
<td>3899</td>
</tr>
<tr>
<td>Detection accuracy</td>
<td>80.98%</td>
</tr>
</tbody>
</table>

Table 2: Summary of traffic sign classification results.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of regions detected as signs</td>
<td>16606</td>
</tr>
<tr>
<td>Number of regions correctly classified as signs</td>
<td>13568</td>
</tr>
<tr>
<td>Number of wrongly classified signs</td>
<td>3038</td>
</tr>
<tr>
<td>Accuracy</td>
<td>81.71%</td>
</tr>
</tbody>
</table>

After determining the 3D circle, we need to place it on the stereo imaging plane; this is done by

$$s'(\varphi) = \frac{1}{Z} K S'(\varphi).$$

(5)

Here, $K$ is the intrinsic calibration matrix of the scene stereo system. Finally, we obtain the attentional area of the driver as an ellipse, which is our region of interest.
3.4. Assessment of Driver Attention Based on Gaze. After detecting the regions of interest and classifying the signs, the last stage is to combine the driver’s gaze with the results of our detection system. This is done using the computed attentional area of the driver. In the following, we describe the algorithm used to evaluate the driver attention based on gaze. To do this, we want to determine

(i) Whether a sign that has been detected and recognized is in the attentional area of the driver

(ii) What the type of the sign is

To determine if the driver has seen or missed a sign, we make the assumption that if the center of the bounding box of a sign is inside the attentional area of the driver, then the driver has seen the sign. The attentional area of the driver is a 2D ellipse. By using the equation of the 2D ellipse, we can identify whether the center of the bounding box around the sign is inside the attentional area of the driver or not. If we let $(x_{lb}, y_{lb})$ be the center coordinate of the bounding box, let $(c_x, c_y)$ be the center coordinate of the gaze, let $\theta$ be the gaze angle, and let $a, b$ be the radiuses of the ellipsoid of the attentional area. Then, a sign is seen or not by the driver given by (6) and (7):

$$
\epsilon = \frac{[(x_{lb} - c_x) \times \cos(\theta) + (y_{lb} - c_y) \times \sin(\theta)]^2}{a^2} + \frac{[(x_{lb} - c_x) \times \sin(\theta) - (y_{lb} - c_y) \times \cos(\theta)]^2}{b^2}, \quad (6)
$$

$$
\begin{cases}
\text{if } \epsilon \leq 1, & \text{sign is seen}, \\
\text{else,} & \text{sign is not seen}. 
\end{cases} \quad (7)
$$

The constraint defined in (7) means the sign has been inside the drivers’ attention area obtained from point of gaze. This constraint is applied to all the detected and recognized signs inside the image frames. Examples of the result of this processing are shown in Figure 6, where the driver’s attention area is a yellow circle, the bounding box of the sign found is green, and the label assigned by the classifier.

4. Results of the Analysis of Sign and Gaze Data

The processing described in Section 3 described the methods used to detect a sign, classify it, and determine whether the sign is under the driver’s point of gaze. However, a single sign would appear in a sequence of frames, not just one. So, further processing was necessary in order to determine the number of frames in which a sign appeared.

4.1. Analysis of Sign and Gaze Data. To do this, a region that was identified as a sign was tracked across the sequence of frames. For two regions in consecutive frames, we compute the Euclidian distance between their centers. If the distance between the two centers is less than a given threshold, we treat these two regions as the same sign. For this work, the threshold was set at 3.2 and determined experimentally as working well. In this way, we could determine the first and last frame which contained a region that was deemed to be the same sign.

Once this was determined, we considered a sign to be “seen” if at least one of the regions for that sign fell under the gaze of the driver. This means that a sign was “missed” if none of those images came under the gaze of the driver. This is a somewhat over-constrained definition of missed and “seen,” but it provides an estimate of just how many signs were missed. Note that this means that a driver does not have to gaze at a sign through the entire sequence; they can gaze at a sign several times, e.g., at a distance and then closer to the sign.

