4-9-2021 10:00 AM

Predictive Model of Driver's Eye Fixation for Maneuver Prediction in the Design of Advanced Driving Assistance Systems

Mohsen Shirpour, *The University of Western Ontario*

Supervisor: Steven Beauchemin, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Computer Science

Abstract

Over the last few years, Advanced Driver Assistance Systems (ADAS) have been shown to significantly reduce the number of vehicle accidents. According to the National Highway Traffic Safety Administration (NHTSA), driver errors contribute to 94% of road collisions. This research aims to develop a predictive model of driver eye fixation by analyzing the driver eye and head information (cephalo-ocular) for maneuver prediction in an Advanced Driving Assistance System (ADAS). Several ADASs have been developed to help drivers to perform driving tasks in complex environments and many studies were conducted on improving automated systems. Some research has relied on the fact that the driver plays a crucial role in most driving scenarios, recognizing the driver's role as the central element in ADASs. The way in which a driver monitors the surrounding environment is at least partially descriptive of the driver's situation awareness. This thesis's primary goal is the quantitative and qualitative analysis of driver behavior to determine the relationship between driver intent and actions. The RoadLab initiative provided an instrumented vehicle equipped with an on-board diagnostic system, an eye-gaze tracker, and a stereo vision system for the extraction of relevant features from the driver, the vehicle, and the environment. Several driver behavioral features are investigated to determine whether there is a relevant relation between the driver's eye fixations and the prediction of driving maneuvers.

Lay Summary

The number of vehicles on our streets and highways increases every day. This fact renders the analysis of traffic situations increasingly complicated. Hence, vehicle manufacturers have been developing Advanced Driver Assistance Systems (ADASs) to avoid 40% of traffic accidents during the driving environment. This research tries to develop a predictive model of driver eye fixation by analyzing the driver eye and head information (cephalo-ocular) for maneuver prediction in an Advanced Driving Assistance System (ADAS). This thesis's primary goal is the quantitative and qualitative analysis of driver behavior to determine the relationship between driver intent and actions. Several driver behavioral features are investigated to determine whether there is a relevant relationship between the driver's eye fixations and the prediction of driving maneuvers.

Acknowledgements

This thesis would not have been possible without support, help and love of many. First and foremost, I am forever grateful to my supervisor, Prof. Steven S. Beauchemin, for his never-ending support, invaluable lessons, and precious guidance throughout the program. I could not have imagined having a better advisor and mentor for my Ph.D study.

Besides my advisor, I would like to thank the rest of my thesis committee: Dr. Krzysztof Czarnecki, Dr. Jagath Samarabandu, Dr. Yalda Mohsenzadeh, and Dr. Anwar Haque for their encouragement, insightful comments, and hard questions.

I would like to thank my fellow labmates in RoadLAB for the stimulating discussions and all the fun we have had in the last few years.

I would also like to express my gratitude to my parents for supporting me spiritually throughout my life.

Last but not least, I must acknowledge my wife, Salimeh, without whose love, encouragement, and patience I would not have finished this thesis.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

According to the World Health Organization (WHO), approximately 1.35 million fatalities and 20 to 50 million injuries occur on the roads every year. Additionally, the WHO predicts that road traffic accidents will rise to become the fifth primary reason for mortality in 2030 [1]. Evidence has demonstrated that a considerable number of accidents are due to driver error. Several Advanced Driver Assistance Systems (ADASs) have been developed to overcome this issue to diminish road fatalities and injuries by minimizing human error.

1.1 Literature Survey

In recent years, various Advanced Driver Assistance Systems (ADAS), such as Automatic Emergency Braking, Forward-Collision Warning, Lane Keep Assist, and Speed Control and Warning, have been designed to assist drivers in performing driving tasks. Improving the reliability and robustness of these systems would certainly have a notable result in reducing the number of collisions and injuries. An ADAS consists of advanced sensors and camera systems

activated when specific conditions arise [2]. Most of these systems are human-centric, as the driver plays an essential role in driving events. Some systems analyze driver behavior in an attempt to predict driver intent in diverse driving situations.

Most drivers have experienced warnings from their passengers to avoid dangerous situations such as accidents with other vehicles or pedestrians. An intelligent ADAS (i-ADAS) works as a co-driver by alerting the driver or even managing the driving task itself. Detecting the driver's eye fixation in relation with the surrounding traffic objects and events could produce meaningful information to efficiently and effectively assist drivers in critical situations [3].

Probably the most significant research area is the determination of driver behavioral features that make ADAS more efficient and effective. The study of identifying objects eliciting visual responses from drivers as it relates to maneuver prediction is known as predictive driver modeling.

This research focuses on driver maneuvers based on a model for driver's eye fixation. The next Section is devoted to driver maneuvers.

1.1.1 Advanced Driver Assistance Systems

ADAS, as the name suggests, are designed to provide safety for drivers in a multitude of driving conditions. These systems assist in minimizing human error, which has been proven to reduce road accidents. ADAS can detect obstacles in the environment by using inputs from several sources such as lidar, radar, and cameras. We provide a summary of ADAS systems in this Section, considering the relationship between the driver's role and these systems, which is classified according to levels of automation. The Society of Automotive Engineers (SAE) has classified driving automation into five levels [4]:

Level 0 Systems

In Level 0 of automation, the driver performs the driving and may be aided by systems that do not monitor the environment or the driving agent itself. An example of this is given by Emergency Braking Systems, that do not technically drive the vehicle but assist the driver in the braking task.

Level 1 Systems

Level 1 ADASs support single functionalities in various driving situations. An example of a Level 1 system is Electronic Stability Control (ESC) that enhances vehicle stability by recognizing and reducing skidding. If the system recognizes a vehicular stability problem, it automatically and temporarily employs braking to stabilize the dynamics of the vehicle.

Level 2 Systems

Level 2 systems can perform various maneuvers, combining longitudinal and lateral dynamic aspects of driving. An example of Level 2 automation is given by Highway Assistance Systems (HAS) that automatically control speed and steering of the vehicle to remain in a particular highway lane.

Level 3 Systems

Systems at this level of automation perform most if not all driving maneuvers by controlling driving actuators, but require driver vigilance. In case of system failure, the driver must be ready to take back vehicular controls. These systems need redundancy in sensors and control units in order to perform driving tasks without driver involvement.

Level 4 Systems

Level 4 systems have the capability of performing some driving tasks without requiring driver vigilance or involvement. An example of a Level 4 system is a Level 3 vehicle equipped with independent valet parking automation. In this scenario, the vehicle is capable of finding a parking area in the absence of the driver.

Level 5 Systems

The final automation step is Level 5, and describes fully automated vehicles. Level 5 vehicles do no require a driver to make decisions and actuate vehicular controls. The driver is considered as a passenger who sets the destination and the vehicle performs the transportation task autonomously.

1.1.2 Driving Maneuver Prediction

Driver maneuver prediction is the primary purpose of driver modeling in ADAS. Authors in [5], [6] and [7] categorize driver maneuvers according to traffic and road infrastructure (See Table 1.1).

To predict driver maneuvers, we need to model the temporal aspects of the driving context and infer driver intent. This task is challenging because driver intent and decisions are not directly measureable, and the interactions between them are poorly understood. Driver behavior is affected by several internal and external factors [7]. These include driving skills, cognitive capabilities, physical features, environmental situations, weather, traffic conditions, and so on. Developing a model to predict driving maneuvers that includes the sum of these factors is complex in practice. Currently the models proposed in the current literature only consider a subset of the above factors.

Table 1.1: Classification of driving maneuvers [5][8]

Reichart	**Tölle**
Follow lane	Start
React to obstacle	Follow
Turn at intersection	Approach vehicle
Cross intersection	Overtake vehicle
Turn into street	Cross intersection
Change lane	Change lane
Turn around	Turn at intersection
Drive backwards	Drive backwards
Choose velocity	Park
Follow vehicle	

Cognitive Driver Model:

These models consider visual perception and attentional features that a driver exhibits while performing maneuvers. From a psychological viewpoint, cognitive driver behavior modeling involves distraction, response time, ability to react, vision, stress, fatigue, and so on [9]. Metari *et al.* [10] examined cephalo-ocular behavior features in different driving scenarios, such as crossing or stopping at an intersection. They showed the cephalo-ocular features play a critical role in effecting maneuvers. Other researchers have studied the influence of human vision on taking actions in specific situations such as driving [11], [12].

The study of driver behavior in a cognitive structure is a valuable source of information to determine the driver's motivation for making an appropriate decision [13]. For instance, when a driver intends to make a left/right turn, the driver's visual behavior indicates the potential intent. Baumann and Krems [14] used driver cognitive structures and presented an operational model of driver situation awareness.

Behaviorist Driver Model:

Such models utilize information in the driver's surrounding environment, including vehicles, pedestrians, and other objects on the road, to find modalities of driver interaction with the surrounding environment. Modern vehicles are equipped with radar [15] for detecting distance, lidar [16] for obstacle detection, visual systems [17] for road object detection, and vehicle navigation systems, such as GPS [18].

The models examined in this literature survey are based on one of these two classes, and it is clear that each group has its own insufficiencies. Obviously, a combination of information from both models can be valuable and practical in understanding driver behavior and predicting the most probable next maneuver.

1.2 Research Overview

Te main objective of this research is to analyse driver eye fixation for maneuver prediction in the context of ADAS. Visual attention and eye fixation plays a crucial role in the research on Driver Safety and Enhanced Driver Awareness (EDA) Systems to inform drivers on incoming traffic conditions, and warn them appropriately.

1.2.1 Primary Conjecture

Cephalo-ocular behavior has shown its usefulness in predicting driver maneuvers [19]. Probably the most efficient approach may be to evaluate and control driver maneuvers to stop or minimize hazardous maneuvers [20]. Based on observation, we believe driver eye fixation and driver visual attention can

be used to build better predictive models of driver behavior in the context of predicting maneuvers [21].

1.2.2 Hypotheses

In this Section, we list and describe the hypotheses on which this research is based. Our objective is to empirically demonstrate their validity.

1. *Driver maneuvers can be partially predicted by Cephalo-ocular behavior features and dynamic vehicle features:* Zabihi *et al.* have demonstrated that driver intent can be predicted by driver behavior features and vehicle dynamics features [22]. They have demonstrated that on their own, neither ocular behavior nor vehicular dynamics are sufficient to predict driver maneuver with adequate accuracy.

2. *Traffic object detection and recognition within the driver's visual attention is possible:* The driver's visual attentional field consists of a circle in 3D space within the plane that contains the Point of Gaze (PoG), perpendicular to the Line of Gaze (LoG). The circle generally projects onto the imaging plane of the stereo sensor as a 2D ellipse. We test this hypothesis by the fact that we can identify objects in the scene. Therefore those objects would fall inside the visual attention area, which has been previously computed by Kowsari et al.[23]. That allows us to detect and recognize which objects within the driver's visual attention area.

3. *The estimation of a confidence interval for the driver's gaze allows for a reliable and robust determination of driver eye fixation:* Driver gaze is not explicitly related to head pose due to the interplay between the head and eye movements. However we believe that head pose is sufficient

for estimation of gaze direction in most situations, as it is devoid from saccadic movements.

4. *Traffic saliency map helps build driver eye fixation model:* It is generally accepted that drivers eye fixations interact with traffic scene objects, which leads to choosing a proper driver maneuver. Therefore, it is crucial to analyze and recognize traffic objects gazed at by drivers, in order to predict intent.

The examination of these hypotheses will increase our knowledge of driver attention and result in the development of predictive driving behavior models for driver maneuver.

RoadLAB Vehicular Instrumentation

To validate the suggested hypotheses, an experimental vehicle equipped with OBD II CANbus channels, a forward stereo vision system, and an eye tracker was provided [3] (see Fig.1.1):

1. The On-Board Diagnostic system (OBD-II) allows sensors to report on current vehicular status in real-time. Several features are extracted from the CANbus, such as vehicle speed, accelerator and brake pedal pressure, steering wheel angle, and turn signals.

2. The stereo vision system located on the vehicle's roof captures the frontal environment at 30Hz.

3. The remote gaze tracker uses a pair of stereo cameras mounted on the dashboard. The remote gaze tracker computes several driver features, including head position and orientation, left and right gaze Euler angles,

Figure 1.1: *RoadLAB Vehicular instrumentation configuration.*

and left and right eye center locations within the tracker's own coordinate system.

Sixteen drivers participated in the data collection experiment, including nine females and seven males. Each participant was recorded by our instrumented vehicle on a pre-determined 28.5km route within the city of London, ON, Canada. Each sequence represented a driving time of approximately one hour. Sequences were recorded in different circumstances, including scenery (downtown, urban, suburban) and traffic conditions, varying from low-traffic to high-traffic situations. They were recorded in various weather conditions (sunny, partially-cloudy, cloudy) and at different day times.

Data Stream

Our driver behavior model includes a Cognitive State of the Driver (head pose, gaze direction, etc.), a Contextual Features Set (road lanes, traffic signs), and a Vehicle Odometry Set, expressed in the form of Real-Time Descriptors (RTDs):

1. the *Cognitive State of the Driver* (CSD), representative of driving ma-

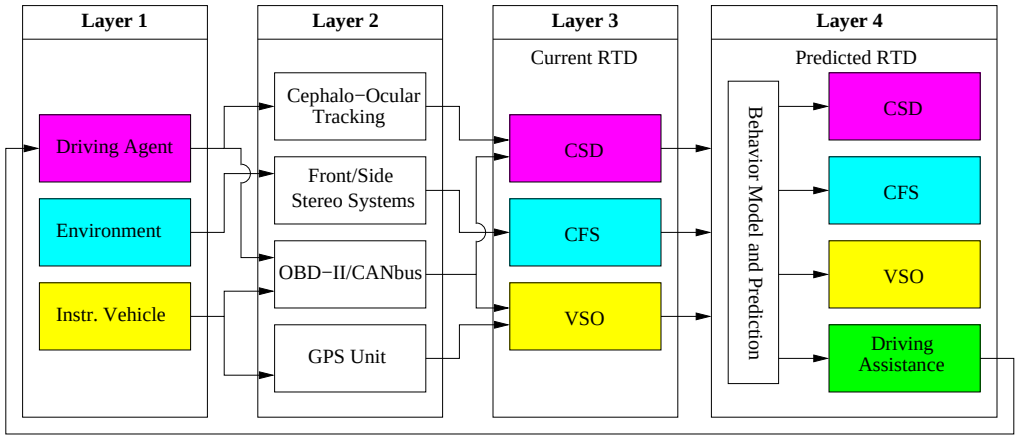


Figure 1.2: *A description of the Driver-Environment-Vehicle parametrization within software layers*

neuvers, usage of vehicle equipment, acknowledged elements within the CFS (by way of intersecting driver 3D gaze direction with elements of the CFS), and level of attention;

2. a *Contextual Feature Set* (CFS), representative of driving environments such as traffic signs, pedestrians, vehicles, lanes, and road boundaries as obtained with on-board sensors, cameras, and vision processes;

3. the *Vehicle State and Odometry* (VSO), representative of dynamic vehicle features such as current speed, acceleration, steering wheel angels, brake pedal pressure, and other information related to the vehicle.

These elements describe the information required in creating an extensive RTD suited for our plans (see Figure 1.2). Both current and predicted RTDs help determine the driver's status. These structures are esential for validating the research hypotheses regarding driver intent and action prediction.

1.3 Contributions

This dissertation is a part of the RoadLAB research program, instigated by Professor Steven Beauchemin, and is entirely concerned with vehicular instrumentation to study driver intent. Chapters 2, 3, 4, 5 and 6 have been published (or in the process of) in recognized peer-reviewed venues. In what follows, I describe my contributions with regards to each publication within the thesis:

1. Chapter 2: N. Khairdoost, M. Shirpour, M.A. Bauer, S.S. Beauchemin, *Real-Time Driver Maneuver Prediction Using LSTM.* IEEE Transactions on Intelligent Vehicles, vol. 5, no. 4, pp. 714-724, Dec. 2020.

    - After initial discussions with Professor Beauchemin about maneuver prediction, N. Khairdoost and I developed a driver behavior model to predict driver maneuvers some seconds before they occur.

2. Chapter 3: M. Shirpour, N. Khairdoost, M.A. Bauer, S.S. Beauchemin, *Traffic Object Detection and Recognition: A Survey and an Approach Based on the Attentional Visual Field of Drivers.* Submitted to IEEE Transactions on Intelligent Vehicles, 2019.

    - N. Khairdoost and I provide a vision-based framework that detects and recognizes traffic objects inside and outside drivers' attentional visual areas.

3. Chapter 4: M. Shirpour, S.S. Beauchemin, M.A. Bauer, *A Probabilistic Model for Visual Driver Gaze Approximation from Head Pose Estimation*, accepted in IEEE 3rd Connected and Automated Vehicles Symposium (CAVS), 2020.

- I presented a new stochastic method for the detection of gaze areas, given driver head pose estimates. Rather than estimating the gaze precisely, which relies on the driver's visual cognitive tasks, the method computes a probabilistic visual attention map describing the probability of finding the actual gaze over the stereo system's imaging plane.

4. Chapter 5: M. Shirpour, S.S. Beauchemin, M.A. Bauer, *Driver's Eye Fixation Prediction by Deep Neural Network,* accepted in 16th International Conference on Computer Vision Theory and Applications (VISAPP 2021) Conference, Vienna Austria, 2021.

    - I proposed convolution neural networks to predict the potential saliency maps in the driving environment and then employed our previous research results to estimate the probability of driver gaze direction as a top-down factor. We statistically combined bottom-up and top-down factors to obtain accurate driver visual fixation predictions.

5. Chapter 6: M. Shirpour, S.S. Beauchemin, M.A. Bauer, *What Does Visual Gaze Attend to During Driving?* submitted to 7th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS 2021) Conference, Prague, Czech Republic, 2020.

    - We analyzed driver gaze behavior and vanishing points with respect to vehicle speed.

1.4 Thesis Organization

The thesis is organized as follows: in Chapter 2, we present a prediction model to anticipate the most likely maneuver a driver will effect a few seconds ahead of time. In Chapter 3, contributions related to traffic objects detected and recognized within the drivers' attentional visual area are presented. In Chapter 4, we propose a probabilistic method for describing the visual attention of drivers. This method applies a Gaussian Process Regression (GPR) technique that estimates the driver gaze direction probability. In Chapter 5, we propose convolution neural networks to predict the potential saliency regions in the driving environment, and then use the probability of the driver gaze direction, given head pose as a top-down factor to predict the driver's eye fixation. In Chapter 6, we analyse the driver's gaze behavior and road vanishing point with the vehicle speed. Finally, in Chapter 7 we suggest conclusions and outlines paths for future research.

Bibliography

[1] Organization WH, et al. Global status report on road safety 2018: Summary. World Health Organization; 2018.

[2] Kim IH, Bong JH, Park J, Park S. Prediction of driver's intention of lane change by augmenting sensor information using machine learning techniques. Sensors. 2017;17(6):1350.

[3] Beauchemin SS, Bauer MA, Kowsari T, Cho J. Portable and Scalable Vision-Based Vehicular Instrumentation for the Analysis of Driver Intentionality. Instrumentation and Measurement, IEEE Transactions on. 2012;61(2):391–401.

[4] Galvani M. History and future of driver assistance. IEEE Instrumentation & Measurement Magazine. 2019;22(1):11–16.

[5] Reichart G. Menschliche zuverlässigkeit beim führen von kraftfahrzeugen. VDI-Verlag; 2001.

[6] Tölle W. Ein Fahrmanöverkonzept für einen maschinellen Kopiloten. PhD thesis, Universität Karlsruhe; 1996.

[7] Bauer C. A driver specific maneuver prediction model based on fuzzy logic. Freie Universität Berlin; 2012.

[8] Okuno A, Fujita K, Kutami A. Visual navigation of an autonomous on-road vehicle: autonomous cruising on highways. In: Vision-based vehicle guidance. Springer; 1992. p. 222–237.

[9] Hamdar S. Driver Behavior Modeling. In: Handbook of Intelligent Vehicles. Springer; 2012. p. 537–558.

[10] Metari S, Prel F, Moszkowicz T, Laurendeau D, Teasdale N, Beauchemin S, et al. A computer vision framework for the analysis and interpretation of the cephalo-ocular behavior of drivers. Machine vision and applications. 2013;24(1):159–173.

[11] Land MF. Eye movements and the control of actions in everyday life. Progress in retinal and eye research. 2006;25(3):296–324.

[12] Lethaus F, Rataj J. Do eye movements reflect driving manoeuvres. IET Intelligent Transport Systems. 2007;1(3):199–204.

[13] Salvucci DD. Modeling driver behavior in a cognitive architecture. Human Factors: The Journal of the Human Factors and Ergonomics Society. 2006;48(2):362–380.

[14] Baumann MRK, Krems JF. A comprehension based cognitive model of situation awareness. In: Digital Human Modeling. Springer; 2009. p. 192–201.

[15] Abou-Jaoude R. ACC radar sensor technology, test requirements, and test solutions. Intelligent Transportation Systems, IEEE Transactions on. 2003;4(3):115–122.

[16] Lu M, Wevers K, Van Der Heijden R. Technical feasibility of advanced driver assistance systems (ADAS) for road traffic safety. Transportation Planning and Technology. 2005;28(3):167–187.

[17] Vlacic L, Parent M, Harashima F. Intelligent vehicle technologies: theory and applications. Butterworth-Heinemann; 2001.

[18] Löwenau JP, Venhovens PJ, Bernasch JH. Advanced vehicle navigation applied in the BMW real time light simulation. Journal of Navigation. 2000;53(01):30–41.

[19] Zabihi SJ, Zabihi SM, Beauchemin SS, Bauer MA. Detection and recognition of traffic signs inside the attentional visual field of drivers. In: 2017 IEEE Intelligent Vehicles Symposium (IV). IEEE; 2017. p. 583–588.

[20] Donmez B, Boyle LNG, Lee JD. Safety implications of providing real-time feedback to distracted drivers. Accident Analysis & Prevention. 2007;39(3):581–590.

[21] Ballard DH, Hayhoe MM, Pelz JB. Memory representations in natural tasks. Journal of Cognitive Neuroscience. 1995;7(1):66–80.

[22] Zabihi SM, Beauchemin SS, Bauer MA. Real-time driving manoeuvre prediction using IO-HMM and driver cephalo-ocular behaviour. In: Intelligent Vehicles Symposium (IV), 2017 IEEE. IEEE; 2017. p. 875–880.

[23] Kowsari T, Beauchemin SS, Bauer MA, Laurendeau D, Teasdale N. Multi-depth cross-calibration of remote eye gaze trackers and stereoscopic scene systems. In: 2014 IEEE Intelligent Vehicles Symposium Proceedings. IEEE; 2014. p. 1245–1250.

Chapter 2

Driver Maneuver Prediction

This Chapter is a reformatted version of the following article:

N. Khairdoost, M. Shirpour, M.A. Bauer, S.S. Beauchemin, *Real-Time Driver Maneuver Prediction Using LSTM. IEEE Transactions on Intelligent Vehicles, vol. 5, no. 4, pp. 714-724, Dec. 2020.*

Driver maneuver prediction is of great importance in designing a modern Advanced Driver Assistance System (ADAS). Such predictions can improve driving safety by alerting the driver to the danger of unsafe or risky traffic situations. In this research, we developed a model to predict driver maneuvers, including left/right lane changes, left/right turns and driving straight forward 3.6 seconds on average before they occur in real-time. For this, we propose a deep learning method based on Long Short-Term Memory (LSTM) which utilizes data on the driver's gaze and head pose as well as vehicle dynamics data. We applied our approach on real data collected during drives in an urban environment with an instrumented vehicle. In comparison with previous *IO-HMM* [1] techniques that predicted three maneuvers including left/right turns and driving straight, our prediction model is able to anticipate two more maneuvers. In addition to this, our experimental results show that our model

using identical datasets improved the F1 score by 4% and increased to 84% accuracy.

2.1 Introduction

The number of vehicles on our streets and highways increases every day. This fact renders the analysis of traffic situations increasingly complicated. For example, in the US alone, at least 33,000 people on average die in road accidents every year, with unsuitable maneuvers being reported as the main cause for most of these accidents [2]. Hence, vehicle manufacturers have been developing Advanced Driver Assistance Systems (ADASs) that able to avoid up to 40% of accidents [3]. Examples of ADASs include adaptive cruise control, collision avoidance systems, traffic warning systems, lane departure warning systems, automatic lane centering, blind spot monitoring, etc. Obviously, improving the reliability and robustness of these systems would have a significant impact on decreasing the number of collisions and accident injuries.

An ADAS consists of advanced sensors and camera systems and is activated when some specific predefined conditions are satisfied. Modeling driving behavior in different traffic scenes, in addition to understanding surrounding environments, makes an ADAS more useful for assisting the driver in controlling the vehicle and avoiding collisions. The goal of this research is to model driver behavior such that ADAS can predict the next driving maneuver a few second before it occurs.

In order to predict driver maneuvers, we need to model the temporal aspects of the driving context and infer the driver's intention. This task is still quite challenging because a driving decisions are not directly detectable and the interactions between them are complex.

In this research, we have developed a model to predict driver maneuvers using Long Short-Term Memory (LSTM) neural networks. LSTMs have the ability to model temporal data and long-term dependencies more accurately than traditional Recurrent Neural Networks (RNNs). Consequently, they are more suitable for predicting driver maneuvers [4]. The model learns the parameters from real driving sequences, including vehicle dynamics, driver head movements, and gaze data. The model infers the potential driving maneuvers (namely, left/right turns, left/right lane changes and driving straight forward) by means of generating a probability for each maneuver. The maneuver with the highest probability is considered as the predicted maneuver.

The rest of this contribution is structured as follows. In Section 2.2, we review the literature. In Section 2.3, we describe our vehicle instrumentation. Section 2.4 contains a description of the proposed method. Section 2.5 presents a summary of the datasets used, learning parameters, and the experimental results obtained along with a critical analysis of those results. We discuss several common reasons resulting in incorrect maneuver prediction in Section 2.6. We give conclusions and future research directions in Section 2.7.

2.2 Literature Survey

In general, to anticipate a driver maneuver, a trained model analyzes contextual driving information. This implies that each driver maneuver is predicted by analyzing data on head movements, GPS, vehicle dynamics, driver gaze, etc.

Artificial Neural Networks (ANNs) have a powerful ability to discover implicitly complicated nonlinear relationships among input variables. Hence, ANNs are suitable techniques for pattern recognition and action prediction

applications, provided that sufficient experimental data is available. For instance, Kim *et al.* [5] applied an ANN to measurements from on-board sensors, such as steering wheel angle, yaw rate and throttle position, to classify road conditions and to predict driver intent for a lane change. Leonhardt and Wanielik [6] employed an ANN for lane change prediction. MacAdam and Johnson [7] represented driver steering behavior in path regulation control tasks using elementary neural networks. Mitrovic [8] used neural networks for short-term prediction of lateral and longitudinal vehicle acceleration.

Although traditional ANNs, such as feed-forward neural nets, are powerful machine learning techniques, ANNs are black box learning techniques. They cannot interpret the relationship among the input and output. Moreover, in the standard probabilistic framework, they cannot work with uncertainties. Another disadvantage is that ANNs consider all input data as independent of each other.

A Bayesian Network (BN) is an acyclic directed graph that represents the conditional dependencies among a set of variables, where the directed edges reflect the qualitative relationships between variables and conditional probability distributions are considered as the quantitative relationships. BNs have been employed for driver maneuver recognition such as overtaking, lane changes or left/right turns [9, 10, 11]. Amata *et al.* [12] presented a prediction model for driver behaviors, such as stopping at intersections based on traffic conditions. Tezuka *et al.* [13] used a BN and steering wheel angle data to develop a model to detect lane keeping, normal lane changes and emergency lane changes. In addition, BNs have been utilized for safety systems to recognize turning maneuvers at intersections and red light crossings [14]. BNs have been used for identifying emergency braking situations [15]. BNs are suitable for applications, such as driver maneuver modeling, where considering uncertainties is

essential. However, considering temporal data using BNs is difficult. Li *et al.* [16] used a novel Dynamic Bayesian Network (DBN) in highway scenarios to predict driver maneuvers. DBNs can model temporal changes, although they cause increased complexity in building and analyzing the network.

Temporal behavior analysis of vehicles surrounding the ADAS vehicle plays an essential role in the safety of the driver. Hence, other methods have been proposed to predict the intent of surrounding vehicles. For example, Kim *et al.* [17] used an LSTM to propose a trajectory prediction technique for analyzing the temporal behavior of surrounding vehicles and their future positions. Alternatively, Khosroshahi *et al.* [18] proposed a framework to classify maneuvers of observed vehicles at four-way intersections using LSTM and 3D trajectory cues. again using LSTM, a method has been introduced by Patel *et al.* [19] to predict lane changes of surrounding vehicles in highway driving. An RNN-based model was presented to interpret the time series data from an observed vehicles at signal-less intersections in order to classify their intentions [20].

Many researchers have utilized Hidden Markov Models (HMMs) for similar purpose. Kuge *et al.* [21] developed steering behavior models for normal/emergency lane changes, and lane keeping using HMMs. Another approach was proposed by Tran *et al.* [22] to predict driver maneuvers, including stop/non-stop, left/right lane changes and left/right turns in both urban and highway driving environments. They employed different input sets to investigate model performance. He *et al.* [23] developed a double-layer HMM structure to model driving behavior and driving intent in the lower and upper layers, respectively. Amsalu and Homaifar [24] employed a Genetic Algorithm (GA) for predicting driver intent when the vehicle approaches an intersection. Aoude *et al.* [25] developed two SVM- and HMM-based approaches

to estimate driver behaviors at road intersections. Their results showed that the SVM-based approach often outperformed the HMM-based model. Jain *et al.* [26] proposed a maneuver prediction model based on an Autoregressive Input-Output Hidden Markov Model (AIO-HMM), which jointly exploits the information inside and outside of the vehicle.

Similarly, Zabihi *et al.* [1] developed a maneuver prediction model using an Input-Output Hidden Markov Model (IO-HMM) that learns relevant parameters from natural driving sequences. They combined vehicle dynamics features and driver cephalo-ocular behavior, including gaze direction and head pose for detecting driver intent. We followed the work of Kowsari *et al.* [27] and Zabihi *et al.* [1] for feature extraction. We refer the reader to these publications for more details.

Researchers also focused on driver maneuver prediction at (urban) intersections. Klingelschmitt et al. [28] created two separate Bayesian Network and Logistic Regression-based models for a vehicle's driving situation and its behavior respectively. Then, they combined them in a single Bayesian Network to design a model able to predict driver intent. An online learning approach using a Bernoulli-Gaussian Mixture Model (BGMM) for feature-based maneuver prediction was presented in [29]. They employed a BGMM to approximate a joint probability density function where predictions are made from a conditional probability distribution function. In [30], an indicator-based approach for driver intent prediction was proposed. They combined context information with vehicular data. The authors in [31] proposed a new approach for intersection maneuver prediction that was based on personalized incremental learning. In other words, they continuously improved the model accuracy by incorporating individual driving history. Liebner et al. [32] proposed an approach to predict driver intent including straight intersection crossing and

Figure 2.1: **a) (left):** *3D infrared gaze tracker;* **b) (center):** *Forward stereoscopic vision system on rooftop;* **c) (right):** *Driver PoG and LoG expressed in the reference frame of stereoscopic vision system and corresponding depth map.*

right turn with the presence or abscence of a preceding vehicle. Their model was based on an explicit parametric model for the longitudinal velocity of preceding vehicles.

Recently, Recurrent Neural Networks (RRNs), Long Short-Term Memories (LSTMs) and Convolutional Neural Networks (CNNs) have been utilized in different applications of ADAS and they have shown promising results for driver activity prediction [33, 34]. Jain *et al.* [33] employed a RNN with LSTM units to preserve long dependencies over the time. They applied their proposed model on a real dataset to predict driver maneuvers. Olabiyi *et al.* [34] proposed a method for anticipating driver action using a deep bidirectional RNN that discovers the relationships between sensor information and future driver maneuvers. They used a fusion of the past and future contexts.

Deep learning has been employed for other ADAS applications, and has brought significant improvements, such as classifying a vehicle's situation for lane changes as safe/unsafe [35] and detecting a driver's confusion level [36].

2.3 Vehicular Instrumentation

We instrumented a research vehicle capable of recording driver-initiated vehicular actuation and relating the 3D driver gaze direction with environ-

Figure 2.2: *Map of predetermined route for drivers, located in London Ontario, Canada. The path length is approximately 28.5 and includes urban and suburban driving areas.*

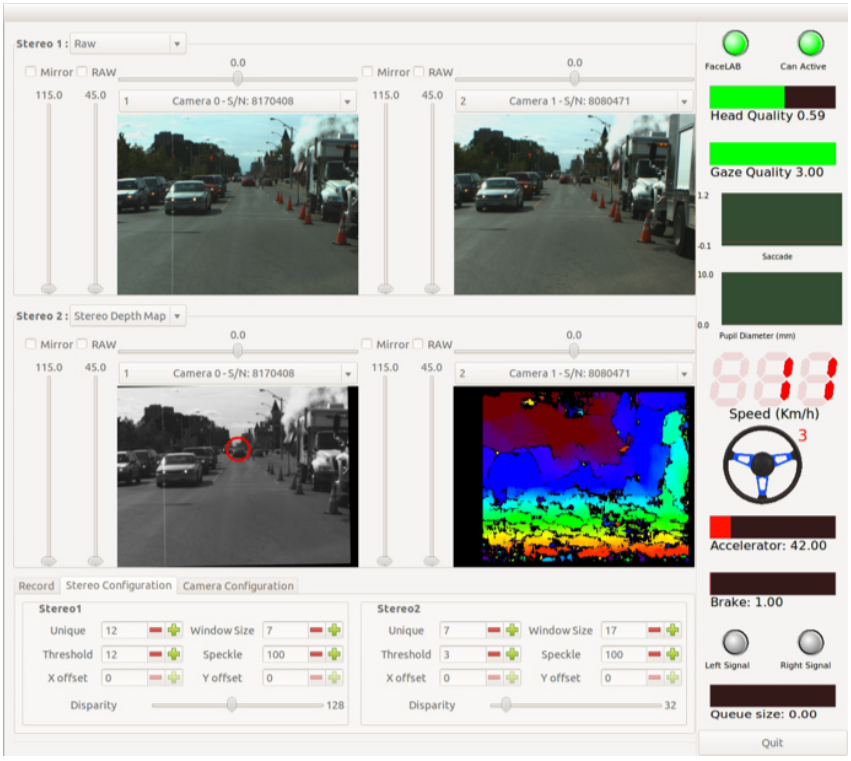


Figure 2.3: *The on-board data recorder interface displaying depth maps, driver PoG, vehicular dynamics, and eye tracker data.*

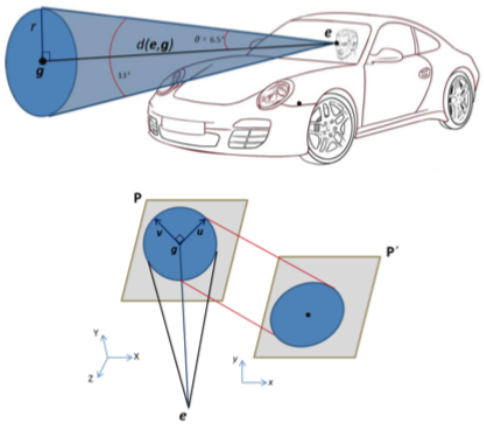


Figure 2.4: *The attentional visual area of driver is defined as the base of the cone located at the depth of sighted features.*

Figure 2.5: *Two projections of the visual attention cone base on the stereo imaging plane.*

mental stereo imagery. The instrumented vehicle was used to collect data sequences with 16 drivers on a pre-determined 28.5km course within the city of London Ontario, Canada. (See Figures 2.1 and 2.2). 3TB of driving sequences were recorded, containing forward stereo imaging and depth, 3D PoG and head pose, and vehicular dynamics obtained with the OBDII CANBus interface (See Figure 2.3). Data frames are collected at a rate of 30Hz.

Our research vehicle is instrumented in such a way as to find whether driver maneuvers could be predicted ahead of time. The vehicle is fitted with a non-contact infra-red 3D gaze and head pose tracker working at 60Hz. Its purpose is to record head movements and gaze direction as they happen while driving. Both head pose and gaze are recorded in the reference frame of the tracker (See Figure 2.1 a) for a depiction of the tracker). A forward stereoscopic vision system is mounted on the roof of the vehicle to provide dense stereo depth maps at 30 Hz. Depth maps are expressed in the frame of reference of the forward stereo system. Details concerning this instrumentation were described by Beauchemin *et al.* [37] and Kowsari *et al.* [27].

We devised a cross-calibration technique to transform the 3D driver gaze and head pose, expressed in the tracker coordinates, in the reference frame

of the forward stereoscopic vision system. As a result, the 3D Point of Gaze (PoG) and Line of Gaze (LoG) of the driver into the surrounding environment are known in absolute 3D coordinates within the frame of reference of the stereo vision system. The attentional visual area of the average driver is defined as the cone from the eye along the LoG (See Figure 2.4). Here, we briefly describe the procedure we used to determine the attentional visual area, whose contour is defined as an ellipse. We first transform the eye position $e = (e_x, e_y, e_z)$ and the 3D PoG $g = (g_x, g_y, g_z)$ into the frame of reference of the forward stereo system, and form a cone with apex e that contains the LoG at its center. This cone has an opening of 6.5° with respect to the LoG [38]. Next, we define a plane perpendicular to the LoG that contains the PoG, and compute the intersection this plane makes with the cone, resulting in a 2D circle located in 3D space. The radius of this circle representing the attentional gaze area is obtained as:

$$r = tan(\theta)d(e, g) \tag{2.1}$$

where

$$d(e, g) = \sqrt{((e_x - g_x)^2 + (e_y - g_y)^2 + (e_z - g_z)^2)} \tag{2.2}$$

The circle is reprojected onto the imaging plane of the forward stereo vision system where it becomes a 2D ellipsoid, as pictured in Figure 2.4. The identification of objects in the scene that elicit an ocular response from the driver can then be identified within this area (Figure 2.5). The cross calibration procedure was devised by Kowsari *et al.* [27]. At the time of its deployment, this was the first publicly known vehicle capable of identifying the 3D PoG of the driver in real-time and in absolute 3D coordinates within the surrounding environment.

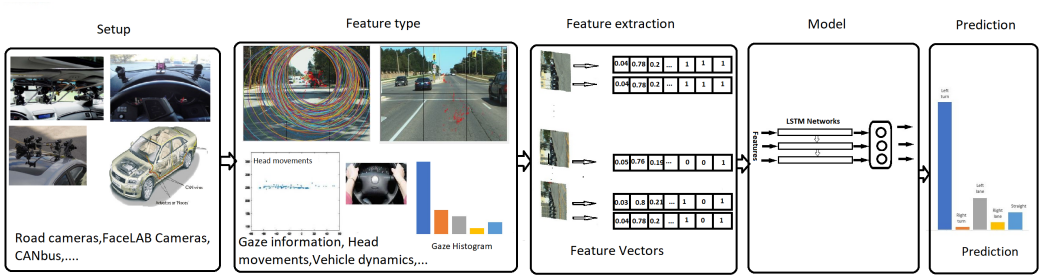


Figure 2.6: *Overview of the proposed approach for predicting driver maneuvers*

2.4 Proposed Method

In order to anticipate driver maneuvers, we need to jointly model the temporal aspects of the driving context and the driver intent. For this purpose, we employed LSTM as it has the ability to model time series data with their long-term dependencies.

In general, the aim of driver maneuver prediction is to anticipate the driver's future maneuvers some time before they occur, given information on driving context. In the model training stage, a set of sequences of observations are fed into the model, where at the end of the sequence, an event happens. In our application, the event can be one of five driver maneuvers: a left/right lane change, a left/right turn, or going forward. The model receives an observation at each time slice so as to predict the driver's future maneuver as early as possible. In other words, the model needs to predict the event by only receiving partial observations from a data sequence. To be exact, each time slice consists of the information of a pre-determined number of frames. Hence, by processing the information available up to current time slice, the observation can be represented as a feature vector (described in Section 2.4.2). We discuss our choice for the size of time slices in Section 2.5.2. Finally, for each time slice, the model outputs the SoftMax probability of each maneuver. Then,

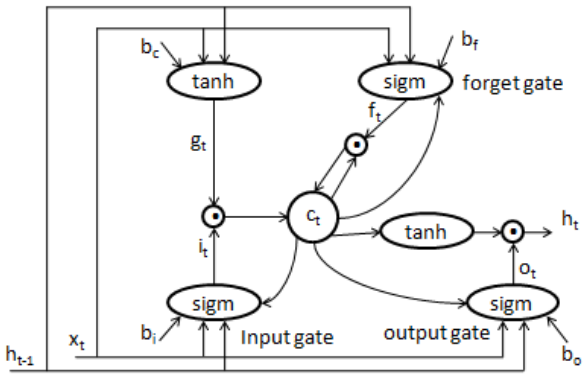


Figure 2.7: *The internal view of an LSTM unit*

the maneuver that has the highest probability is proposed as the predicted maneuver, provided that its probability is higher than a preassigned threshold value, otherwise the system makes no prediction. The choice for this threshold value is justified in Section 2.5.3. Algorithm 1 depicts the complete procedure of our prediction model using LSTM. We refer the reader to Zyner *et al.* [20] and Jain *et al.* [33] for more details on this particular technique. Figure 2.6 provides an overview of our proposed method. Below we present an overview of a standard LSTM unit, illustrated in Figure 2.7.

2.4.1 Long Short-Term Memories (LSTM)

We focus on driver maneuver prediction using LSTMs [39]. LSTM is a particular form of RNNs which is suitable for time series data. Figure 2.7 shows the internal structure of the LSTM unit. An LSTM is able to keep the information of previous input data in its memory, called a *cell*. Hence, it can overcome the vanishing gradient problem in order to remember long-term dependencies. LSTMs have been employed in different ADAS applications [17, 18, 33].

We proceed to describe the equations of an LSTM unit [33, 39]. An LSTM unit has a memory cell and three gates, including an input gate i, a forget gate f, and an output gate o. At each time step, given the observation x_t, the hidden status from the previous time step h_{t-1}, and the previous cell state c_{t-1}, the unit computes i_t and f_t and then updates c_{t-1} to c_t in order to obtain o_t and h_t. Unlike a RNN, the forget gate in the LSTM unit allows the network to throw away part of memory or learn new information. The following recursive equations encode the mechanism:

$$f_t = sigm(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \tag{2.3}$$

$$i_t = sigm(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{2.4}$$

$$g_t = tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{2.5}$$

$$c_t = c_{t-1} \odot f_t + i_t \odot g_t \tag{2.6}$$

$$o_t = sigm(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{2.7}$$

$$h_t = o_t \odot tanh(c_t), \tag{2.8}$$

where $sigm$, $tanh$ and $\odot$ are the sigmoid function, the hyperbolic tangent function, and the element-wise product, respectively. W and b stand for the weight matrix and bias vector. For multi-class applications, we employ a SoftMax layer in which the SoftMax function is applied on a linear transformation of h_t. The following notation describes the internal working of a recurrent LSTM unit concisely. In Section 2.4.2, we describe how we reach an observation x (our features):

$$(c_t, h_t) = LSTM(x_t, c_{t-1}, h_{t-1}). \tag{2.9}$$

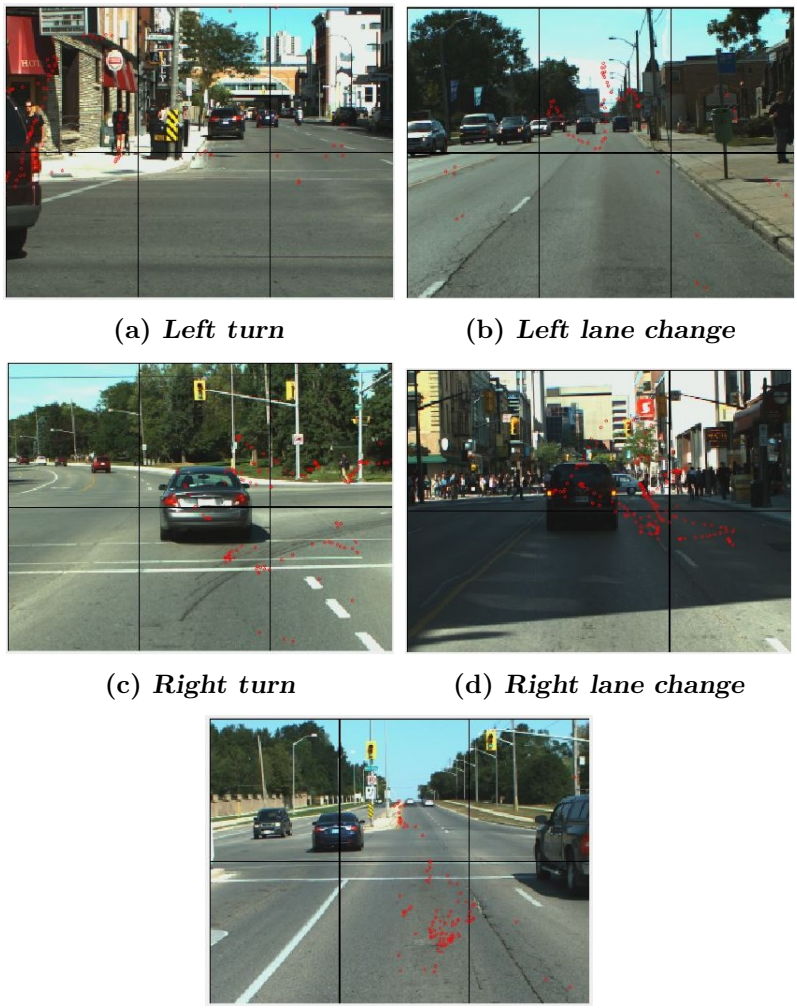
2.4.2 Features for Driver Maneuver Prediction

We proceed with describing the features that are extracted for maneuver prediction. These features are divided into two major categories called driver cephalo-ocular behavioral features and vehicle dynamics features. These features are aggregated and normalized for each time slice (i.e. after receiving 20 consecutive frames in every 0.67 seconds of driving) and their combination constitutes the feature vector, to be fed into the LSTM model. In what follows, we discuss the extracted features for both categories.

Cephalo-Ocular Behavioral Features

It is generally believed that 3D gaze direction plays a significant role in predicting maneuvers since the driver is observing and focusing on the environment moments before performing a maneuver [40],[1]. Hence, two features of the cephalo-ocular behavior of the driver including 3D Point of Gaze (PoG) in absolute coordinates also the horizontal head motion have been utilized to predict driver maneuvers. In order to find the 3D PoG of the driver corresponding to its 3D LoG, we used a cross-calibration method proposed by Kowsari *et al.* [27]. This method combines a binocular eye gaze tracker with a binocular scene stereo system and still remains precise for large distances. Once the cross-calibration step is done, the Line of Gaze (LoG) expressed in the coordinates of the eye-tracker is projected onto the imaging plane of the forward stereo system of the instrumented vehicle. Finally, the 3D PoG is identified as the region obtained by intersecting this projected 3D LoG onto the imaging plane of the stereo system with a valid depth estimate.

To extract 3D PoG features, the frame is separated into six non-overlapping equal parts (as shown in Figure 2.8). We create a histogram of 3D PoGs falling



(a) ***Left turn*** **(b)** ***Left lane change***

(c) ***Right turn*** **(d)** ***Right lane change***

(e) ***Going straight***

Figure 2.8: *Gaze points are shown on the driving frames over the last 5 seconds before a left/right turn, left/right lane change, or going straight maneuver occurs. Frames are divided into six areas.*

into these parts. Figure 2.8 illustrates the PoGs over the last 5 seconds before a maneuver occurs. As can be seen, when drivers are deciding to perform one of the five manuvers, they observe different parts of the frame. For clarification, we discussed the positions of PoGs during a sequence of time slices for a sample of right lane change maneuvers (See Figure 2.9). As shown in Figure 2.9, the driver at first is looking forward, then decides to verify potential obstacles in the right lane before performing the maneuver and then again looks forward. Finally, the driver performs the maneuver while paying attention toward the right lane.

Vehicle Dynamics Features

In 2011, Beauchemin et al. [37] instrumented a vehicle with OBD-II CAN-Bus. As a matter of fact, all vehicles manufactured after 1996 equipped with on-board diagnostic (OBD-II) systems, which allow physical scan devices by means of vehicle sensors to gather and monitor certain vehicle data on the current status via the OBD-II port. Moreover, since 2008, the CANBus protocol (ISO 15765) has been mandatory for OBD-II in all vehicles sold in the US. As a result, this standardization simplifies the examination of real-time vehicle data (which are generally captured with frequencies between 20 and 200 Hz) for researchers and industries to create or improve the performance of intelligent ADAS (i-ADAS) applications.