With a method to determine which regions represent the same sign across multiple frames, we used our system to analyze the driving sequences acquired from 10 different drivers. The number of drivers we have used is comparable to other studies involving actual drivers, e.g., Vincente et al. [6] used 12 drivers and Tawari et al. [8] used 11 drivers.

Table 3 provides a summary of the types of signs we considered (our types of signs) and the information about signs detected and classified across all drivers:

(i) Frames with signs: the number of frames from the sequences in which a sign was detected; that is, a region was identified as a sign by the detection processing. As noted, the detection algorithm was about 80% accurate.

(ii) Sign images under gaze: the number of regions identified as signs (via the detection stage) that fell within the driver’s field of attention and under the point of gaze of the driver. Note that this is done on a per frame basis.

(iii) Sign images not in gaze: the number of regions identified as signs that did not fall within a driver’s field of attention.

(iv) Signs seen: as per our definition, a region identified as a sign is considered to be “seen” if it falls under the driver’s point of gaze in at least one of the frames deemed to have that particular sign.

(v) Signs missed: this is the number of signs where the gaze of the driver did not fall on any of the regions deemed to be the particular sign. This number plus the number of signs seen constitute the number of signs of a particular type that were identified across all 10 drivers.

(vi) Percent missed: percentage of signs missed of the total was identified.

(vii) Average duration: the average duration (seconds) for the gaze of a driver on individual signs for each type of sign.

(viii) Classification accuracy: accuracy of classifier for each type of sign.

In just looking at the computed data provided in Table 3, we can see that roughly 20% of the traffic signs were not seen by drivers. Again, our assumption is that if a driver’s gaze does not fall onto a sign, as computed by our attention area,
Table 3: Summary of traffic sign classification results.