Vehicle dynamics-based data include vehicle speed, steering wheel angle, left/right turn signals, brake pedal pressure, gas pedal pressure and the speeds of all wheels. We integrated features to benefit from the sum of them simultaneously. For each time slice, we made a histogram of steering wheel angles and encoded the minimum, average and maximum values of vehicle speed, brake

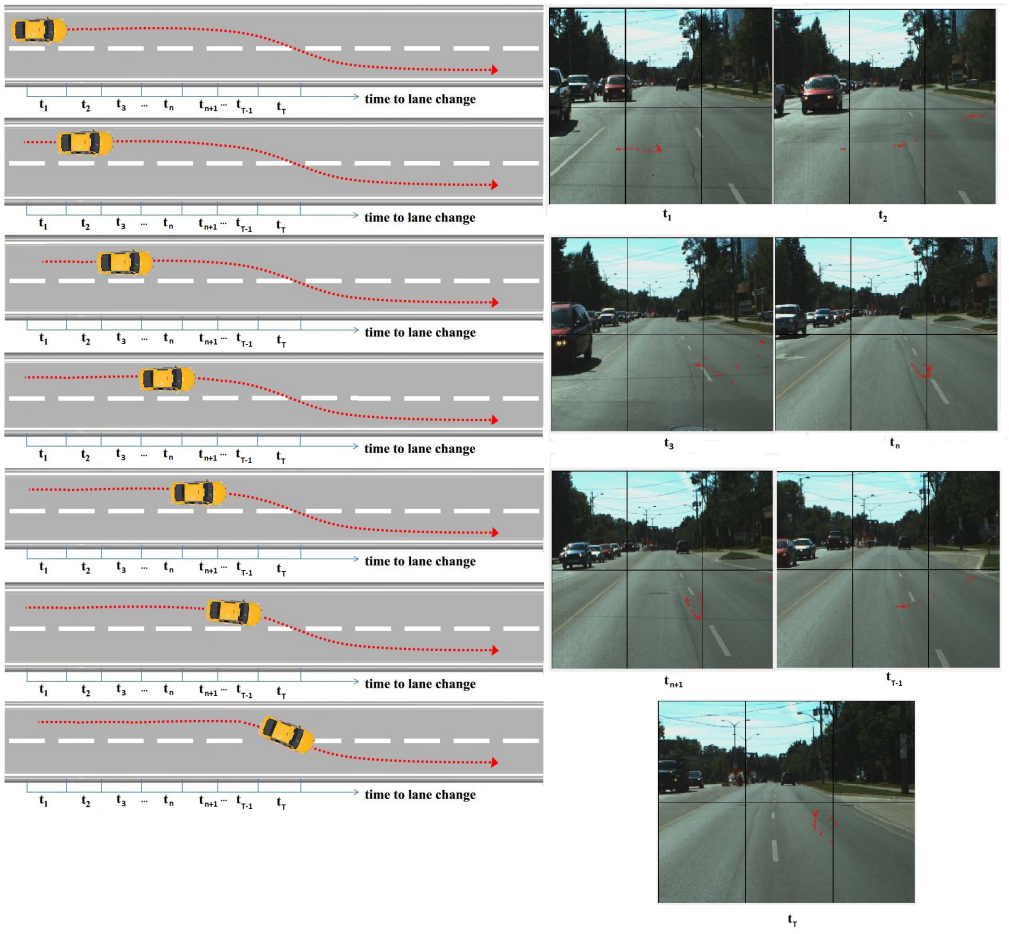


Figure 2.9: *A sequence of time slices belonging to a right lane change event. (t_1): Driver goes straight and looks forward. (t_2 and t_3): Driver decides to initiate an attempt to change lane, and searches visually for potential obstacles in the right lane. (t_n and t_{n+1}): Attention of the driver returns to the current lane and the driver still goes straight. (t_{T-1}): The driver makes the final decision to change lane and looks at the right lane. (t_T): Right lane change event has occurred.*

pedal pressure, gas pedal pressure, indicating independent wheel speeds. Finally, for left and right turn signals, we considered a binary feature for each. This feature value is 1 if the turn signal is on, and 0 otherwise.

Algorithm 1 *Driver Maneuver Prediction Using LSTM*

Input: Cephalo-Ocular Behavior and Vehicle Dynamics Features; Prediction Threshold P_{th}
Output: Predicted Maneuver M; Time-to-Maneuver
  while t = 1 to T **do**
    Observe features available up to current time slice
    Max Probability = Calculate and find the maximum of probabilities of each maneuver using LSTM model
    if Max Probability $> P_{th}$ **then**
      M = Corresponding maneuver with Max Probability
      Time-to-Maneuver = T - t
      break
    end if
  end while
  Return M, Time-to-Maneuver

2.5 Experimental Results

We first give an overview of our maneuver dataset. Then, we explain how we tuned different parameters of the proposed model. Finally, we report our experimental results for maneuver prediction in details.

2.5.1 Dataset

To investigate our proposed model, we applied our approach to driving sequences recorded with the RoadLAB instrumented vehicle in the city of London, Ontario, Canada [37], with the aim of comparing our results with those obtained by Zabihi *et al.* [1], using the same driving sequences as they

did. Table 2.1 provides details on the sequences that have been collected by different drivers for our experiments. These driving sequences contain the data, including GPS, 3D driver gaze, head pose, vehicle speed, and steering wheel angle, among others. We used a total of 325 events obtained from our sequences containing 65 left lane changes, 40 right lane changes, 65 left turns, 75 right turns, and 80 randomly sampled instances of driving straight. Each event is considered as one sample.

Table 2.1: Data Description (each sequence belongs to one driver)

Sequence	Date of Capture	Temperature	Weather
Seq. 8	Sep. 12 2012	27 °C	Sunny
Seq. 9	Sep. 17 2012	24 °C	Partially cloudy
Seq. 10	Sep. 19 2012	8 °C	Sunny
Seq. 11	Sep. 19 2012	12 °C	Sunny
Seq. 13	Sep. 21 2012	19 °C	Partially sunny
Seq. 14	Sep. 24 2012	7 °C	Sunny
Seq. 15	Sep. 24 2012	13 °C	Partially sunny

2.5.2 Learning Parameters

We used a 5-fold cross-validation process to tune network parameters and thresholds on probabilities for driver maneuver prediction by evaluating ranges for the given different parameters. We selected sets of parameters providing us with the highest F1-score on the validation set. Finally, we tested the model on pre-separated, unseen data that consists of a set of randomly selected samples. We performed this strategy several times to estimate the accuracy and generality of the proposed model. In addition, researchers have reported different numbers of frames for the size of time slices such as 10 [31], 15 [29] and 20 [33]. We investigated our performance with time slices of 10, 15, 20, 25 and

30 consecutive frames, and reached better results by employing 20 consecutive frames. Here, we briefly report on other fine-tuned parameters.

Our proposed model consists of 3 hidden LSTM layers. The number of hidden units for the 3 layers was set to 100. We added a dense layer with 5 units for the 5 output classes (including left/right lane changes, left/right turns and driving straight). We employed 0.25, 100 and 10 for the parameters of validation split, epochs and batch size, respectively. The *tanh* activation function for the LSTM layers was used in our experiments. We also used a SoftMax activation function, mean squared error, and Adam method for the dense layer, loss function and stochastic optimization, respectively. Dropout is very important to avoid over-fitting, and we used 0.2, 0.3 and 0.2 for the first, second, and third LSTM layers respectively. The threshold value in our experiments was set to 0.80.

2.5.3 Maneuver Prediction Results

In the test step, the model predicts the driver maneuver every 20 frames and we expect the prediction system to anticipate the maneuver using only partial observations of a sequence. Previously, Zabihi *et al.* [1] proposed an IO-HMM-based model to anticipate three maneuvers of left/right turns and driving straight using our real driving dataset. To compare the performance of our model with theirs, as a first experiment, we employed our approach to predict Zabihi's maneuvers only. In the second experiment, in addition to the aforementioned maneuvers, we utilized our method to predict the maneuvers of left/right lane changes. For each time slice (after receiving 20 frames), the model generates the probability for each maneuver. Then, the maneuver with the highest probability is chosen as the predicted maneuver only if it is higher

than a preset threshold. If the highest probability is less than the threshold (0.8), the system cannot predict the driver maneuver and requires reception of additional features from the next time slice to perform its task. Note that if the maneuver occurs and the system still has not predicted it, the system makes no prediction. We verified the performance of our model by calculating the measures of precision and recall for each maneuver. These measures are defined as follows:

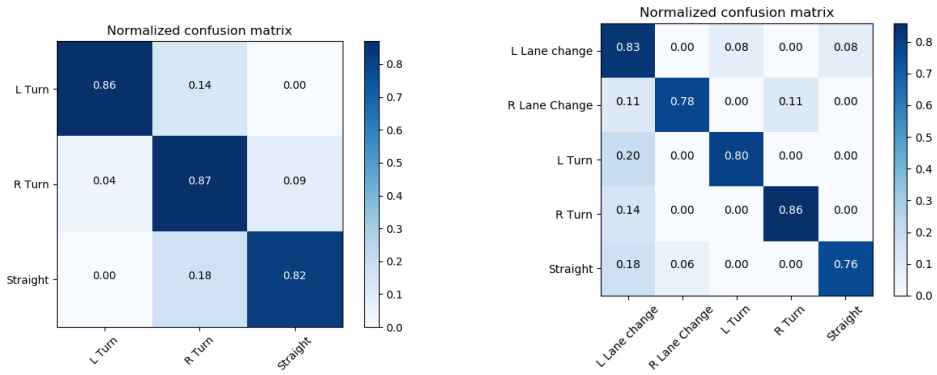
$$Pr = \frac{TP}{TP + FP} \tag{2.10}$$

and

$$Re = \frac{TP}{TP + FN}, \tag{2.11}$$

where, for each maneuver m, TP is the number of correctly predicted instances of maneuver m, FP is the number of incorrectly predicted instances of maneuver m, and FN is the number of instances of maneuver m that are wrongly not predicted or the system is not able to choose any maneuver. Precision is the number of correctly predicted instances of maneuver m divided by the number of instances that were predicted as maneuver m. Recall is the number of instances of correctly predicted maneuver m divided by the total number of instances of maneuver m the average of precision and recall and the average of time-to-maneuver, for true predictions (TP), which indicates the interval between the time of algorithm's prediction and the start of the maneuver. Zabihi *et al.* [1] performed several experiments and reported that utilizing *IO-HMM* with the data on the driver's gaze and head pose *(IO-HMM G+H)* made the better model in terms of precision, recall and Time-to-Maneuver.

Table 2.2 compares our results (considering three and five maneuvers) with Zabihi *et al.* [1]. As can be seen, our LSTM-based model outperformed their prediction model. To be exact, precision and recall of our model for the three



(a) *Model with three maneuvers* **(b) *Model with five maneuvers***

Figure 2.10: *Confusion matrices of our prediction model*

maneuvers are 6.1% and 0.8% higher for these three maneuvers. However, their method can predict the three maneuvers 0.16s earlier on average than ours. The last row in Table 2.2 shows the results of extending our model with two more maneuvers. In this case, we expect increased complexity for the problem and results show that precision, recall and time-to-maneuver have decreased slightly in comparison with our method for predicting only three maneuvers.

Figure 2.10 shows the confusion matrices for our prediction system for three and five maneuvers. In these matrices, a row represents an instance of the actual maneuver class, whereas a column represents an instance of the predicted maneuver class. The values of the diagonal elements represent the degree of correctly predicted classes.

Figure 2.11 compares the changes of the F1-score when we employ our model and the IO-HMM-based model, with different values for the threshold. The F1-score is the harmonic mean of Pr and Re, where it can reach 1 with perfect precision and recall, and 0 in the worst case. The prediction threshold

Table 2.2: Result of different models of driver maneuver prediction on our data set.

	Pr (%)	**Re (%)**	**Time-to-maneuver(s)**
IO-HMM G+H (for three maneuvers)	79.5	83.3	3.8
Our model (for three maneuvers)	85.6	84.1	3.64
Our model (for five maneuvers)	84.2	82.9	3.56

is a useful parameter to find a trade-off between the precision and recall. The F1-score is defined as follows:

$$F1 = \frac{2PrRe}{Pr + Re} \tag{2.12}$$

The trend of F1-scores for the *IO-HMM* model remains roughly stable when the threshold changes. However, when we choose 0.8, the LSTM-based prediction model achieves a significantly higher F1-score in comparison with *IO-HMM* model. In Table 2.2, we used the threshold values which gave us the highest F1-score. Our model predicts maneuvers every 0.67 seconds (20 frames) in 2.8 milliseconds on average, on a 3.40 GHz Core $i7-6700$ CPU with Windows 10.

Finally, we briefly mention here the results of several previous works that have addressed the driver maneuver prediction problem, using their own dataset and features. For instance, Morris *et al.* [40] accomplished a binary classification of lane changes and driving straight maneuvers. They employed a Relevance Vector Machine (RVM; a Bayesian extension to the popular SVM). In addition, Jain *et al.* [33] evaluated some algorithms for the same purpose (including SVM, Bayesian Network and variants of their deep learning model).

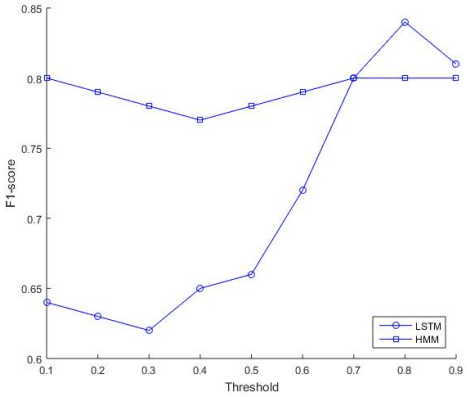


Figure 2.11: *The effect of the threshold on the F1-score for IO-HMM and LSTM models.*

The methods listed in Table 2.3 use identical feature vectors, which guarantees a fair comparison[1]. As can be observed, the SVM classification does not model the temporal aspect of the data, and its performance is poor as a result.

Table 2.3: MANEUVER ANTICIPATION RESULTS OF SEVERAL PREVIOUS METHODS.

Method	**Pr (%)**	**Re (%)**	**Time-to-maneuver(s)**
SVM[40]	43.7±2.4	37.7±1.8	1.20
IO-HMM[26]	74.2±1.7	71.2±1.6	3.83
AIO-HMM[26]	77.4±2.3	71.2±1.3	3.53
S-RNN[33]	78.0±1.5	71.1±1.0	3.15
F-RNN-UL[33]	82.2±1.0	75.9±1.5	3.75
F-RNN-EL[33]	84.5±1.0	77.1±1.3	3.58

[1]The methods listed in the Table are: SVM: Support Vector Machine, IO-HMM: Input-Output Hidden Markov Model, AIO-HMM: Auto-Regressive Input Output Hidden Markov Model, S-RNN: Simple Recurrent Neural Network, F-RNN-UL: Fusion-Recurrent Neural Network Uniform Loss, F-RNN-EL: Fusion-Recurrent Neural Network Exponential Loss.

2.6 Common Reasons for Wrong Maneuver Anticipations

We discuss some major reasons that can generally result in wrong anticipations in the driver maneuver prediction problem. For example, when a driver is interacting with other passengers, head and gaze features are not reliable enough to be taken into account. Also, a driver may be distracted when watching videos, programming a GPS, using a cell phone, adjusting the radio, smoking and etc. In such situations, wrong anticipation is common as the driver may not be fully focused on the road. Moreover, different drivers have different driving styles. For example, during lane change maneuver, some drivers may merge slowly, while others may merge quickly: in this case, the driver has not provided the system with enough data and time to predict the maneuver. Hence, in this situation, other features such as speed, acceleration, steering wheel angle can be significant to predict an accurate maneuver. As another example, when drivers rely on their recent perception of the traffic scene, they probably do not check blind spots and the surroundings carefully, resulting in a lack of head information. A similar driving situation arises when a driver is driving in left/right-turn-only lanes. In this case, the driver might not display helpful head information.

2.7 Conclusion and Future Work

We presented an LSTM-based model to predict driver maneuvers several seconds before they are performed. We employed driver cephalo-ocular behavioral information and vehicle dynamics data as features to train our model. Our experimental results show that our model outperformed the previous IO-

HMM model [1]. It improved the precision from 79.5% to 85.6% and recall from 83.3% to 84.1%. Moreover, we expanded the prediction model to anticipate two more maneuvers (left/right lane changes). For predicting the five maneuvers, our model achieved 84.2% and 82.9% for precision and recall respectively. Our results show that driver maneuvers can be predicted. Several limitations exist for improving the accuracy and generality of the model. We suppose that adding more features from the environment such as the lane in which the driver is locate and where the driver is gazing during the driving maneuver could improve the accuracy of the model. In terms of generality, the tests have been conducted in this research on limited number of drivers and under specific weather and environmental conditions. Collecting new datasets under different situations could help the generality of the model. Hence, for the commercial use of this model, the mentioned items need to be considered. Lastly, this research area is still challenging and more research and efforts must be performed by researchers to be practical in commercial uses. As for future work, we plan to study the extraction of features from video within the attentional visual area of the driver. We believe that utilizing LSTM trained with a combination of these features, with cephalo-ocular behavior and the vehicle dynamics will improve current prediction results.

Bibliography

[1] Zabihi S, Beauchemin S, Bauer M. Real-time driving manoeuvre prediction using IO-HMM and driver cephalo-ocular behaviour. In: Intelligent Vehicles Symposium (IV), 2017 IEEE. IEEE; 2017. p. 875–880.

[2] motor vehicle crashes:. N. Highway Traffic Safety Administration, Wash-

ington, D.C. Tech Rep. 2013;.

[3] Gietelink O, Ploeg J, De Schutter B, Verhaegen M. Development of advanced driver assistance systems with vehicle hardware-in-the-loop simulations. Vehicle System Dynamics. 2006;44(7):569–590.

[4] Azzouni A, Pujolle G. A long short-term memory recurrent neural network framework for network traffic matrix prediction. arXiv preprint arXiv:170505690. 2017;.

[5] Kim IH, Bong JH, Park J, Park S. Prediction of driver's intention of lane change by augmenting sensor information using machine learning techniques. Sensors. 2017;17(6):1350.

[6] Leonhardt V, Wanielik G. Neural network for lane change prediction assessing driving situation, driver behavior and vehicle movement. In: Intelligent Transportation Systems (ITSC), 2017 IEEE 20th International Conference on. IEEE; 2017. p. 1–6.

[7] MacAdam CC, Johnson GE. Application of elementary neural networks and preview sensors for representing driver steering control behaviour. Vehicle System Dynamics. 1996;25(1):3–30.

[8] Mitrovic D. Machine learning for car navigation. In: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer; 2001. p. 670–675.

[9] Huang T, Koller D, Malik J, Ogasawara G, Rao B, Russell SJ, et al. Automatic symbolic traffic scene analysis using belief networks. In: AAAI. vol. 94; 1994. p. 966–972.

[10] Kasper D, Weidl G, Dang T, Breuel G, Tamke A, Wedel A, et al. Object-oriented Bayesian networks for detection of lane change maneuvers. IEEE Intelligent Transportation Systems Magazine. 2012;4(3):19–31.

[11] Meyer-Delius D, Plagemann C, Von Wichert G, Feiten W, Lawitzky G, Burgard W. A probabilistic relational model for characterizing situations in dynamic multi-agent systems. In: Data analysis, machine learning and applications. Springer; 2008. p. 269–276.

[12] Amata H, Miyajima C, Nishino T, Kitaoka N, Takeda K. Prediction model of driving behavior based on traffic conditions and driver types. In: Intelligent Transportation Systems, 2009. ITSC'09. 12th International IEEE Conference on. IEEE; 2009. p. 1–6.

[13] Tezuka S, Soma H, Tanifuji K. A study of driver behavior inference model at time of lane change using Bayesian networks. In: Industrial Technology, 2006. ICIT 2006. IEEE International Conference on. IEEE; 2006. p. 2308–2313.

[14] Zhang J, Roessler B. Situation analysis and adaptive risk assessment for intersection safety systems in advanced assisted driving. In: Autonome Mobile Systeme 2009. Springer; 2009. p. 249–258.

[15] Schneider J, Wilde A, Naab K. Probabilistic approach for modeling and identifying driving situations. In: Intelligent Vehicles Symposium, 2008 IEEE. IEEE; 2008. p. 343–348.

[16] Li J, Li X, Jiang B, Zhu Q. A maneuver-prediction method based on dynamic bayesian network in highway scenarios. In: 2018 Chinese Control And Decision Conference (CCDC). IEEE; 2018. .

[17] Kim B, Kang CM, Lee SH, Chae H, Kim J, Chung CC, et al. Probabilistic vehicle trajectory prediction over occupancy grid map via recurrent neural network. arXiv preprint arXiv:170407049. 2017;.

[18] Khosroshahi A, Ohn-Bar E, Trivedi MM. Surround vehicles trajectory analysis with recurrent neural networks. In: Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on. IEEE; 2016. p. 2267–2272.

[19] Patel S, Griffin B, Kusano K, Corso JJ. Predicting Future Lane Changes of Other Highway Vehicles using RNN-based Deep Models. arXiv preprint arXiv:180104340. 2018;.

[20] Zyner A, Worrall S, Nebot E. A Recurrent Neural Network Solution for Predicting Driver Intention at Unsignalized Intersections. IEEE Robotics and Automation Letters. 2018;3(3):1759–1764.

[21] Kuge N, Yamamura T, Shimoyama O, Liu A. A driver behavior recognition method based on a driver model framework. SAE Technical Paper; 2000.

[22] Tran D, Sheng W, Liu L, Liu M. A Hidden Markov Model based driver intention prediction system. In: Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), 2015 IEEE International Conference on. IEEE; 2015. p. 115–120.

[23] He L, Zong Cf, Wang C. Driving intention recognition and behaviour prediction based on a double-layer hidden Markov model. Journal of Zhejiang University SCIENCE C. 2012;13(3):208–217.

[24] Amsalu SB, Homaifar A. Driver behavior modeling near intersections using Hidden Markov Model based on genetic algorithm. In: Intelligent Transportation Engineering (ICITE), IEEE International Conference on. IEEE; 2016. p. 193–200.

[25] Aoude GS, Desaraju VR, Stephens LH, How JP. Behavior classification algorithms at intersections and validation using naturalistic data. In: Intelligent Vehicles Symposium (IV), 2011 IEEE. IEEE; 2011. p. 601–606.

[26] Jain A, Koppula HS, Raghavan B, Soh S, Saxena A. Car that knows before you do: Anticipating maneuvers via learning temporal driving models. In: Proceedings of the IEEE International Conference on Computer Vision; 2015. p. 3182–3190.

[27] Kowsari T, Beauchemin SS, Bauer MA, Laurendeau D, Teasdale N. Multi-depth cross-calibration of remote eye gaze trackers and stereoscopic scene systems. In: Intelligent Vehicles Symposium Proceedings, 2014 IEEE. IEEE; 2014. p. 1245–1250.

[28] Klingelschmitt S, Platho M, Groß HM, Willert V, Eggert J. Combining behavior and situation information for reliably estimating multiple intentions. In: 2014 IEEE Intelligent Vehicles Symposium Proceedings. IEEE; 2014. p. 388–393.

[29] Wiest J, Karg M, Kunz F, Reuter S, Kreßel U, Dietmayer K. A probabilistic maneuver prediction framework for self-learning vehicles with application to intersections. In: 2015 IEEE Intelligent Vehicles Symposium (IV). IEEE; 2015. p. 349–355.

[30] Rodemerk C, Winner H, Kastner R. Predicting the driver's turn intentions at urban intersections using context-based indicators. In: 2015 IEEE Intelligent Vehicles Symposium (IV). IEEE; 2015. p. 964–969.

[31] Losing V, Hammer B, Wersing H. Personalized maneuver prediction at intersections. In: 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC). IEEE; 2017. p. 1–6.

[32] Liebner M, Klanner F, Baumann M, Ruhhammer C, Stiller C. Velocity-based driver intent inference at urban intersections in the presence of preceding vehicles. IEEE Intelligent Transportation Systems Magazine. 2013;5(2):10–21.

[33] Jain A, Singh A, Koppula HS, Soh S, Saxena A. Recurrent neural networks for driver activity anticipation via sensory-fusion architecture. In: Robotics and Automation (ICRA), 2016 IEEE International Conference on. IEEE; 2016. p. 3118–3125.

[34] Olabiyi O, Martinson E, Chintalapudi V, Guo R. Driver Action Prediction Using Deep (Bidirectional) Recurrent Neural Network. arXiv preprint arXiv:170602257. 2017;.

[35] Scheel O, Schwarz L, Navab N, Tombari F. Situation Assessment for Planning Lane Changes: Combining Recurrent Models and Prediction. arXiv preprint arXiv:180506776. 2018;.

[36] Hori C, Watanabe S, Hori T, Harsham BA, Hershey J, Koji Y, et al. Driver confusion status detection using recurrent neural networks. In: Multimedia and Expo (ICME), 2016 IEEE International Conference on. IEEE; 2016. p. 1–6.