<table>
<thead>
<tr>
<th>Sign type</th>
<th>Frames with signs</th>
<th>Sign images under gaze</th>
<th>Sign images not in gaze</th>
<th>Signs seen</th>
<th>Signs missed</th>
<th>Percent missed (%)</th>
<th>Average duration (seconds)</th>
<th>Classification accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bicycle</td>
<td>3750</td>
<td>1626</td>
<td>2124</td>
<td>29</td>
<td>10</td>
<td>25.64</td>
<td>1.06</td>
<td>83.23</td>
</tr>
<tr>
<td>Closed lane</td>
<td>11814</td>
<td>7048</td>
<td>4766</td>
<td>45</td>
<td>8</td>
<td>15.09</td>
<td>1.61</td>
<td>84.83</td>
</tr>
<tr>
<td>Construction</td>
<td>19489</td>
<td>12104</td>
<td>7385</td>
<td>58</td>
<td>7</td>
<td>10.77</td>
<td>1.70</td>
<td>90.17</td>
</tr>
<tr>
<td>Exit only</td>
<td>1220</td>
<td>340</td>
<td>880</td>
<td>6</td>
<td>1</td>
<td>14.29</td>
<td>1.26</td>
<td>84.41</td>
</tr>
<tr>
<td>Keep to the right of traffic</td>
<td>61544</td>
<td>24896</td>
<td>36648</td>
<td>225</td>
<td>62</td>
<td>21.60</td>
<td>1.39</td>
<td>93.94</td>
</tr>
<tr>
<td>Lane ahead is closed</td>
<td>7689</td>
<td>4747</td>
<td>2942</td>
<td>43</td>
<td>5</td>
<td>10.42</td>
<td>1.52</td>
<td>83.15</td>
</tr>
<tr>
<td>Lane for two-way left turns</td>
<td>1090</td>
<td>430</td>
<td>660</td>
<td>12</td>
<td>3</td>
<td>20.00</td>
<td>1.10</td>
<td>86.97</td>
</tr>
<tr>
<td>Left not allowed</td>
<td>13559</td>
<td>5384</td>
<td>8175</td>
<td>39</td>
<td>13</td>
<td>25.00</td>
<td>1.60</td>
<td>82.34</td>
</tr>
<tr>
<td>No enter</td>
<td>2129</td>
<td>832</td>
<td>1297</td>
<td>17</td>
<td>10</td>
<td>37.04</td>
<td>1.11</td>
<td>82.04</td>
</tr>
<tr>
<td>No truck</td>
<td>10536</td>
<td>3249</td>
<td>7287</td>
<td>16</td>
<td>2</td>
<td>11.11</td>
<td>1.53</td>
<td>87.50</td>
</tr>
<tr>
<td>One green sign</td>
<td>2712</td>
<td>842</td>
<td>1870</td>
<td>9</td>
<td>1</td>
<td>10.00</td>
<td>1.28</td>
<td>91.78</td>
</tr>
<tr>
<td>Park not allowed</td>
<td>35366</td>
<td>15736</td>
<td>19630</td>
<td>366</td>
<td>119</td>
<td>24.54</td>
<td>0.89</td>
<td>90.85</td>
</tr>
<tr>
<td>Parking</td>
<td>11476</td>
<td>3810</td>
<td>7666</td>
<td>61</td>
<td>25</td>
<td>29.07</td>
<td>1.09</td>
<td>92.65</td>
</tr>
<tr>
<td>Pedestrian crossover</td>
<td>918</td>
<td>554</td>
<td>364</td>
<td>20</td>
<td>0</td>
<td>0.00</td>
<td>0.90</td>
<td>88.34</td>
</tr>
<tr>
<td>Railway crossing ahead</td>
<td>5656</td>
<td>2885</td>
<td>2771</td>
<td>19</td>
<td>5</td>
<td>20.83</td>
<td>1.43</td>
<td>89.14</td>
</tr>
<tr>
<td>Right turn not allowed</td>
<td>10536</td>
<td>2874</td>
<td>7662</td>
<td>31</td>
<td>6</td>
<td>16.22</td>
<td>1.33</td>
<td>89.57</td>
</tr>
<tr>
<td>Road work ahead</td>
<td>4288</td>
<td>2264</td>
<td>2024</td>
<td>26</td>
<td>2</td>
<td>7.14</td>
<td>1.37</td>
<td>91.69</td>
</tr>
<tr>
<td>School crossing</td>
<td>10770</td>
<td>5645</td>
<td>5125</td>
<td>73</td>
<td>13</td>
<td>15.12</td>
<td>1.30</td>
<td>89.03</td>
</tr>
<tr>
<td>Slight bend on the road</td>
<td>465</td>
<td>87</td>
<td>378</td>
<td>2</td>
<td>2</td>
<td>50.00</td>
<td>0.88</td>
<td>84.54</td>
</tr>
<tr>
<td>Speed 40</td>
<td>1080</td>
<td>646</td>
<td>434</td>
<td>10</td>
<td>2</td>
<td>16.67</td>
<td>1.35</td>
<td>81.49</td>
</tr>
<tr>
<td>Speed 50</td>
<td>6190</td>
<td>4350</td>
<td>1840</td>
<td>50</td>
<td>9</td>
<td>15.25</td>
<td>1.34</td>
<td>83.21</td>
</tr>
<tr>
<td>Speed 60</td>
<td>15565</td>
<td>8874</td>
<td>6691</td>
<td>115</td>
<td>18</td>
<td>13.53</td>
<td>1.26</td>
<td>88.16</td>
</tr>
<tr>
<td>Speed 70</td>
<td>2260</td>
<td>1227</td>
<td>1033</td>
<td>21</td>
<td>2</td>
<td>8.70</td>
<td>1.28</td>
<td>92.08</td>
</tr>
<tr>
<td>Speed 80</td>
<td>534</td>
<td>344</td>
<td>190</td>
<td>6</td>
<td>0</td>
<td>0.00</td>
<td>1.43</td>
<td>85.63</td>
</tr>
<tr>
<td>Stop</td>
<td>2233</td>
<td>829</td>
<td>1404</td>
<td>9</td>
<td>7</td>
<td>43.75</td>
<td>1.32</td>
<td>81.86</td>
</tr>
<tr>
<td>Traffic light</td>
<td>8970</td>
<td>5946</td>
<td>3024</td>
<td>67</td>
<td>5</td>
<td>6.94</td>
<td>1.31</td>
<td>86.80</td>
</tr>
<tr>
<td>Traffic in one direction</td>
<td>3252</td>
<td>1189</td>
<td>2063</td>
<td>22</td>
<td>5</td>
<td>18.52</td>
<td>1.02</td>
<td>83.54</td>
</tr>
<tr>
<td>Truck</td>
<td>1913</td>
<td>218</td>
<td>1695</td>
<td>7</td>
<td>1</td>
<td>12.50</td>
<td>0.91</td>
<td>82.67</td>
</tr>
<tr>
<td>U turn not allowed</td>
<td>2957</td>
<td>1114</td>
<td>1843</td>
<td>18</td>
<td>7</td>
<td>28.00</td>
<td>1.28</td>
<td>85.02</td>
</tr>
<tr>
<td>Yield</td>
<td>9933</td>
<td>2732</td>
<td>7201</td>
<td>16</td>
<td>4</td>
<td>20.00</td>
<td>1.40</td>
<td>85.53</td>
</tr>
<tr>
<td>Overall</td>
<td>269894</td>
<td>122822</td>
<td>147072</td>
<td>1438</td>
<td>354</td>
<td>19.75</td>
<td>1.28</td>
<td>86.74</td>
</tr>
</tbody>
</table>