[37] Beauchemin SS, Bauer MA, Kowsari T, Cho J. Portable and Scalable Vision-Based Vehicular Instrumentation for the Analysis of Driver Intentionality. IEEE Transactions on Instrumentation and Measurement. 2012;61(2):391–401.

[38] Takagi K, Kawanaka H, Bhuiyan MS, Oguri K. Estimation of a three-dimensional gaze point and the gaze target from the road images. In: 2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC). IEEE; 2011. p. 526–531.

[39] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997;9(8):1735–1780.

[40] Morris B, Doshi A, Trivedi M. Lane change intent prediction for driver assistance: On-road design and evaluation. In: Intelligent Vehicles Symposium (IV), 2011 IEEE. IEEE; 2011. p. 895–901.

[41] Ponziani R. Turn signal usage rate results: A comprehensive field study of 12,000 observed turning vehicles. SAE Technical Paper; 2012.

Chapter 3

Traffic Object Detection and Recognition

This Chapter is a reformatted version of the following article:

M. Shirpour, N. Khairdoost, M.A. Bauer, S.S. Beauchemin, *Traffic Object Detection and Recognition: A Survey and an Approach Based on the Attentional Visual Field of Drivers. Submitted to IEEE Transactions on Intelligent Vehicles, 2019.*

Traffic object detection and recognition systems play an essential role in Advanced Driver Assistance Systems (ADAS) and Autonomous Vehicles (AV). In this research, we focus on four important classes of traffic objects: traffic signs, road vehicles, pedestrians, and traffic lights. We first review the major traditional machine learning and deep learning methods that have been used in the literature to detect and recognize these objects. We provide a vision-based framework that detects and recognizes traffic objects inside and outside the attentional visual area of drivers. This approach uses the driver 3D absolute coordinates of the gaze point attained through the combined, cross-calibrated use of a front-view stereo imaging system and a non-contact 3D gaze tracker. A combination of multi-scale HOG-SVM and Faster R-CNN-based models are utilized in the detection stage. The recognition stage is performed with a

ResNet-101 network to verify sets of generated hypotheses. We applied our approach on real data collected during drives in an urban environment with the RoadLAB instrumented vehicle. Our framework achieved 91% of correct object detections and provided promising results in the object recognition stage.

3.1 Introduction

Advanced Driver Assistance Systems (ADAS) have attracted the attention of many researchers and vehicle manufacturers for several decades. Achieving higher performance levels for ADAS also imposes strict requirements for robust perception of the driving environment. Hence, vision-based traffic scene perception which refers to the identification of the position of traffic objects such as pedestrians, vehicles, traffic signs, etc is of great importance in designing a modern ADAS. However, in practice, many traffic scene issues, such as occlusions, weather conditions, shadows and distant object identification affect the performance of such systems. Improving the accuracy and adaptability of such methods is still a challenging area of research [1]. In this study, we focus on four essential categories of objects: traffic signs, vehicles, traffic lights, and pedestrians. Problems encountered include variations in viewpoints, object shape, size, color, distance from sensors, illumination conditions, and object occlusion [2] [3] [4].

In this contribution, we propose the first traffic object detection and recognition framework that performs its tasks within the attentional visual field of the driver. This is an important aspect of ADAS, as it allows to identify objects possibly seen by the driver, among other things.

This contribution is organized as follows: In Section 3.2, we review the

related literature. Section 3.3 describes the datasets we used and the proposed method. Section 3.4 presents the experimental results obtained along with a critical analysis. Conclusions and future research directions are described in Section 3.5.

3.2 Related Works

3.2.1 Generic Object Detection

Generic object detection is a challenging task in computer vision that attempts to locate and identify existing objects in one image in order to label them and estimate their extent with bounding boxes. Generic object detection algorithms can be divided into two major types of traditional and deep learning-based methods. In this Section, we briefly review these generic object detection methods. Several object detection surveys can be found in [5], [6], [7], [8], [9] and [10].

Among the traditional object detectors we find the framework proposed by Viola and Jones which employs searches based on sliding-windows and AdaBoost classifiers [11]. Another popularly used framework is the linear Support Vector Machine (SVM) classifier with such features as Histograms of Oriented Gradients (HOG), Scale Invariant Feature Transforms (SIFT), and Local Binary Patterns (LBP). For example, in [12] and [13], researchers employed SVM and a multi-scale detection framework with HOG features to detect birds and pedestrians respectively. In addition, Aggregated Channel Features (ACF) can be mentioned as another successful detection framework that has been proposed by [14]. This method also uses sliding-window searches and AdaBoost to detect objects in a multi-scale fashion [15], [16].

Unlike traditional object detection algorithms that benefit from prior knowledge, deep learning-based object detection methods attempt to learn high-level features from massive amounts of data. As a result, they are less sensitive to illumination changes, deformations, and geometric transformations [17]. There are two major types of deep learning-based object detection methods: region-based methods and regression-based methods. The former generates region proposals at first and then classifies them into different object categories, while the latter transforms the object detection problem into a regression problem and predicts locations and class probabilities directly from the whole image [5]. The region-based methods mainly include R-CNN [10], Fast R-CNN [18], Faster R-CNN [19], R-FCN [20], SPP-net [21] and Mask R-CNN [22]. The regression-based methods mainly include AttentionNet [22], G-CNN [23], SSD [24], YOLO [25], YOLOv2 [26], YOLOv3 [27], DSOD [28] and DSSD [29].

3.2.2 Traffic Sign Detection and Recognition

Sign detection methods can be generally categorized into color-based, shape-based and hybrid approaches [30], [31]. Color-based methods use color information as the main attribute to localize image regions containing traffic signs in the image. Color thresholding segmentation is the more common approach among color-based methods as it reduces the search area by ignoring untargeted regions [32], [33]. These methods are generally sensitive to variations in illumination and the distance to traffic signs [34]. Traffic signs also have specific shapes that can be searched for by shape-based methods. The Hough Transform is one of the most common shape-based methods [35], [36], as it is relatively robust against illumination change and image noise. Similarity detection [37] and Distance Transform matching [37] also constitute shape-based

methods. Hybrid approaches take advantage of both sign color and shape [38], [39]. Classification stages mostly employ template matching [40], [41], SVM [42], [43], Genetic Algorithm (GA) [44], Artificial Neural Network (ANN) [45], [46], AdaBoost [47], [48] and deep learning-based methods. In recent years, deep learning methods have increasingly attracted a great deal of attention. Convolutional Neural Networks (CNNs) constitute a subset of deep neural network models that have the power to learn robust and discriminative features from raw data. There is a variety of CNN that have been employed for traffic sign recognition such as small-scale CNN [49], multi-scale CNN [50], a committee of CNN [51], multi-column CNN [52], multi-task CNN [53], and CNN-SVM [54], [55], among others. A number of traffic sign datasets have been created in the past decade. Most of them were recorded in European countries. Consequently, methods that have been proposed in the literature are mostly based on European datasets. As Traffic signs in North America have different colors and shapes, the methods that have been proposed based on European traffic signs are not directly suitable in the North American context [56].

3.2.3 Vehicle Detection

Many traditional vehicle detection approaches comprise a Hypothesis Generation (HG) step followed by a Hypothesis Verification (HV) step. With regards to HG, there are various methods that can be divided into three basic categorizes including knowledge-based, stereo-based, and motion-based [57]. Knowledge-based methods use prior knowledge including shadows [58], symmetry [59], horizontal/vertical edges [60], color [61], texture [62], corners [63], and vehicle lights [64]. Stereo-based approaches usually exploit the Inverse Perspective Mapping (IPM) [65] or disparity maps [66] to localize vehicles,

while motion-based methods detect vehicles with optical flow [67]. HV approaches can be classified into two major categories [57]: template-based and appearance-based. The former employs predefined vehicle patterns and estimates the correlation between templates and candidate image regions [68], while the latter uses machine learning methods such as SVM [38], [39], ANN [69], and AdaBoost [70] to classify hypotheses into vehicle and non-vehicle categories.

Classifiers such as SVM [71], ANN [69], and AdaBoost [70] learn the characteristics of vehicle appearance to draw a decision boundary between vehicle and non-vehicle classes. In HV, a number of local feature descriptors such as HOG [72], PHOG [59], Harr-like [73], Gabor [74], and SURF [75] have shown a remarkable ability in collecting contextual information. Additionally, different vehicle detection approaches that employ deep learning-based methods discussed in Section 3.2.1 have been proposed. For instance, in [76], the authors provided a comparative study on the performance of Alex Net and Faster R-CNN models. Also, in [77], the authors exploited the fine-tuned YOLO [25] for vehicle detection. In [78], vehicles are detected with a simplified Fast R-CNN.

3.2.4 Pedestrian Detection

Many traditional methods for pedestrian detection have been proposed with the majority of them using features such as HOG [79], Haar-Like [79], Viola-Jones [80], and LBP [81], followed by a classification stage using either SVM, ANN, or AdaBoost. Additionally, pedestrian detection methods using deep learning can be categorized as either single-stage or two-stage techniques. RPN+BF [82], Fast R-CNN [83], and Faster R-CNN [19] are examples where the authors employed a two-stage approach. Examples of single-stage ap-

proaches have been proposed: For instance, Lan *et al.* [84] modified YOLO-v2 into a single-stage network called YOLO-R for pedestrian detection. Comprehensive surveys on pedestrian detection are provided in [85] and [86].

3.2.5 Traffic Light Detection

Color segmentation is a method often used to reduce the search space in traffic scene images. For example, in [87] and [88], the authors employed HSV and YCbCr color spaces respectively to detect traffic lights. In some studies, shape-based methods such as the circular Hough transform [89] was used after color segmentation to find round traffic lights. Blob detection is another approach to detect traffic lights that analyses the size and aspect ratio of the traffic lights to eliminate regions likely to produce false positives [90]. In [91], saliency maps are employed to detect traffic lights. In [92], GPS data and digital maps are used to identify traffic lights in urban areas. Feature descriptors such as HOG [87], Haar-like [93], and Gabor Wavelet [88] have been extensively used to detect traffic lights. To recognize the state of traffic lights, several methods have been employed, mostly including SVM [94], fuzzy algorithms [95] and more recently, deep learning methods. A simple CNN was used by Lee and Park [96] for traffic light classification. Behrendt *et al.* [2] applied YOLO-v1 for traffic light detection and classification. In [97], YOLO-9000 [26] was applied to the LISA traffic light dataset. The authors in [98] exploited DeepTLR networks for real-time traffic light detection and classification. A novel Faster R-CNN hierarchical architecture was proposed in [99] and trained on a joint traffic light and sign dataset.

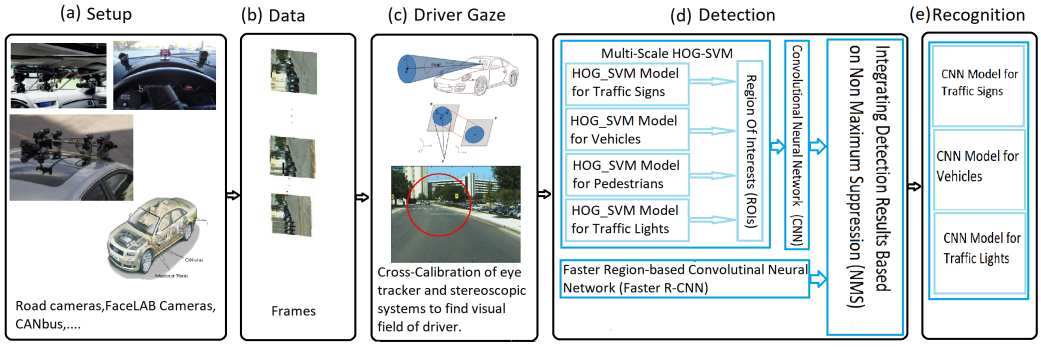


Figure 3.1: **Framework Overview.** *Our framework detects and recognizes traffic objects inside the visual field of driver.* **(from left to right: a)** *The RoadLAB vehicle with forward stereoscopic and eye-tracking systems.* **b)** *Dataset created with the RoadLAB experimental vehicle.* **c)** *Computing the radius of driver's view as attentional gaze cone and locating the re-projected 2D ellipse of the visual field of the driver.* **d)** *We used two different model types in the detection stage of the framework; Model A consists of two steps including multi-scale HOG-SVM followed by applying a CNN, and Model B is a Faster Region-based CNN. Detection results are integrated by a NMS-based algorithm.* **e)** *For the recognition stage, we separately trained three independent models on traffic signs, vehicles, and traffic lights.*

3.3 Proposed Method

In this Section, we describe our proposed method for traffic object detection and recognition based on the attentional visual field of the driver. First, the dataset used in this research is introduced. Following this, we describe the method employed to find the attentional gaze area of the driver in the forward stereo imaging system. Next, in the object detection stage, our trained models and the methods used for enriching our data set are described. We then discuss the Region of Interest (ROI) integration method we used. Finally, the object recognition stage is presented. Figure 3.1 illustrates our proposed framework.

3.3.1 The RoadLAB Dataset

An essential element of deep learning-based object detection systems is the availability of a large number of sample images. In this Section, we present our own object dataset from the RoadLAB experimental data sequences [100]. The RoadLAB project included an instrumented vehicle capable of recording the following items:

- front view of the driving environment using calibrated stereo cameras mounted on the roof of the vehicle
- vehicle dynamic features such as odometry and steering wheel angle
- driver cephalo-ocular behavioral features including head pose and 3D driver gaze direction

Sixteen driving sequences were collected by our experimental vehicle on a pre-determined, 28.5-kilometer course in the city of London, ON, Canada (details provided in Table 3.1). Figure 3.2 illustrates the forward stereoscopic system and the eye tracking system as part of the vehicular instrumentation.

As one of our contributions in this study, in order to train, validate and test our models, we collected 13,546 sample images to detect and recognize traffic objects including traffic signs, vehicles, pedestrians and traffic lights. Our dataset contains 3,225 sample images for the background class in addition to 5,172, 1,984, 1,290 and 1,875 sample images for the object classes of traffic sign, vehicle, pedestrian and traffic light respectively. The vehicle class consists of 3 distinct classes including car, bus and truck. The traffic light class consists of 4 distinct classes including red, yellow, green and not clear. Finally, the raffic sign class includes 19 distinct classes of traffic signs. Additionally, some traffic sign classes include more than one sign type such as “Maximum Speed Limit”,

Figure 3.2: **(top):** *Forward stereoscopic vision system on rooftop.* **(bottom):** *Infrared gaze tracker.*

Table 3.1: DATA DESCRIPTION (EACH SEQUENCE CORRESPONDES TO ONE DRIVER.)

Seq. #	Date of Capture	Weather Conditions	Gender
1	2012-08-24	29 °C Sunny	M
2	2012-08-24	31 °C Sunny	M
3	2012-08-30	23 °C Sunny	F
4	2012-08-31	24 °C Sunny	M
5	2012-09-05	27 °C Partially Cloudy	F
6	2012-09-10	21 °C Partially Cloudy	F
7	2012-09-12	21 °C Sunny	F
8	2012-09-12	27 °C Sunny	M
9	2012-09-17	24 °C Partially Cloudy	F
10	2012-09-19	8 °C Sunny	M
11	2012-09-19	12 °C Sunny	F
12	2012-09-21	18 °C Partially Cloudy	F
13	2012-09-21	19 °C Partially Cloudy	M
14	2012-09-24	7 °C Sunny	F
15	2012-09-24	13 °C Partially Cloudy	F
16	2012-09-28	14 °C Partially Cloudy	M

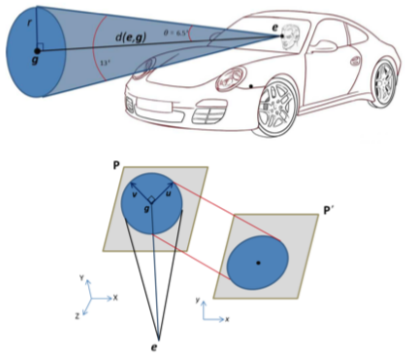


Figure 3.3: **(top):** *Depiction of the driver attentional gaze cone.* **(bottom):** *Re-projection of the 3D attentional circle into the corresponding 2D ellipse on image plane of the forward stereo scene system.*

"Construction", "Parking", etc. Our samples for traffic signs can be considered as a complete sign dataset including warning signs, regulatory signs, direction signs, and temporary signs. The main point of comparison work is to compare with the Roadlab dataset and the performance achieved by zabihi et al. [56] for the reason of practically.

3.3.2 Driver Gaze Localization

The visual attentional field of the driver consists of a circle in 3D space within the plane that contains the Point of Gaze (PoG), perpendicular to the Line of Gaze (LoG). The radius of the circle is determined by the angular opening of the cone of visual attention as shown in Figure 3.3. The circle generally projects onto the imaging plane of the stereo sensor as a 2D ellipse. We describe the procedure we employed, as per Kowsari *et al.* [101].

First, both the eye position $e = (e_x, e_y, e_z)$ and the 3D PoG $g = (g_x, g_y, g_z)$ are transormed into the reference frame of the forward stereo sensor. Next, the radius of the circular attentional gaze area is obtained by computing the

Euclidean distance between e and g (θ is set to 6.5 °: [102]).

$$r = \tan(\theta)\|e - g\|_2 \tag{3.1}$$

We re-project the obtained circle contained in the 3D plane perpendicular to the LoG onto the image plane of the forward stereo imaging sensor where it becomes an ellipse. The coordinates of the ellipse are obtained as:

$$(X, Y, Z) = g + r(\cos\phi\mathbf{u} + \sin\phi\mathbf{v}) \tag{3.2}$$

where $\mathbf{u}$=(u_x, u_y, u_z) and $\mathbf{v}$=(v_x, v_y, v_z) are two orthonormal vectors in the plane orthogonal to the LoG and $\phi \in [0, 2\pi]$. Using perspective projection $x = \frac{X}{Z}$ and $y = \frac{Y}{Z}$ and applying the intrinsic calibration matrix of the stereo scene system from [101] yields the 2D ellipse on the image plane of the forward stereo sensor. The mathematical details are found in [101] and [103]. Figure 3.4 illustrates several attentional visual areas for several sample frames.

3.3.3 Object Detection Stage

To detect traffic objects of interest inside and outside of the attentional field of the driver, we employed a framework consisting of two different model types that we proceed to describe:

Model A

The first model consists of two steps that include a multi-scale HOG-SVM followed by the use of a ResNet-101 network. We used the multi-scale HOG-SVM because of the model's simplicity compared to the other model, such as cascade-RCNN. The multi-scale HOG-SVM descriptor counts occurrences of

Figure 3.4: *Examples of attentional gaze areas projected onto the forward stereo sensor of the vehicle.*

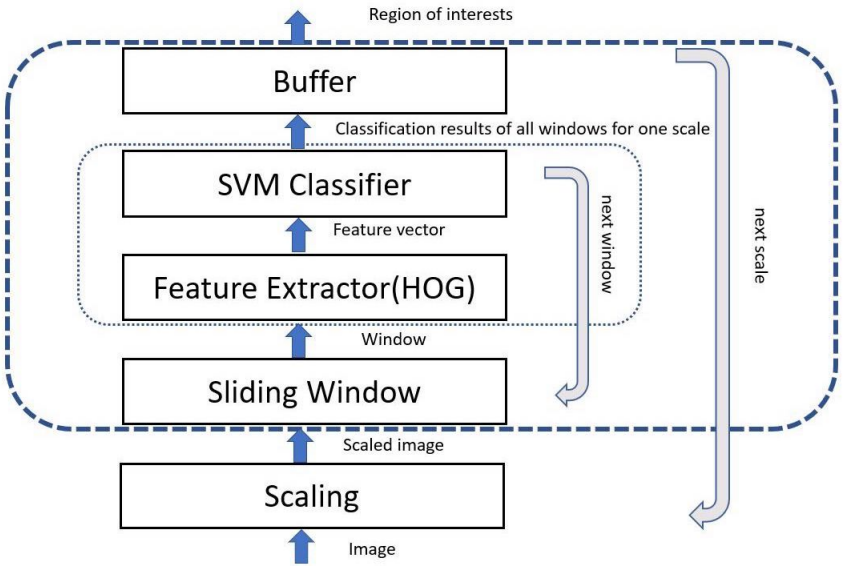


Figure 3.5: *Internal view of a multi-scale HOG-SVM*

gradient orientations in an image region followed by a block-normalization algorithm that results in better invariance to edge contrast and shadows. Since it operates on local cells, it is also relatively invariant to geometric and photometric transformations. In general, the detection algorithm is based on an overlapping sliding window approach. Since the Region of Interest (RoI) contains objects that vary in size, we used a multi-scale method for the object detection problem. We treat the HOG features extracted from each sliding window at each level as independent samples prior to feeding them to the SVM classifier. Figure 3.5 illustrates the internal view of multi-scale HOG-SVM.

We trained four independent multi-scale HOG-SVM models to find RoIs, for our four types of traffic objects (signs, vehicles, pedestrians, and traffic lights). The model moves a sliding window across the images and HOG features are extracted. The model follows this strategy at several imaging scales. Typically, SVM outputs conventional binary decision labels. However, it can also provide a probabilistic confidence score [104] for each sliding window, which we use to threshold on RoIs. With the use of HOG-SVM, we discard the RoIs labelled as background while other candidates are transferred to the next stage of processing.

The remaining ROIs from the HOG-SVM classifier were categorized into five classes: background, traffic sign, vehicle, pedestrian and traffic light. In the second stage we applied ResNet-101 [105], which is a popular CNN that has been already trained with more than a million images from the ImageNet database [106]. Figure 3.6 illustrates sample results obtained with this model. However, we noted the multi-scale HOG-SVM sometimes had difficuly localizing vehicles occupying a large part of the image (Figure 3.7 illustrates this problem). Hence, we also used a Faster R-CNN model to detect vehicles in

Figure 3.6: *Model A output examples.*

Figure 3.7: *Examples of model A missing large vehicle objects.*

Figure 3.8: *Model B output examples.*

parallel with Model A.

Model B

We trained a Faster R-CNN model on our dataset to localize vehicles. We observed that Model B outperforms Model A for detecting vehicles that occupy a large image area, or that are very close to the instrumented vehicle. Conversely, Faster R-CNN cannot effectively detect objects that are low in resolution or small in size [107], [108] and [109]. We integrated the results from both Models A and B to cirumvent their respective weaknesses. The hypotheses generated in this stage are directly transferred to an integration stage where detection results are merged. Figure 3.8 displays vehicle detections obtained with Model B.

3.3.4 Data Augmentation

In addition to collecting over 10,000 sample object images, to further enrich our training dataset, we employed a data augmentation technique and a boosting algorithm. Through data augmentation, we made our dataset greater by adding the translated, rotated, scaled, and sheared versions of our original samples resulting in increased performance at the detection stage. To boost the performance of our models, we employed an advanced learning method known as Hard Examples Mining (HEM). HEM refers to examples that are mislabeled by the current version of the model. We trained the SVM, Resnet, and Faster R-CNN models in an iterated procedure on a portion of the training data, and at each iteration, the detector models were applied to a number of unseen images from the training data. Then, we corrected the mislabeled results in preparation for the next iteration. We finally provided the models with additional key samples which made them more robust.

3.3.5 Integrating Detection Results

After completing the detection stage on test images, in order to improve the detection performance, we eliminated redundant detections and merged the remaining ones into a set of integrated results. For this, we used a method that is based on Non Maximum Suppression (NMS) [56], [30]. When multiple bounding boxes overlap, NMS retains the highest-scored bounding box and eliminates any other whose overlap ratio exceeds a preset threshold. We employed Pascal's overlap score [110] to find the overlap ratio a between them. This ratio is obtained as:

$$a = \frac{area(B_1 \cap B_2)}{area(B_1 \cup B_2)} \tag{3.3}$$

where B_1 and B_2 are two overlapping bounding boxes.

The NMS algorithm is not practical in all situations. For instance, consider a situation in which a vehicle is partially occluded by a pedestrian, and both of them are detected. If their overlap ratio is greater than the threshold, NMS wrongly eliminates the lower-scored object. To address this case, we integrated all bounding boxes in three steps. We considered a lower bound and a upper bound threshold for the overlap ratio. In the first step, we employ NMS to merge bounding boxes that belong to the same class. In this step, NMS eliminates the lower-scored bounding boxes whose overlap ratios are between the lower bound and the upper bound thresholds. In the second step, if bounding boxes belong to the same class and their overlap ratio is greater than the upper bound threshold, they are merged into a larger bounding box. In the last step, all remaining bounding boxes are merged without employing NMS to generate the final set of detected hypotheses.