Figure 6: Examples of signs in attention area of driver and not.
then they did not see it. With the exception of parking not allowed and parking, all the signs convey driving information to drivers. If one excludes these, which is reasonable since the drivers were not looking to park, then the percentage of missed signs is about 17%. While signs have been missed, the results do speak positively of the drivers: none of the pedestrian crossover signs were missed, only 20% of the railway crossing ahead signs were missed, and only 15% of the school crossing signs were missed. As well, only 13% of the speed limit signs were missed.

We looked to see if there was any statistical differences among the classes of signs. The Kolmogorov–Smirnov test for normality indicated that the distribution of the percent of signs missed was not significantly different from a normal distribution (K-S statistic: 0.11131; p value: 0.812). We computed the mean percentage missed (18.3%) and the 99% confidence interval (13.0%, 23.5%). Table 4 shows the types of signs outside this interval and whether they were high or low in terms of the percentage of missed signs. Those flagged as “Low” mean that drivers missed a lower percentage of those signs, while those labelled “High” were more frequently missed. As with the cursory analysis of Table 3, those missed most infrequently are from classes of signs particularly important during driving, e.g., lane closed, pedestrian crossing, road work, speed limits, and traffic lights, a positive indication of driver attention to important signs.

We also computed the average duration of eye gaze on signs for each type of sign (see Table 3). As noted, we track a sign across a number of consecutive frames and then can use that to compute the duration of the gaze (frames at 30 fps). The results show that, for most types of signs, the duration is on the order of 1 second.

As noted in Section 1, understanding eye fixation, reaction, and recognition has been an area of study for cognitive scientists and neuroscientists. Their work [1, 2] has suggested that 250 ms is needed for the eye to fixate and react to the sensory input, that is, for the human optical system and brain to react to an object. This process is necessary for recognition of what that object is. The time for recognition would be additional time and is still an area of research for cognitive scientists and neuroscientists. If we take 250 ms as a minimum time for required for recognition and we consider any sign which has been under a driver’s gaze for 8 frames or less (250 ms/30 fps) as missing, then the percentage of missed signs increases to 29%.

4.2. Discussion. From Table 3, one can see that some signs were missed a significant percentage of the time, for example, Stop signs and No Enter signs (43.75% and 37.04%, respectively). This might be of some concern, but there may also be other environmental cues that the driver reacts to without necessarily gazing directly on the sign. For example, some stop signs occur at the end of a side street where the only way to proceed onto the cross street is by taking a left or right turn. In such cases, the driver may just “naturally” stop without gazing at the sign. Also, in our analysis, there were only 16 Stop signs identified. Similarly, roads with No Enter signs may have physical barriers as well as signs and, again, the driver does not explicitly note the sign. Thus, even if some signs were missed, this may not mean that a driver is unaware of the sign’s information in a particular driving situation and that more of the environmental driving context needs to be considered. Another example would be roads approaching railway crossings where there are typically multiple signs along the road. If a driver sees one, it may be sufficient to make them aware of an impending crossing, even if they miss others.