3.3.6 Object Recognition Stage

The output of the detection stage is a list of candidate objects that have been labeled with the class they belong to (traffic sign, vehicle, traffic light, and pedestrian). Except for pedestrian objects, the remaining objects from the list are considered for further analysis at this stage. We separately trained three independent models on traffic signs, vehicles, and traffic lights by using ResNet-101 for recognizing the remaining objects. After feeding the candidate object (hypothesis) into its corresponding model, the classifier decides whether the object in the list is either a rejected object or a recognized object and, in this case, the classifier responds with the appropriate class name. More precisely, the traffic light recognizer is able to classify traffic light hypotheses

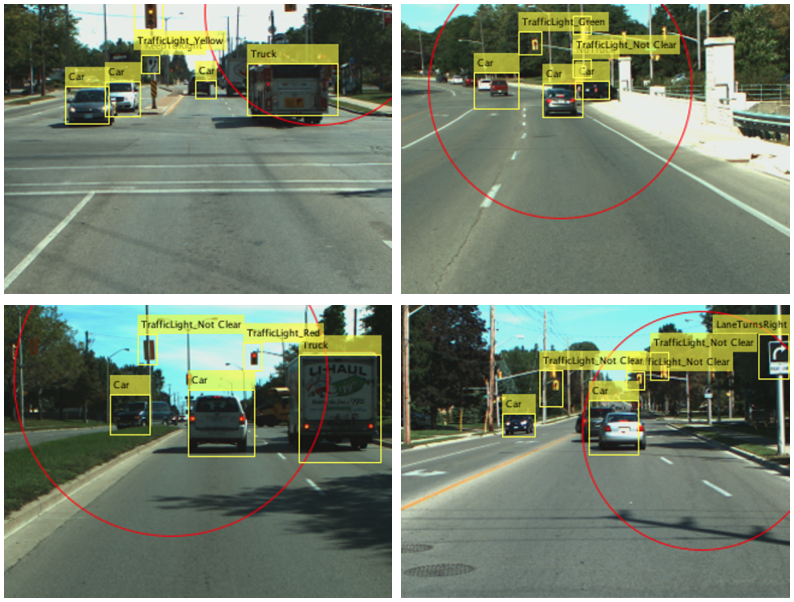


Figure 3.9: *Output samples from the proposed framework superimposed on the attentional visual field of the driver*

into five classes, the vehicle recognizer is able to classify vehicle hypotheses into four classes, while the traffic sign recognizer classifies traffic sign hypotheses into twenty classes. Fig 3.9 shows a sample of results from the proposed framework for four classes of traffic objects.

3.4 Experimental Results

We employed the driving sequences captured with the RoadLAB experimental vehicle [100] and our dataset as described in Section 3.3.1. The proposed method was used to detect and recognize traffic objects inside and outside of the attentional visual area of the driver. We provide the parameters which have been used in our experiments. Then we report on our experimental results for the proposed detection and recognition stages in detail.

3.4.1 Parameters

Table 3.2: DESCRIPTION OF DATA AUGMENTATION

Method	Description	Range
Translate	Each image is translated in the horizontal and vertical direction by a distance, in pixels	$(-10, 10)$
Rotate	Each image is rotated by an amount, in degrees	$(-15, 15)$
Scale	Each image is scaled in the horizontal and vertical direction by a factor	$(0.5, 1.5)$
Shear	Each image is sheared along the horizontal or vertical axis by a factor	$(-30, 30)$

To obtain fine-tuned parameters for each classifier model, we used cross-validation experiments on our training dataset. We divided the training data into a basic training set and a validation set. Then, the basic training set was used to train the classifier and subsequently, the validation set was used to evaluate the model. By exploring various ranges for the tuning parameters, we selected the parameter settings that resulted in maximum validation accuracy. Next, the classifier was re-trained on the complete training set using the fine-tuned parameters. Finally, we tested the models on the pre-separated unseen data that consists of a set of randomly selected samples.

We applied a threshold to the score values that each SVM model provided, and RoIs were considered for post-processing only if their SVM score was higher than the threshold value. These score values ranged from 0 (definitely negative) to 1 (definitely positive). We selected the threshold that allowed a maximum of true positives. While some false positives passed this stage, they could mostly be elimiated in the following stage of processing.

Threshold values of 0.50, 0.40, 0.40, and 0.60 were applied to the SVM

models for detection of traffic signs, pedestrians, traffic lights, and vehicles respectively. These values provided the best results. We also utilized different augmentation methods to improve the performance of our models. Table 3.2 lists the methods we have used to augment our data.

3.4.2 Results for the Object Detection Stage

In the following Subsections, we discuss the results we obtained for the object detection stage in detail.

Assessing the Accuracy of the Trained ResNet-101 CNN Model

As described in Section 3.3.3, after localizing RoIs by way of multi-scale HOG-SVM, a ResNet-101 CNN was trained and used on our dataset to verify and categorize RoIs into our five classes of traffic objects. We computed the confusion matrix from the ResNet-101 model on the test data (See Figure 3.10). The model classifies the test data correctly in 94.1% of cases. Notably, 10% of vehicles have been incorrectly classified as background by ResNet 101. As a result, we employed a Faster R-CNN-based model to detect vehicles besides Model A.

Assessing the Accuracy of the Object Detection Stage

To verify the accuracy of the object detection stage, we report the Detection Rate (DR) and the number of False Positives Per Frame (FPPF), defined as follows:

$$\mathrm{DR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{3.4}$$

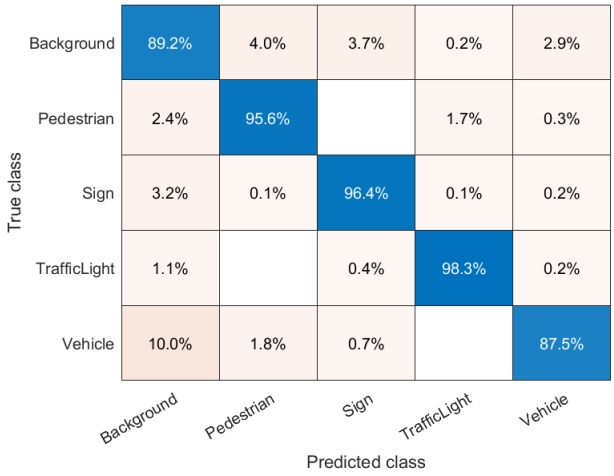


Figure 3.10: *Confusion matrix from trained ResNet-101 for labelling of traffic object classes.*

Table 3.3: DESCRIPTION OF DETECTION RESULTS

Description	**DR**	**FPPF**
traffic lights	0.93	0.03
pedestrians	0.88	0.11
traffic signs	0.91	0.06
vehicles	0.92	0.04
object detection stage, 4 object classes	0.91	0.06
previous work [56] for traffic signs	0.84	0.04

$$\text{FPPF} = \frac{\text{FP}}{\text{F}} \tag{3.5}$$

where TP is the number of correctly detected objects, FN is the number of objects that are wrongly not detected, FP is the number of incorrectly detected objects, and F is the total number of frames.

As shown in Table 3.3, our proposed object detection framework achieved 0.91 and 0.06 for DR and FPPF respectively. Previously, Zabihi *et al.* [56]

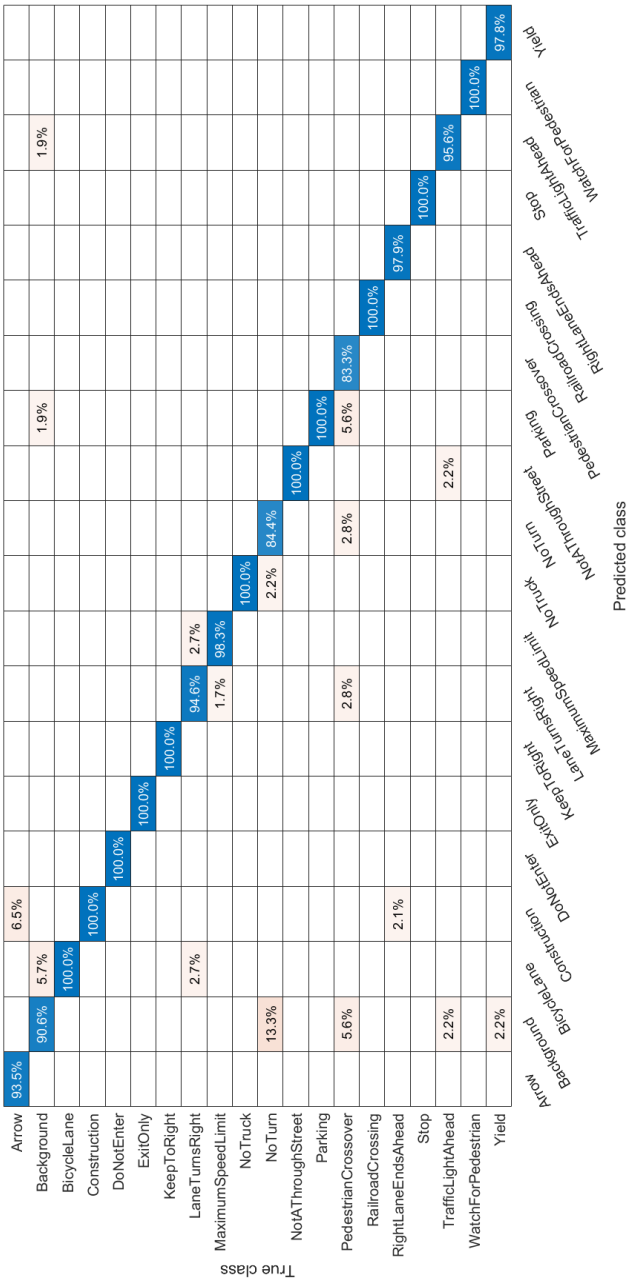


Figure 3.11: *Confusion matrix from trained ResNet-101 for traffic sign recognition.*

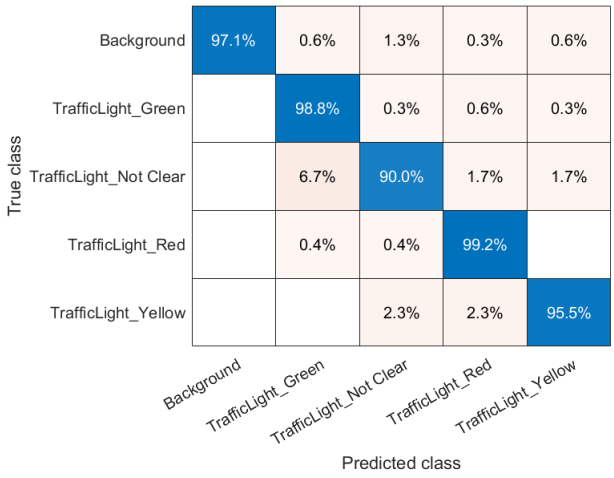


Figure 3.12: *Confusion matrix from trained ResNet-101 for traffic light recognition.*

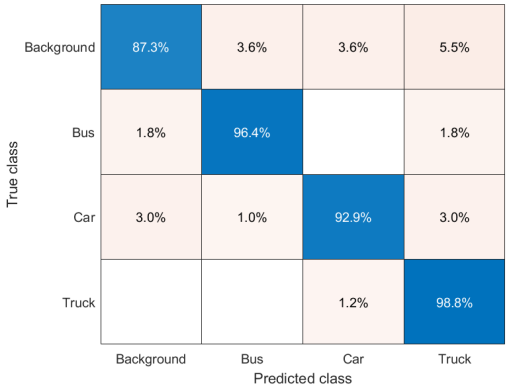


Figure 3.13: *Confusion matrix from trained ResNet-101 for vehicle recognition.*

detected traffic signs only from the RoadLAB dataset and reported 0.84 for DR and 0.04 for FPPF (last row of Table 3.3). Our model for traffic sign detection, when compared with the work from Zabihi *et al.* [56], has reached 0.07 more accuracy for DR and shows an increase in FPPF of 0.02.

3.4.3 Results for Object Recognition Stage

The object recognition stage is applied to the output of the object detection stage to recognize hypotheses and to provide a classification result. We trained three separate ResNet-101 models for classes corresponding to traffic signs, traffic lights, and vehicles using our training dataset. To verify the accuracy of the object recognition stage, we computed the confusion matrix for each class, as displayed in Figures 3.11, 3.12, and 3.13.

Results for traffic sign recognition (Fig 3.11) show that the model reached 96.1% accuracy with our Canadian traffic sign dataset. The largest values along the main diagonal indicate that the majority of the test sign images were classified correctly. The lowest correct response of 83.3% was obtained for the class *PedestrianCrossover*.

Fig 3.12 illustrates the confusion matrix for traffic light recognition. The results show that the model has reached 96.2% of overall correct classification. As can be seen, the lowest degree of correctly categorized classes belongs to class *NotClear* while classes *Green* and *Red* obtained 98.8% and 99.2% respectively.

The results shown in Figure 3.13 indicate that the vehicle recognizer model achieved 94.8% of overall correct classification. This confusion matrix shows that this model is able to discriminate vehicle objects (i.e. vehicle, bus, and truck) with less than 3% of mislabeling error. The *background* class achieved

the least accuracy with 87.3%.

3.5 Conclusion

We conducted a literature review of detection and recognition approaches for four important classes of traffic objects including traffic signs, vehicles, pedestrians and traffic lights. Generally, the availability of suitable and adequate training data is a vital element in the learning process, in order to achieve a discriminative model. In this work, we collected over 10,000 object sample images from sequences belonging to the RoadLAB initiative [100]. We also enriched our training data using augmentation and a HEM strategy. We localized the attentional visual area of the driver onto the imaging plane of the forward stereoscopic system, and a framework for the detection and recognition of traffic objects located inside and outside the attentional visual field of drivers was devised. We considered 3, 4, and 19 different classes for vehicles, traffic lights, and traffic signs respectively. The object detection stage was built from a combination of both traditional and deep learning-based models to detect objects at various scales. Finally, in the recognition stage, by means of trained ResNet-101 networks, our framework achieved 96.1%, 96.2% and 94.8% of correct classification for traffic signs, traffic lights, and vehicles respectively.

Bibliography

[1] Tian B, Morris BT, Tang M, Liu Y, Yao Y, Gou C, et al. Hierarchical and Networked Vehicle Surveillance in ITS: A Survey. IEEE Transactions on Intelligent Transportation Systems. 2017;18(1):25.

[2] Behrendt K, Novak L, Botros R. A deep learning approach to traffic lights: Detection, tracking, and classification. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE; 2017. p. 1370–1377.

[3] Diaz M, Cerri P, Pirlo G, Ferrer MA, Impedovo D. A survey on traffic light detection. In: International Conference on Image Analysis and Processing. Springer; 2015. p. 201–208.

[4] Fairfield N, Urmson C. Traffic light mapping and detection. In: 2011 IEEE International Conference on Robotics and Automation. IEEE; 2011. p. 5421–5426.

[5] Zhao Z, Zheng P, Xu S, Wu X. Object detection with deep learning: A review. IEEE transactions on neural networks and learning systems. 2019;.

[6] Zhiqiang W, Jun L. A review of object detection based on convolutional neural network. In: 2017 36th Chinese Control Conference (CCC). IEEE; 2017. p. 11104–11109.

[7] Tiwari M, Singhai R. A review of detection and tracking of object from image and video sequences. Int J Comput Intell Res. 2017;13(5):745–765.

[8] Zou Z, Shi Z, Guo Y, Ye J. Object Detection in 20 Years: A Survey. arXiv preprint arXiv:190505055. 2019;.

[9] Wang X, Yang M, Zhu S, Lin Y. Regionlets for generic object detection. In: Proceedings of the IEEE international conference on computer vision; 2013. p. 17–24.

[10] Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2014. p. 580–587.

[11] Viola P, Jones MJ. Robust real-time face detection. International journal of computer vision. 2004;57(2):137–154.

[12] Kumar R, Kumar A, Bhavsar A. Bird region detection in images with multi-scale HOG features and SVM scoring. In: Proceedings of 2nd International Conference on Computer Vision & Image Processing. Springer; 2018. p. 353–364.

[13] Dürre J, Paradzik D, Blume H. A HOG-based real-time and multi-scale pedestrian detector demonstration system on FPGA. In: Proceedings of the 2018 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays. ACM; 2018. p. 163–172.

[14] Dollár P, Appel R, Belongie S, Perona P. Fast feature pyramids for object detection. IEEE transactions on pattern analysis and machine intelligence. 2014;36(8):1532–1545.

[15] Ohn-Bar E, Trivedi M. Fast and robust object detection using visual subcategories. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2014. p. 179–184.

[16] Mathias M, Timofte R, Benenson R, Van Gool L. Traffic sign recognition—How far are we from the solution? In: The 2013 international joint conference on Neural networks (IJCNN). IEEE; 2013. p. 1–8.

[17] Suhao L, Jinzhao L, Guoquan L, Tong B, Huiqian W, Yu P. Vehicle type detection based on deep learning in traffic scene. Procedia computer science. 2018;131:564–572.

[18] Girshick R. Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision; 2015. p. 1440–1448.

[19] Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems; 2015. p. 91–99.

[20] Dai J, Li Y, He K, Sun J. R-fcn: Object detection via region-based fully convolutional networks. In: Advances in neural information processing systems; 2016. p. 379–387.

[21] He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence. 2015;37(9):1904–1916.

[22] He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 2961–2969.

[23] Najibi M, Rastegari M, Davis LS. G-cnn: an iterative grid based object detector. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 2369–2377.

[24] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C, et al. Ssd: Single shot multibox detector. In: European conference on computer vision. Springer; 2016. p. 21–37.

[25] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 779–788.

[26] Redmon J, Farhadi A. YOLO9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 7263–7271.

[27] Redmon J, Farhadi A. Yolov3: An incremental improvement. arXiv preprint arXiv:180402767. 2018;.

[28] Shen Z, Liu Z, Li J, Jiang Y, Chen Y, Xue X. Dsod: Learning deeply supervised object detectors from scratch. In: Proceedings of the IEEE International Conference on Computer Vision; 2017. p. 1919–1927.

[29] Fu C, Liu W, Ranga A, Tyagi A, Berg AC. Dssd: Deconvolutional single shot detector. arXiv preprint arXiv:170106659. 2017;.

[30] Hu Q, Paisitkriangkrai S, Shen C, van den Hengel A, Porikli F. Fast detection of multiple objects in traffic scenes with a common detection framework. IEEE Transactions on Intelligent Transportation Systems. 2015;17(4):1002–1014.

[31] Wali SB, Abdullah MA, Hannan M, Hussain A, Samad SA, Ker PJ, et al. Vision-based traffic sign detection and recognition systems: current trends and challenges. Sensors. 2019;19(9):2093.

[32] De la Escalera A, Armingol J, Mata M. Traffic sign recognition and analysis for intelligent vehicles. Image and vision computing. 2003;21(3):247–258.

[33] Kuo W, Lin C. Two-stage road sign detection and recognition. In: 2007 IEEE international conference on multimedia and expo. IEEE; 2007. p. 1427–1430.

[34] Saadna Y, Behloul A. An overview of traffic sign detection and classification methods. International Journal of Multimedia Information Retrieval. 2017;6(3):193–210.

[35] Nandi D, Saif S, Prottoy P, Zubair K, Shubho S. Traffic sign detection based on color segmentation of obscure image candidates: a comprehensive study. International Journal of Modern Education and Computer Science. 2018;10(6):35.

[36] Yin S, Ouyang P, Liu L, Guo Y, Wei S. Fast traffic sign recognition with a rotation invariant binary pattern based feature. Sensors. 2015;15(1):2161–2180.

[37] Saxena P, Gupta N, Laskar S, Borah P. A study on automatic detection and recognition techniques for road signs. Int J Comput Eng Res. 2015;5(12):24–28.

[38] De La Escalera A, Moreno L, Salichs M, Armingol JM. Road traffic sign detection and classification. IEEE transactions on industrial electronics. 1997;44(6):848–859.

[39] Prisacariu V, Timofte R, Zimmermann K, Reid I, Van Gool L. Integrating object detection with 3D tracking towards a better driver assistance system. In: 2010 20th International Conference on Pattern Recognition. IEEE; 2010. p. 3344–3347.

[40] Torresen J, Bakke JW, Sekanina L. Efficient recognition of speed limit signs. In: Proceedings. The 7th International IEEE Conference on Intelligent Transportation Systems (IEEE Cat. No. 04TH8749). IEEE; 2004. p. 652–656.

[41] Greenhalgh J, Mirmehdi M. Traffic sign recognition using MSER and random forests. In: 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO). IEEE; 2012. p. 1935–1939.

[42] Xing M, Chunyang M, Yan W, Xiaolong W, Xuetao C. Traffic sign detection and recognition using color standardization and Zernike moments. In: 2016 Chinese Control and Decision Conference (CCDC). IEEE; 2016. p. 5195–5198.

[43] Zhang H, Wang B, Zheng Z, Dai Y. A novel detection and recognition system for Chinese traffic signs. In: Proceedings of the 32nd Chinese Control Conference. IEEE; 2013. p. 8102–8107.

[44] Kobayashi M, Baba M, Ohtani K, Li L. A method for traffic sign detection and recognition based on genetic algorithm. In: 2015 IEEE/SICE International Symposium on System Integration (SII). IEEE; 2015. p. 455–460.

[45] Hannan MA, Hussain A, Samad SA, Ishak KA. A unified robust algorithm for detection of human and non-human object in intelligent safety application. International Journal of Computer and Information Engineering. 2008;2(11).

[46] Hechri A, Mtibaa A. Automatic detection and recognition of road sign

for driver assistance system. In: 2012 16th IEEE Mediterranean Electrotechnical Conference. IEEE; 2012. p. 888–891.

[47] Chen L, Li Q, Li M, Mao Q. Traffic sign detection and recognition for intelligent vehicle. In: 2011 IEEE Intelligent Vehicles Symposium (IV). IEEE; 2011. p. 908–913.

[48] Lin C, Wang M. Road sign recognition with fuzzy adaptive preprocessing models. Sensors. 2012;12(5):6415–6433.

[49] Zhu Y, Zhang C, Zhou D, Wang X, Bai X, Liu W. Traffic sign detection and recognition using fully convolutional network guided proposals. Neurocomputing. 2016;214:758–766.

[50] Lee H, Kim K. Simultaneous traffic sign detection and boundary estimation using convolutional neural network. IEEE Transactions on Intelligent Transportation Systems. 2018;19(5):1652–1663.

[51] Sermanet P, LeCun Y. Traffic sign recognition with multi-scale Convolutional Networks. In: IJCNN; 2011. p. 2809–2813.

[52] Fang C, Fuh C, Yen PS, Cherng S, Chen S. An automatic road sign recognition system based on a computational model of human recognition processing. Computer vision and Image understanding. 2004;96(2):237–268.

[53] Kumar AD. Novel deep learning model for traffic sign detection using capsule networks. arXiv preprint arXiv:180504424. 2018;.

[54] Lim K, Hong Y, Choi Y, Byun H. Real-time traffic sign recognition based on a general purpose GPU and deep-learning. PLoS one. 2017;12(3):e0173317.

[55] Lai Y, Wang N, Yang Y, Lin L. Traffic Signs Recognition and Classification based on Deep Feature Learning. In: ICPRAM; 2018. p. 622–629.

[56] Zabihi SJ, Zabihi SM, Beauchemin SS, Bauer MA. Detection and recognition of traffic signs inside the attentional visual field of drivers. In: 2017 IEEE Intelligent Vehicles Symposium (IV). IEEE; 2017. p. 583–588.

[57] Sun Z, Bebis G, Miller R. On-road vehicle detection: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2006;28(5):694–711.

[58] Tzomakas C, von Seelen W. Vehicle detection in traffic scenes using shadows. In: Ir-Ini, Institut fur Nueroinformatik, Ruhr-Universitat. Citeseer; 1998. .

[59] Khairdoost N, Monadjemi SA, Jamshidi K. Front and rear vehicle detection using hypothesis generation and verification. Signal & Image Processing. 2013;4(4):31.

[60] Mu K, Hui F, Zhao X, Prehofer C. Multiscale edge fusion for vehicle detection based on difference of Gaussian. Optik. 2016;127(11):4794–4798.

[61] Xiong T, Debrunner C. Stochastic car tracking with line-and color-based features. IEEE Transactions on Intelligent Transportation Systems. 2004;5(4):324–328.

[62] Bucher T, Curio C, Edelbrunner J, Igel C, Kastrup D, Leefken I, et al. Image processing and behavior planning for intelligent vehicles. IEEE Transactions on Industrial electronics. 2003;50(1):62–75.

[63] Bertozzi M, Broggi A, Castelluccio S. A real-time oriented system for vehicle detection. Journal of Systems Architecture. 1997;43(1-5):317–325.

[64] Chen D, Lin Y, Peng Y. Nighttime brake-light detection by Nakagami imaging. IEEE Transactions on Intelligent Transportation Systems. 2012;13(4):1627–1637.

[65] Broggi A, Bertozzi M, Fascioli A, Bianco C, Piazzi A. Visual perception of obstacles and vehicles for platooning. IEEE Transactions on Intelligent Transportation Systems. 2000;1(3):164–176.

[66] Franke U, Kutzbach I. Fast stereo based object detection for stop&go traffic. In: Proceedings of Conference on Intelligent Vehicles. IEEE; 1996. p. 339–344.

[67] Liu Y, Lu Y, Shi Q, Ding J. Optical flow based urban road vehicle tracking. In: 2013 Ninth International Conference on Computational Intelligence and Security. IEEE; 2013. p. 391–395.

[68] Handmann U, Kalinke T, Tzomakas C, Werner M, Seelen WV. An image processing system for driver assistance. Image and Vision Computing. 2000;18(5):367–376.

[69] Goerick C, Noll D, Werner M. Artificial neural networks in real-time car detection and tracking applications. Pattern Recognition Letters. 1996;17(4):335–343.

[70] Song G, Lee K, Lee J. Vehicle detection by edge-based candidate generation and appearance-based classification. In: 2008 IEEE Intelligent Vehicles Symposium. IEEE; 2008. p. 428–433.

[71] Truong Q, Lee B. Vehicle detection algorithm using hypothesis generation and verification. In: International Conference on Intelligent Computing. Springer; 2009. p. 534–543.

[72] Yan G, Yu M, Yu Y, Fan L. Real-time vehicle detection using histograms of oriented gradients and AdaBoost classification. Optik. 2016;127(19):7941–7951.

[73] Wen X, Shao L, Fang W, Xue Y. Efficient feature selection and classification for vehicle detection. IEEE Transactions on Circuits and Systems for Video Technology. 2014;25(3):508–517.

[74] Ming Q, Jo K. Vehicle detection using tail light segmentation. In: Proceedings of 2011 6th International Forum on Strategic Technology. vol. 2. IEEE; 2011. p. 729–732.

[75] Lin B, Chan Y, Fu L, Hsiao P, Chuang L, Huang S, et al. Integrating appearance and edge features for sedan vehicle detection in the blind-spot area. IEEE Transactions on Intelligent Transportation Systems. 2012;13(2):737–747.

[76] Espinosa JE, Velastin SA, Branch JW. Vehicle detection using alex net and faster R-CNN deep learning models: a comparative study. In: International Visual Informatics Conference. Springer; 2017. p. 3–15.

[77] Zhou Y, Nejati H, Do T, Cheung N, Cheah L. Image-based vehicle analysis using deep neural network: A systematic study. In: 2016 IEEE International Conference on Digital Signal Processing (DSP). IEEE; 2016. p. 276–280.

[78] Hsu S, Huang C, Chuang C. Vehicle detection using simplified fast R-CNN. In: 2018 International Workshop on Advanced Image Technology (IWAIT). IEEE; 2018. p. 1–3.

[79] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Conference on Computer Vision and Pattern Recognition (CVPR). vol. 1. IEEE; 2005. p. 886–893.

[80] Viola P, Jones MJ, Snow D. Detecting pedestrians using patterns of motion and appearance. International Journal of Computer Vision. 2005;63(2):153–161.

[81] Wang X, Han T, Yan S. An HOG-LBP human detector with partial occlusion handling. In: 2009 IEEE 12th international conference on computer vision. IEEE; 2009. p. 32–39.

[82] Zhang L, Lin L, Liang X, He K. Is faster r-cnn doing well for pedestrian detection? In: European conference on computer vision. Springer; 2016. p. 443–457.

[83] Brazil G, Yin X, Liu X. Illuminating pedestrians via simultaneous detection & segmentation. In: Proceedings of the IEEE International Conference on Computer Vision; 2017. p. 4950–4959.

[84] Lan W, Dang J, Wang Y, Wang S. Pedestrian Detection Based on YOLO Network Model. In: 2018 IEEE International Conference on Mechatronics and Automation (ICMA). IEEE; 2018. p. 1547–1551.

[85] Brunetti A, Buongiorno D, Trotta GF, Bevilacqua V. Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. Neurocomputing. 2018;300:17–33.

[86] Ahmed S, Huda MN, Rajbhandari S, Saha C, Elshaw M, Kanarachos S. Pedestrian and Cyclist Detection and Intent Estimation for Autonomous Vehicles: A Survey. Applied Sciences. 2019;9(11):2335.

[87] Chen Q, Shi Z, Zou Z. Robust and real-time traffic light recognition based on hierarchical vision architecture. In: 2014 7th International Congress on Image and Signal Processing. IEEE; 2014. p. 114–119.

[88] Cai Z, Li Y, Gu M. Real-time recognition system of traffic light in urban environment. In: 2012 IEEE Symposium on Computational Intelligence for Security and Defence Applications. IEEE; 2012. p. 1–6.

[89] Omachi M, Omachi S. Traffic light detection with color and edge information. In: 2009 2nd IEEE International Conference on Computer Science and Information Technology. IEEE; 2009. p. 284–287.

[90] Zhou Y, Chen Z, Huang X. A system-on-chip FPGA design for real-time traffic signal recognition system. In: 2016 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE; 2016. p. 1778–1781.

[91] John V, Yoneda K, Liu Z, Mita S. Saliency map generation by the convolutional neural network for real-time traffic light detection using template matching. IEEE transactions on computational imaging. 2015;1(3):159–173.

[92] John V, Yoneda K, Qi B, Liu Z, Mita S. Traffic light recognition in varying illumination using deep learning and saliency map. In: 17th International IEEE Conference on Intelligent Transportation Systems (ITSC). IEEE; 2014. p. 2286–2291.

[93] Gong J, Jiang Y, Xiong G, Guan C, Tao G, Chen H. The recognition and tracking of traffic lights based on color segmentation and camshift for intelligent vehicles. In: 2010 IEEE Intelligent Vehicles Symposium. Ieee; 2010. p. 431–435.

[94] Shi Z, Zou Z, Zhang C. Real-time traffic light detection with adaptive background suppression filter. IEEE Transactions on Intelligent Transportation Systems. 2015;17(3):690–700.

[95] Almeida T, Vasconcelos N, Benicasa A, Macedo H. Fuzzy model applied to the recognition of traffic lights signals. In: 2016 8th Euro American Conference on Telematics and Information Systems (EATIS). IEEE; 2016. p. 1–4.

[96] Lee G, Park BK. Traffic light recognition using deep neural networks. In: 2017 IEEE international conference on consumer electronics (ICCE). IEEE; 2017. p. 277–278.

[97] Jensen MB, Nasrollahi K, Moeslund TB. Evaluating state-of-the-art object detector on challenging traffic light data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2017. p. 9–15.

[98] Weber M, Wolf P, Zöllner JM. DeepTLR: A single deep convolutional network for detection and classification of traffic lights. In: 2016 IEEE Intelligent Vehicles Symposium (IV). IEEE; 2016. p. 342–348.

[99] Pon A, Adrienko O, Harakeh A, Waslander SL. A hierarchical deep architecture and mini-batch selection method for joint traffic sign and

light detection. In: 2018 15th Conference on Computer and Robot Vision (CRV). IEEE; 2018. p. 102–109.

[100] Beauchemin SS, Bauer MA, Kowsari T, Cho J. Portable and scalable vision-based vehicular instrumentation for the analysis of driver intentionality. IEEE Transactions on Instrumentation and Measurement. 2011;61(2):391–401.

[101] Kowsari T, Beauchemin SS, Bauer MA, Laurendeau D, Teasdale N. Multi-depth cross-calibration of remote eye gaze trackers and stereoscopic scene systems. In: 2014 IEEE Intelligent Vehicles Symposium Proceedings. IEEE; 2014. p. 1245–1250.

[102] Takagi K, Kawanaka H, Bhuiyan S, Oguri K. Estimation of a three-dimensional gaze point and the gaze target from the road images. In: 2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC). IEEE; 2011. p. 526–531.

[103] Zabihi SM, Beauchemin SS, De Medeiros EAM, Bauer MA. Frame-rate vehicle detection within the attentional visual area of drivers. In: 2014 IEEE Intelligent Vehicles Symposium Proceedings. IEEE; 2014. p. 146–150.

[104] Lin H. SVM;. `http://www.work.caltech.edu/~htlin/program/libsvm/`.

[105] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–778.

[106] Net I. ImageNet;. `http://www.image-net.org`.

[107] Ji H, Gao Z, Mei T, Li Y. Improved Faster R-CNN With Multiscale Feature Fusion and Homography Augmentation for Vehicle Detection in Remote Sensing Images. IEEE Geoscience and Remote Sensing Letters. 2019;.

[108] Hoang Ngan Le T, Zheng Y, Zhu C, Luu K, Savvides M. Multiple scale faster-rcnn approach to driver's cell-phone usage and hands on steering wheel detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2016. p. 46–53.

[109] Hu GX, Yang Z, Hu L, Huang L, Han JM. Small Object Detection with Multiscale Features. International Journal of Digital Multimedia Broadcasting. 2018;2018.

[110] Everingham M, Eslami A, Van Gool L, Williams C, Winn J, Zisserman A. The pascal visual object classes challenge: A retrospective. International journal of computer vision. 2015;111(1):98–136.

Chapter 4

Visual Driver Gaze Approximation

This Chapter is a reformatted version of the following article:

M. Shirpour, S.S. Beauchemin, M.A. Bauer, *A probabilistic model for visual driver gaze approximation from head pose estimation, accept In IEEE 3rd Connected and Automated Vehicles Symposium (CAVS), 2020.*

The direction of a vehicle driver's visual attention plays an essential role in the research on Advanced Driving Assistance Systems (ADAS) and autonomous vehicles. How a driver monitors the surrounding environment is at least partially descriptive of the driver's situation awareness. While driver gaze is not explicitly related to head pose due to the interplay between head and eye movements, it may still provide an approximation of the visual attention that is sufficiently accurate for many applications. In this research, we propose a probabilistic method for describing the visual attention of drivers. This method applies a *Gaussian Process Regression (GPR)* technique that estimates the probability of the driver gaze direction, given head pose. We evaluate our model on real data collected during drives with an experimental vehicle in urban and suburban areas. Our experimental result illustrates that 82.5% of drivers' gaze lies within the 95% confidence interval predicted by our

framework.

4.1 Introduction

Over the last few years, Advanced Driver Assistance Systems (ADAS) have been shown to significantly reduce the number of vehicle accidents. According to the *National Highway Traffic Safety Administration* (NHTSA), driver errors contribute to 94% of road collisions [1]. Evidence shows that a large number of accidents are due to driver distraction, drivers whose attention is deflected away from the driving task for more than two seconds at a time, and so on. Hence, real-time monitoring of a driver's visual field of attention is at the core of future safety systems that will be capable of further reducing the number of accidents. In recent years the driver gaze has been studied both in driving simulators and in real driving conditions.

In this work, the driver's visual attention is investigated by analyzing the head pose in the reference frame of the forward stereo system located on the roof of the experimental vehicle. This approach is often described as looking-out, which was coined by [2], [3]. In this contribution, we present a new approach that applies the Gaussian Process Regression (GPR) model for estimating the gaze direction and consequently, the object of visual attention of drivers. The rest of the contribution is structured as follows: First, a summary of related work is given in Section 4.2, followed by a description of the instrumented vehicle and data collection process in Section 4.3. Section 4.4 describes our proposed method. In Section 4.5, we present the experimental results along with a critical analysis. We provide a short conclusion and future research directions in Section 4.6.

Figure 4.1: **(left):** *Forward stereoscopic vision system on rooftop,* **(center):** *3D infra-red gaze tracker,* **(right):** *The faceLAB system interface from SeeingMachines*

4.2 Related Works

We provide a summary of the literature focused on estimating the gaze direction of drivers in order to identify objects of driver visual attention. There are two types of approaches: vision-based methods, and learning-based methods. Vision-based gaze zone estimation falls into one of two categories: gaze estimation inside the vehicle space, and gaze estimation located outside the vehicle, in the the reference frame of the stereo system. Learning-based methods for gaze estimation comprise traditional machine learning and deep learning methods.

Methods presented in [4], [5], [6], [7], [8] estimate the gaze zone inside vehicle space. Doshi *et al.* [4] approximated the frequency of eye gaze location on omnidirectional images. They observed that a classifier based upon head movement has notably more predictive power than one based on the eye's gaze 3 seconds before a lane change, but not 2 seconds before it. Tawari *et al.* [5] proposed a framework to determine if the driver is looking inside or outside the vehicle. They considered coarse eye pose and combined the salience of the scene with the object the driver is focused on at a particular time. Ahlstrom *et al.* [6] employed a dynamic region method in which a 3D model divides

the vehicle into different parts such as windshield, speedometer, rear-view mirrors, dashboard, etc. Another dynamic region-based method based on the work of [7], [8] determines whether the driver gaze is on-road or off-road. They evaluated their approach in stationary vehicles.

Authors in [9], [10], [11], [3], and [2] study driver gaze in relation with object instances in the image space. In particular, Martin *et al.* [9] introduced an architecture that learns to allocate a probability to every object in the view based upon their likelihood of being the driver's object of fixation. Schwehr *et al.* [11] studied various types of gaze trackers calibrated against other sensors in order to evaluate the robustness of techniques that associate a scene object with the gaze of drivers. Kowsari *et al.* [2] introduced a cross-calibration technique to transform the driver gaze from the reference frame of a remote gaze tracker onto the reference frame of a forward stereoscopic vision system. Zabihi *et al.* [3] defined a framework that uses the 3D Point of Gaze (PoG) and Line of Gaze (LoG) of the driver in absolute 3D coordinates. They consider the attentional visual area of the driver as the cone originating from the eye position along the LoG.

Machine learning methods have also been used for similar purposes. For instance, Bar *et al.* [12] used a decision tree to learn how the driver's gaze engages to important objects in a given situation, for the purpose of estimating an awareness confidence level. Fridman *et al.* [13] proposed a method to find facial regions with a combination of histogram of oriented gradients (HOG) and SVM classifiers, and then classify the feature vectors with respect to gaze zones by way of a random forest classifier. Lundgren *et al.* [14] proposed a Bayesian filtering method that models the visual focus of attention in the absence of gaze observations. They estimate the probability the driver is looking in various zones by observing the driver behavior in terms of head rotations. Jha *et*

al. [15] proposed a Gaussian process that estimates the probability of specific points on the windshield, where the driver could be looking at.

Recently, attempts at applying deep learning methods for gaze estimation have been made. These methods need large datasets with annotated gaze labels. Alletto *et al.* [16] provided the Dr(eye)ve dataset, where multiple researchers contributed to the data collection. Jha *et al.* [17] proposed a method based upon deep learning networks for estimation of driver's attention. They gradually up-sample the resolution of the gaze area, which improves the accuracy and the resolution of the prediction. Vora *et al.* [18] used a similar technique to categorize the driver gaze into seven non-overlapping zones.

Our proposed method employs a *Gaussian Process Regression (GPR)* technique to estimate the driver's visual attention, expressed in the reference frame of the forward stereo system, located on the roof of the vehicle.

4.3 Vehicle Equipment and Data Collection

Our vehicle is equipped with a infrared remote gaze tracker. This system [19] uses a pair of IR-sensitive stereo cameras mounted on the dashboard. The system has been used in several experiments for various purposes [20], [3], [2], [21]. The remote gaze tracker computes several driver features, including head position and orientation, left and right gaze Euler angles, and left and right eye center locations within the coordinate system of the tracker. A stereo system located on the vehicle's roof captures the frontal environment. Figure 4.1 depicts the configuration of the experimental vehicle [22].

The experimental vehicle is equipped with an On Board Diagnostic system (OBD-II) that allows sensors to report on current vehicular status. It constitutes the interface through which odometry is made available in real-time.

Figure 4.2: *The on-board software system displays image plane of the forward stereo system, dynamic vehicle features, and eye tracker data.*

Since 2008, the Controller Area Network bus protocol (CANbus) has become mandatory for OBD-II. This standardization simplifies the real-time capture of vehicular data. Several critical elements of vehicular dynamics are extracted from the CANbus as driving indicators, such as vehicle speed, accelerator and brake pedal pressures, steering wheel angle, and state of turn signals [22]. Sixteen driving sequences were recorded by our experimental vehicle with test drivers on a pre-determined 28.5km course around the city of London, ON, Canada (see Table 4.1). A single driving sequence represents a driving time of approximately one hour.

Table 4.1: Description of Driving Sequences Used For Experiments.

Sequence	Capture Date	Weather	Gender
1	2012-08-24	29 °C Sunny	M
2	2012-08-24	31 °C Sunny	M
3	2012-08-30	23 °C Sunny	F
4	2012-08-31	24 °C Sunny	M
5	2012-09-05	27 °C Partially Cloudy	F
6	2012-09-10	21 °C Partially Cloudy	F
7	2012-09-12	21 °C Sunny	F
8	2012-09-12	27 °C Sunny	M
9	2012-09-17	24 °C Partially Cloudy	F
10	2012-09-19	8 °C Sunny	M
11	2012-09-19	12 °C Sunny	F
12	2012-09-21	18 °C Partially Cloudy	F
13	2012-09-21	19 °C Partially Cloudy	M
14	2012-09-24	7 °C Sunny	F
15	2012-09-24	13 °C Partially Cloudy	F
16	2012-09-28	14 °C Partially Cloudy	M

4.4 Methodology

Our interest is to build a stochastic model defining the area visual attention of drivers projected onto the imaging plane of the forward sterescopic system. This approach requires the cross-calibration of the stereo system with the remote gaze tracker. Section 4.4.1 describes the calibration method used to get the point of gaze and Section 4.4.2 discusses a method based upon the Gaussian Process Regression to predict visual driver gaze.

4.4.1 From Calibration to Projection of PoGs Onto the Forward Stereo System

The calibration process between eye tracker and stereo system is an essential step towards building a useful PoG representation. We used a framework proposed in our laboratory to calibrate the systems and to project the PoGs onto the imaging plane of the stereo system (See Figure 4.2). The framework is defined in the following steps [2]:

- Description of Calibration Procedure:
  - *Extraction of Salient Points:* Selection of calibration points that are provided by OpenCV from Hessian salient points.
  - *Depth Estimation:* The driver fixates eyes on preselected salient points for which the depth estimate, the gaze vector, the eye location, and the 3D position of those points are available.
  - *Rotation Matrix and Translation Vector Estimates:* The objective consists of computing an estimate of the rigid body transformation that exists between the stereo system and the eye tracker [5]. These estimates are known as the Extrinsic Parameters of the paired systems.
- *Projection of the Gaze on the Scene Image:* The LoG is projected onto the stereo system's imaging plane. The PoG is identified onto the image region, as long as the line intersects with a valid depth [5].

4.4.2 Gaussian Process Regression

For reasons mentioned above, our interest is to estimate the image area onto the imaging plane of the stereo system that elicits a visual response (fixation) from the driver, using head pose as an approximation. To some extent, driver head pose is indicative of driver attention orientation. This method differs from those that propose to find the exact gaze by directly observing the eyes.

We aim to build a visual heat-map that indicates the stereo image area the driver's visual attention is most likely turned to.

We provide a short overview of GPR for the purpose of implementation [23]. Gaussian Process Regression is a non-parametric approach that specifies a prior probability distribution over a latent function f, where f is a mapping from the input $\mathcal{X}$ to output $\mathcal{Y}$. The marginal distribution $P(f(x_1), f(x_2), ...f(x_n))$ is a multivariate normal distribution, where $x_i\{i :\in 1, 2...n\}$ is an input vector.

The statistical parameters of Gaussian Process are defined by mean and covariance functions $m(x) = E[f(x)]$ and $k(x, x') = E[(f(x)-E[f(x)])(f(x')-E[f(x')])^T]$ respectively, where $x, x' \in X$ are input vectors and E represents expectation value. The GP is defined as follows[24]:

$$y = \mathcal{GP}(m(x), k(x, x')) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2) \tag{4.1}$$

where x and y denote the input vector and noisy observation respectively. We refer the reader to [23] and [24] for more details.

To map a feature vector $X \in \mathbb{R}^6$ containing each of the six head pose parameters to the coordinate output $Y \in \mathbb{R}^2$, we define a training set $D = \{(X_i, Y_i) \mid i = 1, 2, ..., n\}$, where X_i represents an input vector of dimension 6, Y_i represents the output, and n is the number of observations. In order

to make the model predict unseen data X' from training data D, we need to build a function f to predict for all inputs:

$$Y' = f(X') = p(Y', X' \mid D) \tag{4.2}$$

where p is a *posterior* distribution for the training set. The joint distribution of y and y' is:

$$\begin{bmatrix} y \\ y' \end{bmatrix} \sim \mathcal{N}\Big(0, \begin{bmatrix} K(X,X) + \sigma_n^2 & K(X,X') \\ K(X',X) & K(X',X') \end{bmatrix}\Big) \tag{4.3}$$

where X denotes training set and X' denotes testing set. If we assume n data from training set D and m test data, then $K(X,X)$ is a $n \times n$ covariance matrix of input training data, which is a symmetric positive definite matrix. The matrix element $K_{ij} = K(x_i, x_j)$ represents the correlation between x_i and x_j. $K(X,X') = K(X',X)^T$ is an $n \times m$ covariance matrix between training X and testing data X'. $K(X',X')$ is a $m \times m$ covariance matrix of testing data X'. For instance, $K(X,X')$ is given by:

$$\sigma_f^2 \exp\left(\frac{-\mid X - X' \mid^2}{2l^2}\right) \tag{4.4}$$

where σ_f^2 denotes the standard deviation that controls the degree of correlation. The covariance function $K(x_i, x_j)$ reaches the maximum σ_f^2 when the inputs satisfy $x_i = x_j$. In other words, it occurs when $f(x_i)$ and$f(x_j)$ are completely correlated. If x_i and x_j are distant from each other, we obtain $K(x_i, x_j) \approx 0$. l is the length-scale feature, that indicates the correlation level related to differences between inputs.

Figure 4.3: *Various PoGs projected onto the forward stereo scene system of the vehicle, with less than 3-pixel movement in the last 15 frames (1/2 second)* .

4.5 Experimental Evaluation

Here, we employed the driving sequences captured with our experimental vehicle [22], as described in Section 4.3. We applied our proposed method to estimate the most probable areas within the imaging plane of the stereoscopic system that are being gazed at by the driver, given head pose parameters. These probabilities are depicted with confidence intervals [25], [24]. For evaluation purposes, the dataset was divided into a training, a validation, and a test data set.

Table 4.2: Gaze Estimation Results Per Confidence Interval

Sequence	**Confidence Interval**	**Accuracy**	**Sequence**	**Confidence Interval**	**Accuracy**
Subject 2	50% CI	60.4%	Subject 3	50% CI	54.9%
	75% CI	75.1%		75% CI	65.6%
	95% CI	87.6%		95% CI	78.1%
Subject 4	50% CI	53.8%	Subject 5	50% CI	58.7%
	75% CI	62.5%		75% CI	67.9%
	95% CI	80.3%		95% CI	75.1%
Subject 6	50% CI	63.3%	Subject 7	50% CI	51.1%
	75% CI	78.2%		75% CI	62.5%
	95% CI	91.6%		95% CI	76.9%
Subject 8	50% CI	65.3%	Subject 10	50% CI	55.2%
	75% CI	81.1%		75% CI	73.9%
	95% CI	93.6%		95% CI	87.2%
Subject 11	50% CI	53.1%	Subject 12	50% CI	58.5%
	75% CI	63.9%		75% CI	76.4%
	95% CI	77.6%		95% CI	84.1%
Subject 13	50% CI	50.9%	Subject 15	50% CI	55.1%
	75% CI	65.6%		75% CI	66.8%
	95% CI	80.3%		95% CI	78.9%
Average	50% CI	56.6%			
	75% CI	69.9%			
	95% CI	82.5%			

Ideally, an in-vehicle safety system must rely on sensors and cameras to track the behavioral characteristics of drivers. Therefore, it is important to have reliable algorithms estimating the driver head pose. Our gaze tracker performed the head pose estimation and provided a measure of its quality. This quality metric varied from 0 to 2, and we considered the head pose to be reliable when this metric had a value of 1 or greater.

In order to perform an adequate analysis of the accuracy of our method, we extracted the PoGs projected onto the forward stereo system of the vehicle, in the preceding 15 consecutive frames ($\frac{1}{2}$ second) for which the PoGs vary less than 3 pixel positions (see Figure 4.3), and considered those PoGs as target gazes.