In our evaluation, signs are “missed” only if the driver’s gaze does not fall on any of the images for that sign. A driver may gaze on the same sign multiple times within a subsequence and so may not gaze on the image of the sign in others, the sign is still considered “seen.” If the image sequence is broken, say because of a capture error or the sign being occluded by another vehicle, the sign would be considered a separate sign in our analysis. This is possible, but the analysis is quite complex and requires scene understanding and recognition of occlusion. This is an interesting area for future work.

The above discussion suggests that understanding the broader contextual environment is also important in determining whether missing a sign is critical or not, especially if looking to augment ADASs with this type of capability. Furthermore, in using gaze to help determine driver awareness, gaze duration would seem to be a critical element.

5. Conclusion and Future Work

We analyzed a number of driver naturalistic driving sequences to estimate traffic signs missed by drivers. By computing the intersection of bounding boxes for detected signs and the gaze area information, we were able to determine the point of gaze and attention area. Using this, we were able to determine whether the gaze of a driver fell on a particular sign. We were then able to determine whether the sign had been seen or missed, i.e., no gaze, by the drivers and estimate the duration of the gaze.

4.2. Discussion. From Table 3, one can see that some signs were missed a significant percentage of the time, for example, Stop signs and No Enter signs (43.75% and 37.04%, respectively). This might be of some concern, but there may also be other environmental cues that the driver reacts to without necessarily gazing directly on the sign. For example, some stop signs occur at the end of a side street where the only way to proceed onto the cross street is by taking a left or right turn. In such cases, the driver may just “naturally” stop without gazing at the sign. Also, in our analysis, there were only 16 Stop signs identified. Similarly, roads with No Enter signs may have physical barriers as well as signs and, again, the driver does not explicitly note the sign. Thus, even if some signs were missed, this may not mean that a driver is unaware of the sign’s information in a particular driving situation and that more of the environmental driving context needs to be considered. Another example would be roads approaching railway crossings where there are typically multiple signs along the road. If a driver sees one, it may be sufficient to make them aware of an impending crossing, even if they miss others.

In our evaluation, signs are “missed” only if the driver’s gaze does not fall on any of the images for that sign. A driver may gaze on the same sign multiple times within a subsequence and so may not gaze on the image of the sign in others, the sign is still considered “seen.” If the image sequence is broken, say because of a capture error or the sign being occluded by another vehicle, the sign would be considered a separate sign in our analysis. This is possible, but the analysis is quite complex and requires scene understanding and recognition of occlusion. This is an interesting area for future work.

The above discussion suggests that understanding the broader contextual environment is also important in determining whether missing a sign is critical or not, especially if looking to augment ADASs with this type of capability. Furthermore, in using gaze to help determine driver awareness, gaze duration would seem to be a critical element.

5. Conclusion and Future Work

We analyzed a number of driver naturalistic driving sequences to estimate traffic signs missed by drivers. By computing the intersection of bounding boxes for detected signs and the gaze area information, we were able to determine the point of gaze and attention area. Using this, we were able to determine whether the gaze of a driver fell on a particular sign. We were then able to determine whether the sign had been seen or missed, i.e., no gaze, by the drivers and estimate the duration of the gaze.

| Table 4: Traffic sign categories outside 99% confidence interval. |
|-----------------|-------|
| Bicycle         | High  |
| Construction    | Low   |
| Keep to the right of traffic | High  |
| Lane ahead is closed | Low   |
| Left not allowed | High  |
| No enter        | High  |
| No truck        | Low   |
| Park not allowed | High  |
| Parking         | High  |
| Pedestrian crossover | Low   |
| Road work ahead | Low   |
| Straight bend on the road | High  |
| Speed 60        | Low   |
| Speed 70        | Low   |
| Speed 80        | Low   |
| Stop            | High  |
| Traffic light   | Low   |
| Truck           | High  |
Obvious areas of further work include improving the detection and recognition methods for traffic signs. Our methods were about 80% accurate but still resulted in an estimated 20% missed signs over several tens of thousands of images. Our approach made use of an SVM. Using deep learning methods would likely result in improved accuracy of sign recognition and classification and improving the overall analysis, it would also be more important to gather additional driving sequences and to validate the results reported in this study for more drivers and more varied routes. Work by Schwer et al. [41] describes an approach for the evaluation of the performance of object of fixation detection algorithms. While our focus was not on the method of determining gaze fixation, it would be useful to evaluate the performance along the lines outlined in [41].

Another interesting avenue to consider would be to explore signs missed or not by drivers when traversing a familiar route versus one that is very unfamiliar. Is there a difference in the numbers of signs missed? Similarly, the current experiments and analyses focused on driving in an urban environment where there are many signs, both traffic and nontraffic, but would similar results occur in highway driving? A rural environment?

The results show that different signs are missed at different rates. Another interesting direction for future research would be to consider whether the location, type, shape, and color of signs may be factors in whether drivers locate and identify signs while driving. The approaches presented in this study can be used to examine in detail specific road segments in regards to traffic signs. For example, for intersections that are abnormal in the number of accidents, being able to collect driver gaze and attention to signs may reveal problems with sign placement. Similarly, understanding driver gaze at signage at complex highway intersections may shed light on challenges in negotiating those roadways.

While being attentive to traffic signs is important in driving, signs are not the only environmental cues that inform the driver. For example, the presence of bicycle lanes or pedestrian crossings provide road markings that can alert the driver. The results showed that bicycle and pedestrian signs were missed more often (25% and 40%, respectively) than the average. It may be that the driver had noticed on road cues of bicycles and possible pedestrians and was already alerted. The approach adopted in this work can be used to consider other types of objects that fall within the attentional area of the driver, such as pedestrian crossings, traffic lights, vehicles, and bicycles.

Though the main objective of this study was understanding driver’s attention to traffic signs, our longer term objective was the consideration of such approaches in future ADASs. Determining whether a driver did not see a particular sign image within a limited time period is a possible indication of a potential issue; understanding of the driving situation over a longer time period is needed to determine whether there is a problem or not. This likely means developing a model of the environment and driver behavior over a few previous seconds in which to evaluate the impact of not seeing a sign; i.e., a driving context is needed. Our future work will explore how to build and represent such contexts and use them in evaluating driver behavior.

The sign detection and recognition models can also be constructed offline and subsequently used as part of the ADAS system itself for sign detection and recognition. The gaze can be determined while driving and associated with frames from stereo cameras, this is how data were collected in the RoadLab vehicle originally. Determining the point of gaze can be done quickly, and then, we determined if it has fallen on a traffic sign and if one had been identified. Collecting this information over several image frames, the ADAS could hypothesize whether the driver saw the sign or not. As noted above, additional research can explore determining the broader context in which a sign or signs occurred to really assess whether a driver has missed some key information.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was sponsored jointly by the National Science and Engineering Research Council of Canada (Grant no. 06344-2015). The work presented in this paper is based on the MSc dissertation Assessment of Driver’s Attention to Traffic Signs through Analysis of Gaze and Driving Sequences by Shabnam Shabani and is an extension of that work.

References


Y. Aoyagi and T. Asakura, "A study on traffic sign recognition in scene image using genetic algorithms and neural


[40] S. Salti, A. Petrelli, F. Tombari, N. Fioraio, and L. D. Stefano, “A traffic sign detection pipeline based on interest region extraction,” in Proceedings of the Int. Joint Conf. on Neural Networks (IJCNN), pp. 1–7, Dallas, TX, USA, August 2013.