For each head pose we considered accurate, we estimated the confidence intervals using the GPR model for their positions. We evaluated our method with confidence intervals of 50%, 75%, and 95%. It should be noted that the 50% and 75% confidence intervals are subsets of the 95% interval. The size of confidence intervals relies on the uncertainty of the approach, which is a function of the estimated head pose parameters. We proceeded to calculate the proportion of target gazes that found themselves within the image regions corresponding to our confidence intervals.

Table 4.2 illustrates the proportion of correctly estimated target gazes for each driver, including the 50%, 75%, and 95% confidence intervals. We observe that the average proportion of the target gaze for the confidence interval of 95% is 82.5%. It can be seen that the majority of the points of gaze are located inside this confidence region. Furthermore, 69.9% of the target gazes are included in the 75% confidence region. As it is expected for the most accurate region (50%), this confidence interval included only 56.6% of the gazes from the test samples. As mentioned before, while the gaze direction and the

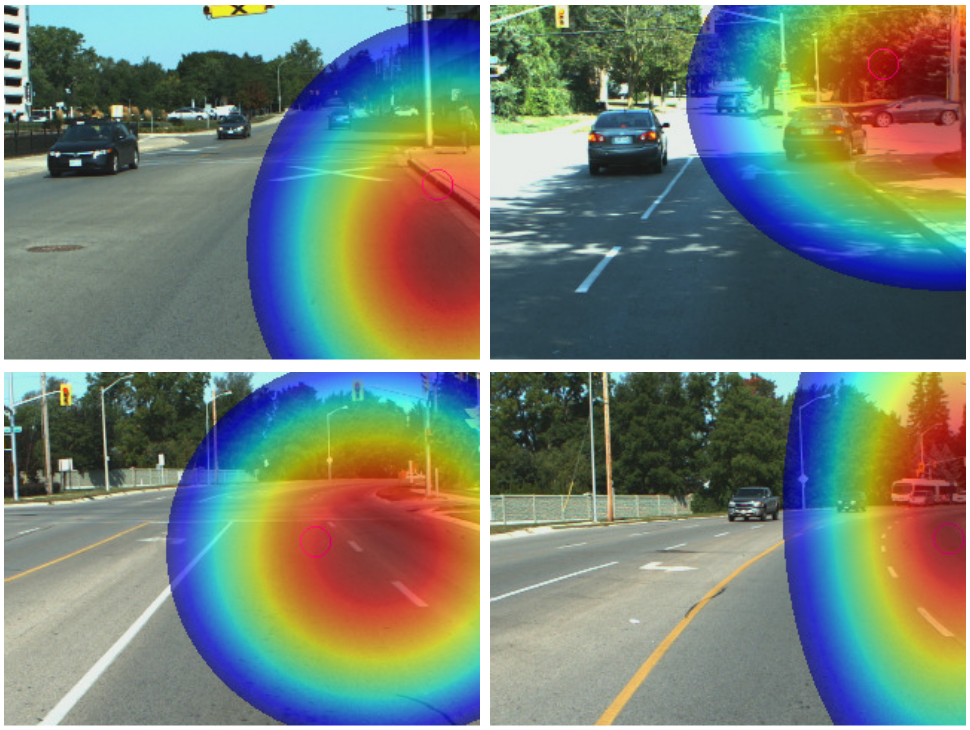


Figure 4.4: *Output samples for which the PoG falls within the confidence regions*

head pose are not explicitly correlated, the proposed model is nonetheless able to provide a coarse, yet mostly correct estimation of gaze localization.

Figure 4.4 presents output samples of the probabilistic model for which the PoG is inside a confidence region. The purple circle is the target gaze location (real gaze), and the heatmap displays the confidence regions, where the color red indicates a high probability of finding the gaze. The blue ellipses represent higher variances. Notice that the sizes of the ellipses do vary, as they are altered according to the uncertainty within the relationship between head pose and gaze. A smaller circle indicates higher confidence in the data, whereas a larger circle indicates greater uncertainty in the driver gaze estimates.

Figure 4.5 presents output samples for which the PoG is outside any of

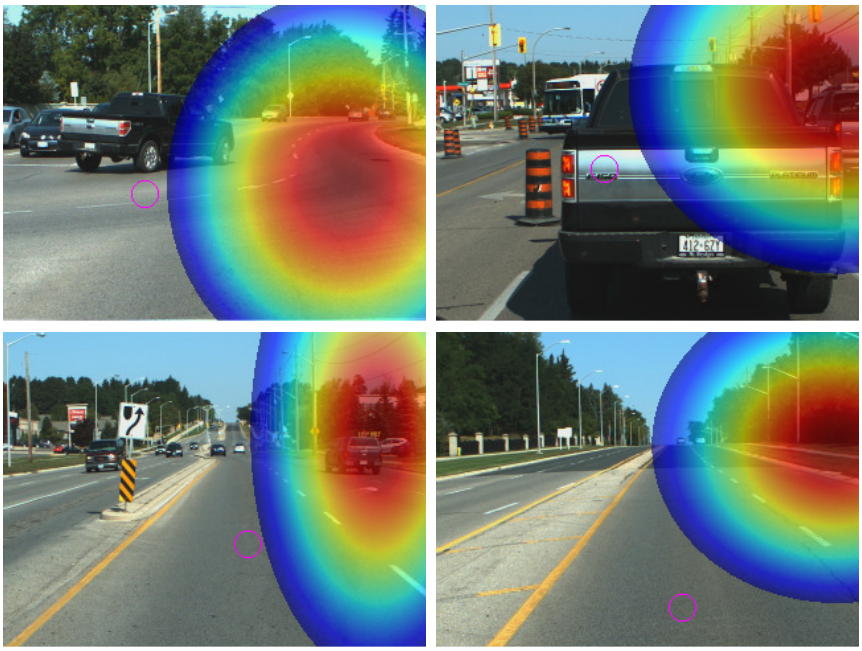


Figure 4.5: *Output samples for which the PoG falls outside of the confidence regions*

the confidence regions. There is a trade-off between the size of the confidence regions and the quality of the approximation that can be obtained from our model. When the area of the confidence regions is large, it signifies that the estimation is uncertain. We expect that the accuracy of our model for visual driver gaze estimation will be improved as the number of drivers increase in the data set.

4.6 Conclusion

We presented a new stochastic method for the detection of gaze areas, given driver head pose estimates. Rather than estimating the gaze precisely, which relies on the driver's visual cognitive tasks, the method computes a probabilistic visual attention map describing the probability of finding the

actual gaze over the imaging plane of the stereo system.

We calculated confidence regions onto the forward stereo system of the vehicle, as they express the uncertainty within the relationship between gaze and head pose. Our approach is capable of estimating driver gaze without explicitly tracking the eyes of drivers, thus simplifying the hardware requirements for applications in which a coarse estimate of gaze suffices, such as with certain applications in traffic safety systems.

Our future work includes object and event identification located within the surroundings of the vehicle that elicit driver visual attention. Our prior research has established that driver gaze estimation is important for driver maneuver prediction [21]. However, of equal importance is the ability to identify the object of driver visual attention on a real-time basis. We believe that this ability plays a crucial role in maneuver prediction because the driver perceives and focuses on environmental features moments before performing a maneuver. To reach this goal, we will be producing a collection of annotated datasets in which the static (road signs, traffic lights) and dynamic (pedestrians, other vehicles) actors within the scene are labelled. Also, the predictive model for driver gaze direction could be used as an input feature in the driving maneuver prediction model.

Bibliography

[1] Administration NHTS, et al. Traffic safety facts: Research note. DOT HS. 2016;812:318.

[2] Kowsari T, Beauchemin SS, Bauer MA, Laurendeau D, Teasdale N. Multi-depth cross-calibration of remote eye gaze trackers and stereoscopic scene

systems. In: 2014 IEEE Intelligent Vehicles Symposium Proceedings. IEEE; 2014. p. 1245–1250.

[3] Zabihi S, Beauchemin SS, De Medeiros E, Bauer MA. Frame-rate vehicle detection within the attentional visual area of drivers. In: 2014 IEEE Intelligent Vehicles Symposium Proceedings. IEEE; 2014. p. 146–150.

[4] Doshi A, Trivedi M. Investigating the relationships between gaze patterns, dynamic vehicle surround analysis, and driver intentions. In: 2009 IEEE Intelligent Vehicles Symposium. IEEE; 2009. p. 887–892.

[5] Tawari A, Møgelmose A, Martin S, Moeslund TB, Trivedi MM. Attention estimation by simultaneous analysis of viewer and view. In: 17th International IEEE Conference on Intelligent Transportation Systems (ITSC). IEEE; 2014. p. 1381–1387.

[6] Ahlstrom C, Kircher K, Kircher A. A gaze-based driver distraction warning system and its effect on visual behavior. IEEE Transactions on Intelligent Transportation Systems. 2013;14(2):965–973.

[7] Vicente F, Huang Z, Xiong X, De la Torre F, Zhang W, Levi D. Driver gaze tracking and eyes off the road detection system. IEEE Transactions on Intelligent Transportation Systems. 2015;16(4):2014–2027.

[8] Tawari A, Chen KH, Trivedi MM. Where is the driver looking: Analysis of head, eye and iris for robust gaze zone estimation. In: 17th International IEEE Conference on Intelligent Transportation Systems (ITSC). IEEE; 2014. p. 988–994.

[9] Martin S, Tawari A. Object of fixation estimation by joint analysis of gaze

and object dynamics. In: 2018 IEEE Intelligent Vehicles Symposium (IV). IEEE; 2018. p. 2042–2047.

[10] Schwehr J, Knaust M, Willert V. How to evaluate object-of-fixation detection. In: 2019 IEEE Intelligent Vehicles Symposium (IV). IEEE; 2019. p. 570–577.

[11] Schwehr J, Willert V. Driver's gaze prediction in dynamic automotive scenes. In: 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC). IEEE; 2017. p. 1–8.

[12] Bär T, Linke D, Nienhüser D, Zöllner JM. Seen and missed traffic objects: A traffic object-specific awareness estimation. In: 2013 IEEE Intelligent Vehicles Symposium (IV). IEEE; 2013. p. 31–36.

[13] Fridman L, Langhans P, Lee J, Reimer B. Driver gaze region estimation without use of eye movement. IEEE Intelligent Systems. 2016;31(3):49–56.

[14] Lundgren M, Hammarstrand L, McKelvey T. Driver-gaze zone estimation using Bayesian filtering and Gaussian processes. IEEE transactions on intelligent transportation systems. 2016;17(10):2739–2750.

[15] Jha S, Busso C. Probabilistic estimation of the driver's gaze from head orientation and position. In: 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC). IEEE; 2017. p. 1–6.

[16] Alletto S, Palazzi A, Solera F, Calderara S, Cucchiara R. Dr (eye) ve: a dataset for attention-based tasks with applications to autonomous and assisted driving. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2016. p. 54–60.

[17] Jha S, Busso C. Probabilistic estimation of the gaze region of the driver using dense classification. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC). IEEE; 2018. p. 697–702.

[18] Vora S, Rangesh A, Trivedi MM. On generalizing driver gaze zone estimation using convolutional neural networks. In: 2017 IEEE Intelligent Vehicles Symposium (IV). IEEE; 2017. p. 849–854.

[19] Online. Available: http://www.seeingmachines.com;.

[20] Fletcher L, Zelinsky A. Driver inattention detection based on eye gaze—Road event correlation. The international journal of robotics research. 2009;28(6):774–801.

[21] Khairdoost N, Shirpour M, Bauer MA, Beauchemin SS. Real-Time Maneuver Prediction Using LSTM. IEEE Transactions on Intelligent Vehicles. 2020;.

[22] Beauchemin SS, Bauer MA, Kowsari T, Cho J. Portable and scalable vision-based vehicular instrumentation for the analysis of driver intentionality. IEEE Transactions on Instrumentation and Measurement. 2011;61(2):391–401.

[23] Rasmussen CE. Gaussian processes in machine learning. In: Summer School on Machine Learning. Springer; 2003. p. 63–71.

[24] Williams CK, Rasmussen CE. Gaussian processes for machine learning. vol. 2. MIT press Cambridge, MA; 2006.

[25] Ebden M. Gaussian processes: A quick introduction. arXiv preprint arXiv:150502965. 2015;.

Chapter 5

Driver's Eye Fixation

This Chapter is a reformatted version of the following article: M. Shirpour, S.S. Beauchemin, and M.A. Bauer, *Driver's Eye Fixation Prediction by Deep Neural Network, accepted in VISAPP 2021 Conference, Vienna, Austria, 2021.*

The driving environment is a complex dynamic scene in which a driver's eye fixation interacts with traffic scene objects. Prediction of a driver's eye fixation plays a crucial role in Advanced Driving Assistance Systems (ADAS) and autonomous vehicles. However, currently, no computational framework has been introduced to combine the bottom-up saliency map with the driver's head pose and gaze direction to estimate a driver's eye fixation. In this work, we first propose convolution neural networks to predict the potential saliency regions in the driving environment, and then use the probability of the driver gaze direction, given head pose as a top-down factor. We evaluate our model on real data gathered during drives in an urban and suburban environment with an experimental vehicle. Our analyses show promising results.

5.1 Introduction

Recently, visual driver attention has become a noticeable element of intelligent Advanced Driver Assistance Systems (i-ADAS). Based on the World Health Organization (WHO) studies, approximately 1.35 million fatalities and anywhere between 20 to 50 million injuries occur every year on the roads. The WHO predicts that road traffic accidents will rise to become the fifth primary reason for mortality in 2030 [1]. Evidence has shown that a considerable number of accidents are due to distraction.

Driver monitoring research has been carried out for years in various research fields, from science to engineering, to protect the driver from dangerous situations. The driver's eye fixation plays a crucial role in the research on Driver Safety System and Enhanced Driver Awareness (EDA) systems to alert drivers on incoming traffic conditions and warn them appropriately. Some driver monitoring systems use head and eye location to evaluate the driver's gaze-direction and gaze-zone [2, 3]. Their purpose is to estimate the driver's intent and predict the driver's maneuvers a few seconds before they occur [4, 5]. Their results illustrate a strong connection between a driver's visual attention and action.

The driver's eye generally fixates on parts of the driving environment that depend on a number of objective and subjective factors that are based on two classes of attentional mechanisms: bottom-up and top-down. Bottom-up mechanisms consider features obtained from the driving scene such as traffic signs, vehicles, traffic lights, and so on. In contrast, top-down mechanisms are driven by internal factors such as a driver's experience or intent [6]. Saliency maps identify essential regions in the scene. In a driving context, top-down factors significantly contribute to the estimation of traffic saliency maps, which

in turn provide an insight as to what a driver's gaze may be fixated on while driving.

In this study, we focus on developing a framework to predict the driver's eye fixation onto the forward stereo system's imaging plane located on the instrumented vehicle's rooftop. This contribution is structured as follows: an overview on the current literature in the field of saliency regions is provided in Section 5.2, followed by a description of the RoadLAB vehicle instrumentation and data collection processes in Section 5.3. Section 5.4 describes our proposed method. In Section 5.5, we present and evaluate the experimental results. We provide a conclusion and areas for further research in Section 5.6.

5.2 Related Works

Traffic saliency methods focus on highlighting salient regions or areas in a given environment. This is an active area in the fields of computer vision and intelligent vehicle systems. We provide a summary of the literature that brings the essential concepts of visual attention and salient regions applied to driving environments.

Saliency, as it relates to visual attention, refers to areas of fixation humans or drivers would concentrate on at a first glance. The modern history of visual saliency goes back to the works of Itti [7]. They considered low-level features, namely intensity, color, and orientation at multiple scales extracted from images, and then normalized and combined them with linear and non-linear methods to estimate a saliency map. Harel *et al.* [8] suggested a saliency method with Graph-Based Visual Saliency (GBVS). They defined the equilibrium distribution of Markov chains from low-level features and then combined them to obtain the final saliency map. Schauerte *et al.* [9] pro-

posed quaternion-based spectral saliency methods that apply the integration of quaternion DCT and FFT-based to estimate spectral saliency for predicting human eye fixations. Li *et al.* [10] proposed a bottom-up factor for visual saliency detection, which is considered a scale-space analysis of amplitude spectra of images. They convolved image spectra with properly scaled low-pass Gaussian kernels to obtain saliency maps. Some research demonstrated that a driver's attention was mainly focused on the vanishing points present in the scene [11, 6] . Deng *et al.* [6] applied the road vanishing point as guidance for the for traffic saliency detection. Subsequently, they proposed a model based on a random forest to predict a driver's eye fixation according to low-level features (color, orientation, intensity) and vanishing points [12]. Details on low-level features for non-deep learning approaches are provided in [13].

Deep learning-based models brought a paradigm shift in computer vision research. Deep-learning methods commonly perform better when compared with classical learning methods. Vig *et al.* [14] introduced one of the early networks that performed large scale searches over different model configurations to predict saliency regions. Liu *et al.* [15] proposed Multi-resolution Convolutional Neural Networks (Mr-CNN) to learn two types of visual features from images simultaneously. The Mr-CNNs were trained to classify image regions for saliency at different scales. Their model used top-down feature factors learned in upper-level layers, and bottom-up features gathered by a combination of information over various resolutions. They then integrated bottom-up and top-down features with a logistic regression layer that predicted eye fixations. Kummerer *et al.* [16] presented the DeepGaze model that applied the VGG-19 deep neural network for feature extraction, where features for saliency prediction were extracted without any additional fine-tuning. Huang *et al.* [17] proposed a deep neural network (DNN) obtained from concatenating

two pathways: the first path considered a large scale image to extract coarse features, and the second path considered a smaller image scale to extract fine ones. This model and similar ones are suitable to extract features at various scales. Wang *et al.* [18] proposed a framework that extracted features from deep coarse-layers with global information and shallow fine layers with local information that captured hierarchical saliency features to predict eye fixation. Subsequently, they designed the Attentive Saliency Network (ASNet) from the fixations to detect salient objects [19].

In the driving context, Palazzi *et al.* [20] proposed a model based on a multi-branch deep neural network on the DR(eye)VE dataset, which consisted of three-stream convolutional networks for color, motion, and semantics. Each stream possessed its parameter set, and the final map aggregated a three-stream prediction. Also, Tawari *et al.* [21] estimated drivers' visual attention with the use of a Bayesian Network model and detected the saliency region with a fully convolutional neural network. Deng *et al.* [22] proposed a model to detect driver's eye fixations based on a convolutional-deconvolutional neural network (CDNN). Their framework could predict the primary fixation location and the second saliency region in the driving context, if it existed.

This contribution aims to apply a Deep Neural Network to our natural driving sequences for the estimation of saliency maps followed by a Gaussian Process Regression (GPR) to estimate the driver's confidence region for the final estimation of driver's eye fixation.

Figure 5.1: *RoadLAB configuration.* **(top):** *vehicular configuration: stereoscopic vision system on rooftop and 3D infrared eye-tracker located on the dashboard.* **(bottom):** *software systems: The on-board system displays frame sequences with depth maps, dynamic vehicle features, and eye-tracker data.*

5.3 Vehicle Instrumentation And Data Collection

5.3.1 Vehicle Configuration

Our experimental vehicle is equipped with a stereo system placed on the vehicle's roof to capture the frontal driving environment. A remote eye-gaze tracker located on the dashboard captures several features related to the driver, including head position and orientation, left and right gaze Euler angles, and left and right eye center locations within the coordinate system of the tracker. Furthermore, the On-Board Diagnostic system (OBD-II) records the current status of vehicular dynamics such as vehicle speed, brake and accelerator pedal pressure, steering wheel angle, etc. Figure 5.1 depicts the RoadLAB experimental vehicle and its software systems as described in [23].

5.3.2 Cross-Calibration Technique

The calibration process between the eye-tracker and stereo system is essential for generating a useful Point of Gaze (PoG). We applied a technique developed in our laboratory to cross-calibrate these systems and project the PoGs onto the stereo system imaging plane. Details are provided in [24].

5.3.3 Participants

Sixteen drivers participated in this experiment, including nine females and seven males. Each participant was recorded by our instrumented vehicle on a pre-determined 28.5km route within the city of London, ON, Canada. Each sequence represented a driving time of approximately one hour. Sequences

were recorded in different circumstances, including scenery (downtown, urban, suburban) and traffic conditions varying from low-traffic to high-traffic situations. They were recorded in various weather conditions (sunny, partially-cloudy, cloudy) and at various times of the day (see Table 5.1).

5.3.4 Driver Gaze-Movement Analysis

Our eye-tracker performed the gaze estimation and provided a confidence measure on its quality. This metric ranged from 0 to 3, and we considered the driver's gaze to be reliable when this metric had a value of 2 or higher. We selected the PoGs projected onto the vehicle's forward stereo system in the preceding 15 consecutive frames. The driver's POG data implemented with the Gaussian distribution (Fig.5.2) were considered the ground-truth data.

Figure 5.2: *An example of PoG and matching fixation saliency map.* **(left):** *PoGs projected onto the forward stereo system of the vehicle obtained with the preceding 15 consecutive frames.* **(right):** *The driver's point of gaze as a 2-D Gaussian distribution.*

Table 5.1: Description of RoadLab Dataset.

Seq#	Date	Weather	Gender
1	2012-08-24	29 °C Sunny	M
2	2012-08-24	31 °C Sunny	M
3	2012-08-30	23 °C Sunny	F
4	2012-08-31	24 °C Sunny	M
5	2012-09-05	27 °C Partially Cloudy	F
6	2012-09-10	21 °C Partially Cloudy	F
7	2012-09-12	21 °C Sunny	F
8	2012-09-12	27 °C Sunny	M
9	2012-09-17	24 °C Partially Cloudy	F
10	2012-09-19	8 °C Sunny	M
11	2012-09-19	12 °C Sunny	F
12	2012-09-21	18 °C Partially Cloudy	F
13	2012-09-21	19 °C Partially Cloudy	M
14	2012-09-24	7 °C Sunny	F
15	2012-09-24	13 °C Partially Cloudy	F
16	2012-09-28	14 °C Partially Cloudy	M

5.4 Driver Fixation

We proposed method to predict a driver's eye fixation in the forward stereo vision reference frame. First, we introduce a model to predict the saliency maps in the driving scene, inspired by [18]. Following this, we use a framework proposed in our laboratory to estimate the probability of driver's gaze

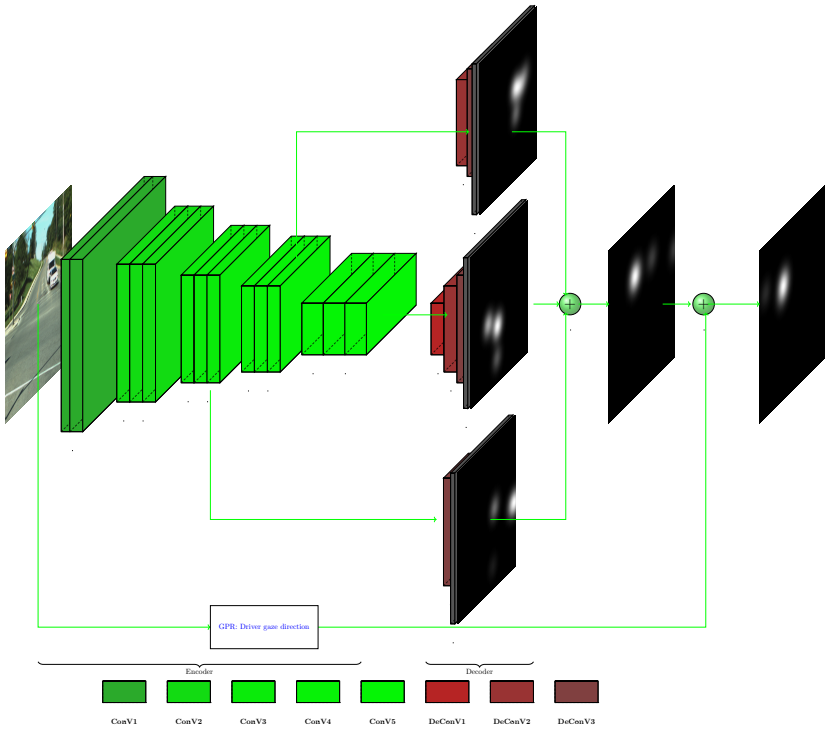


Figure 5.3: *Network configuration*

direction, as top-down information for prediction of driver's eye fixation [3].

5.4.1 Model Architecture

The network configuration selection is a fundamental step when using a neural network. There are various types of deep neural network saliency models, mainly divided into three groups: single stream, multi-stream, and skip layer networks. Our network inherits the advantage of skip layer networks capable of capturing hierarchical features. This network configuration learns multi-scale features inside the model; the low-level layers reflect primitive features such as edges, corners, etc; and the high-level layers represent meaningful information such as parts of objects in various positions. The network archi-

tecture is shown in Fig. 5.3. This architecture promotes performance via:

- the creation of multi-scale saliency features inside the network
- the preservation of high-resolution features from the encoder path

Our network encoder is based on the first five convolutional layers of VGG16 [25], used for feature extraction from input images. The dimensions of the input images are $H \times W \times 3$. The network encoder includes a stack of convolution layers that gradually learns from local to global information. The spatial feature dimensions generated from VGG16 are consequently divided by 2 until, in the last convolution layers, the dimensions reach $H/16 \times W/16$. We choose three feature maps from the encoder path generated by convolution layers $ConV3-3$, $ConV4-3$, and $ConV5-3$ to capture multi-scale saliency information. We use these three-channel feature maps with different dimensions and resolutions to obtain the final saliency prediction.

In the decoder part for each path, we apply multiple deconvolution layers to increase the spatial dimension toward getting a saliency prediction map with dimensions identical to those of the input images. For instance, the feature map in the $ConV3-3$ layer has a $H/4 \times W/4$ spatial dimension (after each convolution block, the spatial dimension size is halved). Its decoder network path includes two deconvolution layers, where the first one doubles the spatial size of feature map to $H/2 \times W/2$, while the second deconvolution increases the spatial size of the feature map to $H \times W$. Each deconvolution in these paths is followed by a Rectified Linear Unit $ReLU$ layer, which learns a nonlinear upsampling. Similarly, the other decoder path related to $ConV4-3$ and $ConV5-3$ layers has three and four deconvolution layers, respectively.

The loss function $L(S_F, S_G)$ is defined as follows:

$$L(S_F, S_G) = \frac{1}{N} \sum_{n=1}^{N} S_{G_i} log(S_{F_i}) + (1 - S_{G_i}) log(1 - S_{F_i}) \tag{5.1}$$

where N is the number of pixels, S_{G_i} is the i^{th} pixel from the ground truth driver's fixation map, and S_{F_i} is the i^{th} pixel from the predicted driver's fixation map.

5.4.2 Top-Down Information

The driver gaze is not explicitly related to the head pose due to the interaction between head and eye movements. Generally, the driver moves both the head and the eyes to obtain a fixation. In our previous research, we suggested a stochastic model for describing a driver's visual attention. This method uses a Gaussian Process Regression (GPR) approach that estimates the driver gaze direction probability, given head pose. We refer the reader to [3] for details on the confidence interval for the driver's gaze direction process.

Based on the driver's head pose information, we propose a traffic saliency maps framework, which utilizes the gaze direction as a top-down constraint. The primary part of the framework is to find top-down features according to the driver's head pose and to estimate the probability of a driver's gaze direction, which is then fused with the saliency map, as follows:

$$S_F(x, y) = w S_{CI}(x, y) + (1 - w) S_m(x, y) \tag{5.2}$$

where w is the weighting factor, $S_{CI}(x, y)$ represents the confidence interval of driver's gaze according to the head pose information, and $S_m(x, y)$ represents the saliency map model. The weight w in 5.2 is a critical parameter of the

framework, as it dictates the importance of the top-down factor in our model. To choose a correct weight, we have shown that the drivers focus most of their attention on the 95% confidence interval region estimated with the driver head pose. Since the top-down saliency area includes 80% of the information that is related to a driver's fixation within the area of the confidence interval of the driver's head pose, we hypothesized that 0.8 was a suitable value for w.

5.5 Experimental Evaluation

In this Section, we describe the training of our proposed network and evaluate its performance both qualitatively and quantitatively.

5.5.1 Qualitative Evaluation

To evaluate our model against a number of cutting-edge methods, we chose various sample frames from challenging driving environments, including difficult situations and conditions, such as traffic objects with different sizes, low contrast scenes, and multiple traffic objects. Figure 5.4 illustrates the comparison of our network against other methods, namely: Graph-based Visual Saliency (GBVS) [8], Image Signature [26], Itti [7], and Hypercomplex Fourier Transform (HFT) [10]. Results clearly demonstrate that our method highlights the drivers' fixation areas more accurately and preserves details compared to other methods. Our model displays excellent prediction of traffic objects such as traffic signs, traffic lights, pedestrians, vehicles, among others. Other models displayed difficulties when attempting to detect relevant information from the driving environments. Conversely, by way of bottom-up and top-down processes, our model accurately predicts the driver's fixation, including the

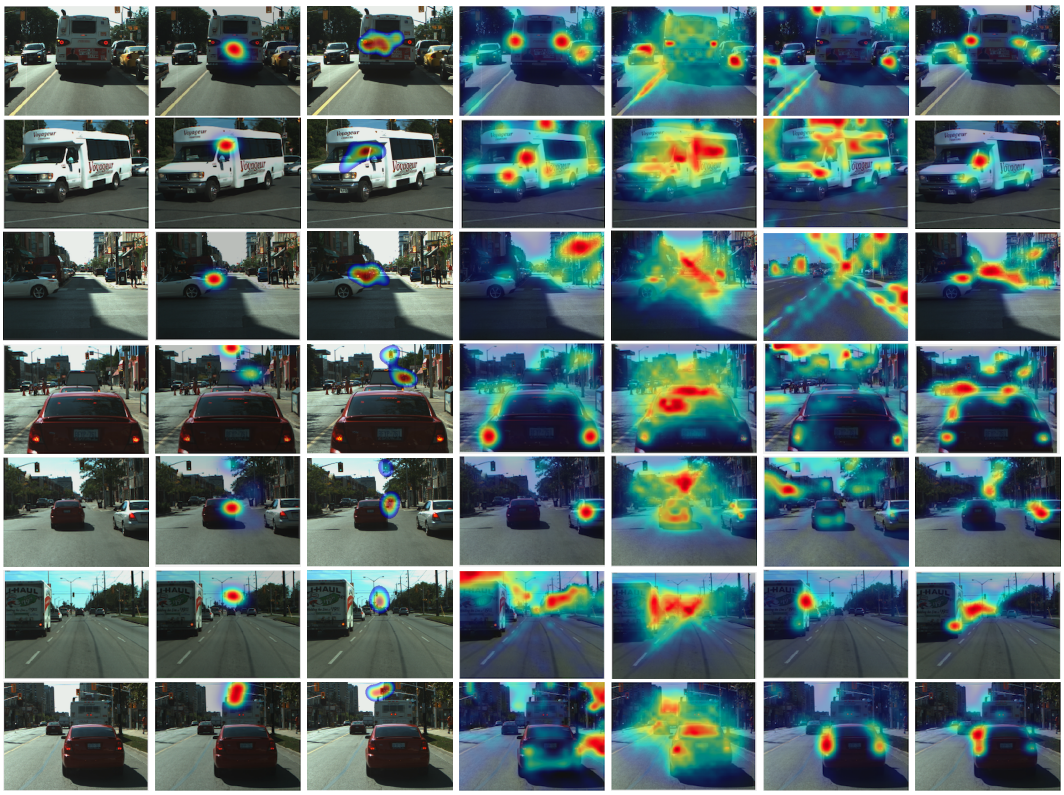


Figure 5.4: **(from left to right:)** input frames, ground truth fixation maps, our predicted saliency maps, and the predictions of Itti [7], GBVS [8], Image Signature [26], and HFT [10]

.

primaray and secondary fixation, if they exist.

5.5.2 Quantitative Evaluation Metrics

We have evaluated our model's performance on various metrics to measure the correspondence between the driver's eye fixation prediction and the ground truth driver's eye fixation.

Some of the metrics considered herein are based on the location of fixation, such as Normalized Scanpath Saliency (NSS) [27], and Area under ROC Curve (AUC-Borji [28], AUC-Judd [29]). They evaluate the similarity between the driver's eye fixation prediction and ground-truth. In contrast, others are based on distributions, such as Earth Movers Distance (EMD) [30], Similarity Metric (SIM) [29], and Linear Correlation Coefficient (CC) [31]. They evaluate the dissimilarity between the model's prediction and ground truth. Let S_G represent the ground-truth driver's eye fixation map and S_F the saliency maps prediction provied by the various methods:

- **Normalized Scanpath Saliency (NSS):** The NSS metric is computed by the average normalized saliency at driver's eye fixation locations, as follows:

$$\mathrm{NSS} = \frac{1}{N} \sum_{n=1}^{N} \frac{S_F(x_n, y_n) - \mu_{S_F}}{\sigma_{S_F}} \tag{5.3}$$

  where N is the number of eye positions, (x_n, y_n) the eye-fixation point location, and μ_{S_F}, and σ_{S_F} are the mean and standard deviation of a driver's eye fixation map predication.

- **Area Under the ROC Curve (AUC):** AUC is commonly used for evaluating estimated saliency maps. With AUC, two types of locations are considered: the true driver fixation points, regarded as the positive

set, versus a negative set consisting of the sum of other fixation points. The driver's eye fixation map is classified into the salient and non-salient regions with a predetermined threshold. Then, the ROC curve is plotted by the true-positive (TP) rate versus the false-positive (FP) rate, as the threshold varies from 0 to 1. Depending on the non-fixation distribution's selection, there are two commonly used types of AUC, namely AUC-Judd and AUC-Borji.

- **Linear Correlation Coefficient (CC):** The CC provides a measure of the linear relationship between S_F and S_G. This metric varies between -1 and 1, and a value close to either -1 or 1 shows alignment between S_F and S_G:

$$\mathrm{CC} = \frac{\mathrm{cov}(S_F, S_G)}{\sigma_{S_F} \times \sigma_{S_G}} \tag{5.4}$$

- **Similarity Metric (SIM):** This metric estimates the similarity between the distributions of predicted and ground truth driver's eye fixation maps by easuring the intersection between two distributions, calculated by a sum of the minimum values at any pixel location from distributions ($S_F(n)$ and $S_G(n)$):

$$\mathrm{SIM} = \sum_{n=1}^{N} \min(S_F(n), S_G(n)) \tag{5.5}$$

where, $S_F(n)$ and, $S_G(n)$ are normalized distributions, and N is the number of locations of interest in the maps. A value close to 1 indicates that the two saliency maps are similar, while the score close to zero denotes little overlap.

- **Earth Mover's Distance (EMD):** This metric computes the spatial

distance between two probability distributions $S_F(n)$ and $S_G(n)$ over a region, as the minimum cost of transforming the probability distribution of the computed driver's eye fixation map $S_F(n)$ into the ground truth $S_G(n)$. A high value for EMD indicates little similarity between the distributions.

Table 5.2: Saliency metric scores of our model as compared with state-of-the-art saliency models on the RoadLab dataset.

Models	NSS	CC	SIM	AUC-Borji	AUC-Judd	EMD
GT	3.26	1	1	0.88	0.94	0
ITTI [7]	1.15	0.23	0.25	0.62	0.64	2.13
GBVS [8]	1.32	0.29	0.32	0.69	0.71	1.91
Image Signature [26]	1.48	0.29	0.30	0.73	0.75	2.06
HFT [10]	1.42	0.42	0.38	0.64	0.66	2.31
ΔQDCT [9]	1.68	0.34	0.32	0.71	0.73	1.72
RARE2012 [32]	1.34	0.31	0.33	0.67	0.68	1.48
ML Net [33]	2.47	0.72	0.66	0.76	0.80	1.43
Wang [18]	2.87	0.78	0.68	0.81	0.85	1.23
Proposed	2.98	0.82	0.72	0.81	0.89	1.06

To illustrate the effectiveness of the saliency map model in predicting a driver's eye fixation, we compared our model with eight state-of-the-art tech-

niques, including six non-AI models: ITTI [7], GBVS [8], Image Signature [26], HFT [10], RARE2012 [32], ΔQDCT [9], and two deep learning-based models: ML-Net [33], and Wang [18]. These models have been introduced in recent years and are often utilized for comparison purposes.

The quantitative results obtained on the RoadLAB dataset [23] are presented in Table 5.2. Our proposed model gives the maximum similarity and minimum dissimilarity with respect to the ground truth data. We conclude that our model predicts the driver's eye fixation maps more accurately than other saliency models.

5.6 Conclusions

We proposed convolution neural networks to predict the potential saliency maps in the driving environment, and then employed our previous research results to estimate the probability of the driver gaze direction, given head pose as a top-down factor. Finally, we statistaically combined bottom-up and top-down factors to obtain accurate drivers' fixation predictions. Due to simplicity, we test out model on selected frames from the RoadLab dataset in which the quality of head and gaze matrices, estimated by remote eye tracker, is more than a predetermined threshold.

Our previous study established that driver gaze estimation is a crucial factor for driver maneuver prediction. The identification of objects that drivers tend to fixate on is of equal importance in maneuver prediction models. We believe that the ability to estimate these aspects of visual behavior contitutes a significant improvement for the prediction of maneuvers, as drivers generally focus on environmental features a few seconds before effecting one or more maneuvers.

Bibliography

[1] Organization WH, et al. Global status report on road safety 2018: Summary. World Health Organization; 2018.

[2] Zabihi S, Beauchemin SS, De Medeiros E, Bauer MA. Frame-rate vehicle detection within the attentional visual area of drivers. In: 2014 IEEE Intelligent Vehicles Symposium Proceedings. IEEE; 2014. p. 146–150.

[3] Shirpour M, Beauchemin SS, Bauer MA. A Probabilistic Model for Visual Driver Gaze Approximation from Head Pose Estimation. In: 2020 IEEE 3nd Connected and Automated Vehicles Symposium (CAVS). IEEE; 2020. .

[4] Khairdoost N, Shirpour M, Bauer MA, Beauchemin SS. Real-Time Maneuver Prediction Using LSTM. IEEE Transactions on Intelligent Vehicles. 2020;.

[5] Jain A, Koppula HS, Raghavan B, Soh S, Saxena A. Car that knows before you do: Anticipating maneuvers via learning temporal driving models. In: Proceedings of the IEEE International Conference on Computer Vision; 2015. p. 3182–3190.

[6] Deng T, Yang K, Li Y, Yan H. Where does the driver look? Top-down-based saliency detection in a traffic driving environment. IEEE Transactions on Intelligent Transportation Systems. 2016;17(7):2051–2062.

[7] Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on pattern analysis and machine intelligence. 1998;20(11):1254–1259.

[8] Harel J, Koch C, Perona P. Graph-based visual saliency. In: Advances in neural information processing systems; 2007. p. 545–552.

[9] Schauerte B, Stiefelhagen R. Quaternion-based spectral saliency detection for eye fixation prediction. In: European Conference on Computer Vision. Springer; 2012. p. 116–129.

[10] Li J, Levine MD, An X, Xu X, He H. Visual saliency based on scale-space analysis in the frequency domain. IEEE transactions on pattern analysis and machine intelligence. 2012;35(4):996–1010.

[11] Shirpour M, Beauchemin SS, Bauer MA. What Does Visual Gaze Attend to During Driving? In: VEHITS; 2021. .

[12] Deng T, Yan H, Li YJ. Learning to boost bottom-up fixation prediction in driving environments via random forest. IEEE Transactions on Intelligent Transportation Systems. 2017;19(9):3059–3067.

[13] Borji A, Cheng MM, Jiang H, Li J. Salient object detection: A benchmark. IEEE transactions on image processing. 2015;24(12):5706–5722.

[14] Vig E, Dorr M, Cox D. Large-scale optimization of hierarchical features for saliency prediction in natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2014. p. 2798–2805.

[15] Liu N, Han J, Zhang D, Wen S, Liu T. Predicting eye fixations using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015. p. 362–370.

[16] Kümmerer M, Wallis TS, Bethge M. DeepGaze II: Reading fixations from deep features trained on object recognition. arXiv preprint arXiv:161001563. 2016;.

[17] Huang X, Shen C, Boix X, Zhao Q. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In: Proceedings of the IEEE International Conference on Computer Vision; 2015. p. 262–270.

[18] Wang W, Shen J. Deep visual attention prediction. IEEE Transactions on Image Processing. 2017;27(5):2368–2378.

[19] Wang W, Shen J, Dong X, Borji A, Yang R. Inferring salient objects from human fixations. IEEE transactions on pattern analysis and machine intelligence. 2019;.

[20] Palazzi A, Abati D, Solera F, Cucchiara R, et al. Predicting the Driver's Focus of Attention: the DR (eye) VE Project. IEEE transactions on pattern analysis and machine intelligence. 2018;41(7):1720–1733.

[21] Tawari A, Kang B. A computational framework for driver's visual attention using a fully convolutional architecture. In: 2017 IEEE Intelligent Vehicles Symposium (IV). IEEE; 2017. p. 887–894.

[22] Deng T, Yan H, Qin L, Ngo T, Manjunath B. How do drivers allocate their potential attention? Driving fixation prediction via convolutional neural networks. IEEE Transactions on Intelligent Transportation Systems. 2019;21(5):2146–2154.

[23] Beauchemin SS, Bauer MA, Kowsari T, Cho J. Portable and scalable vision-based vehicular instrumentation for the analysis of driver in-

tentionality. IEEE Transactions on Instrumentation and Measurement. 2011;61(2):391–401.

[24] Kowsari T, Beauchemin SS, Bauer MA, Laurendeau D, Teasdale N. Multi-depth cross-calibration of remote eye gaze trackers and stereoscopic scene systems. In: 2014 IEEE Intelligent Vehicles Symposium Proceedings. IEEE; 2014. p. 1245–1250.

[25] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556. 2014;.

[26] Hou X, Harel J, Koch C. Image signature: Highlighting sparse salient regions. IEEE transactions on pattern analysis and machine intelligence. 2011;34(1):194–201.

[27] Peters RJ, Iyer A, Itti L, Koch C. Components of bottom-up gaze allocation in natural images. Vision research. 2005;45(18):2397–2416.

[28] Borji A, Sihite DN, Itti L. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. IEEE Transactions on Image Processing. 2012;22(1):55–69.

[29] Judd T, Durand F, Torralba A. A benchmark of computational models of saliency to predict human fixations. 2012;.

[30] Pele O, Werman M. A linear time histogram metric for improved sift matching. In: European conference on computer vision. Springer; 2008. p. 495–508.

[31] Le Meur O, Le Callet P, Barba D. Predicting visual fixations on video based on low-level visual features. Vision research. 2007;47(19):2483–2498.

[32] Riche N, Mancas M, Duvinage M, Mibulumukini M, Gosselin B, Dutoit T. Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. Signal Processing: Image Communication. 2013;28(6):642–658.

[33] Cornia M, Baraldi L, Serra G, Cucchiara R. A deep multi-level network for saliency prediction. In: 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE; 2016. p. 3488–3493.

Chapter 6

Vanishing Points

This Chapter is a reformatted version of the following article: M. Shirpour, S.S. Beauchemin, and M.A. Bauer, *What Does Visual Gaze Attend to During Driving?, submitted in 7th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS) Conference, Prague, Czech Republic, 2021.*

This study aims to analyze driver Cephalo-Ocular behavior features and road vanishing points according with repsect to vehicle speed in urban and suburban areas using data obtained from an instrumented vehicle's eye tracker. This study utilizes two models for driver gaze estimation. The first model estimates the 3D point of the driver's gaze in absolute coordinates obtained through the combined use of an imaging plane of the forward stereo vision system and an eye-gaze tracker system. The second approach uses a stochastic model, known as Gaussian Process Regression (GPR), that estimates the most probable gaze direction given head pose. We evaluated models on real data gathered in an urban and suburban environment with the RoadLAB experimental vehicle.

6.1 Introduction

The human visual system collects about 90% of the information that is needed to adequately perform driving tasks [1]. Driver gaze has been studied for many years in driving simulators and real driving environments. It has been demonstrated that driver gaze direction in relation to the surrounding driving environment is predictive of driver maneuvers [2]. In addition to these results, our aim is to elucidate the rules that govern driver gaze with respect to the characteristics of vehicular dynamics. In particular, this contribution reports on our investigation of the relationship that exists between gaze behavior, vanishing points, and vehicle speed.

6.1.1 Literature Survey

Driver visual attention plays a prominent role in intelligent Advanced Driver Assistance Systems (i-ADAS). Some driver monitoring systems utilize the driver's head pose and eyes to evaluate the driver's gaze-direction and zone [3, 4]. We recently presented a stochastic model that derives gaze direction from head pose data provided by a contactless gaze tracking system [4]. This model computes a probabilistic visual attention map that estimates the probability of finding the actual gaze over the stereo system's imaging plane, with a Gaussian Process Regression (GPR) technique. Subsequently, we proposed a deep learning model to predict driver eye fixation according to driver's visual attention [5]. In addition, other contributions use the direction of gaze to detect 2D image gaze regions [6, 7]. Others have defined a framework that uses the 3D Point of Gaze (PoG) and Line of Gaze (LoG) in absolute coordinates for similar purposes [8].

In other works, the driver's attentional visual area was modelled as in-

tersection of the elliptical region formed by the cone emanating from the eye position with the LoG as its symmetrical axis along its length, with the imaging plane of the forward stereoscopic vision system installed in the experimental vehicle, as depicted in Figure 6.1. Using this mechanism, several authors were able to estimate the driver's most probable next maneuver some time before it occurred [2, 9]. Their evaluation showed a strong relationship between driver gaze behaviour and maneuvers.

In general, a driver concentrates on parts of the driving scene that contain some objective and subjective elements. Objective elements are obtained with bottom-up approaches that consider features extracted from the driving environment such as traffic-related objects. On the other hand, subjective elements are obtained with top-down approaches and are attributed to a driver's internal factors, such as experience or intention [10]. Top-down strategies provide insight into what a driver's gaze could be fixated on while driving.

6.1.2 Human Vision System

The human visual field affords a remarkably broad view of the world, in the range of 90° to the left and right, and more than 60° above and below the gaze [11]. Information within 2° of the gaze is processed in foveal vision. More broadly, parafoveal vision covers up to 6° of visual angle [12]. This implies that the existing information in the parafovea is combined with that from the fovea. The information from the fovea is clearer when compared with the information present in the parafovea [13]. Together, the foveal and parafoveal areas are known as the central visual field, where objects are clearly and sharply seen and used to perform most activities [11].

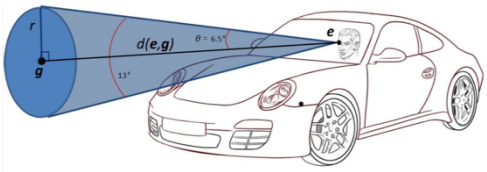


Figure 6.1: *The attentional area is defined as the elliptical region formed by the cross-section of a cone emanating from the eye position with the LoG as its symmetrical axis along its length, and the imaging plane of the forward stereoscopic vision system.*

Figure 6.2: **(left):** *Stereo vision system located on the vehicle's roof;* **(center):** *infrared gaze tracker;* **(right:)** *FaceLAB system interface.*

6.1.3 Experimental Vehicle

Our research vehicle is equipped with instruments that capture driver-initiated vehicular actuation and relate the 3D driver gaze direction on the imaging plane of the forward stereoscopic vision system. The vehicle was used to gather data sequences from 16 different test drivers on a pre-determined 28.5km route within the city of London, Ontario, Canada. 3TB of driving sequences were recorded. The data contains significant driving information, including forward stereo imaging and depth, 3D PoG and head pose, and vehicular dynamics obtained with the OBDII CANBus interface. Image and data frames are collected at a rate of 30Hz. The vehicular instrumentation consists of a non-contact infrared remote gaze and head pose tracker, with two cameras mounted on the vehicle dashboard, operating at 60Hz. This instrument provides head movement and pose, eye position, and gaze direction

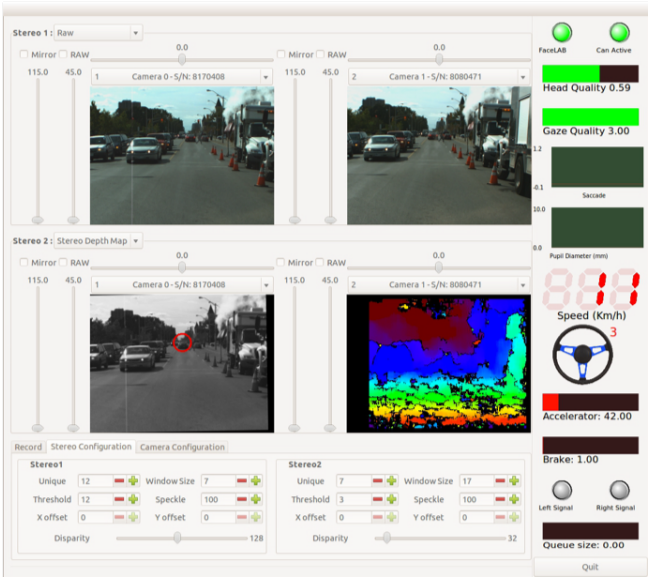


Figure 6.3: *RoadLab software systems: The on-board system displays frame sequences with depth maps, dynamic vehicle features, and eye-tracker data.*

within its own coordinate system. A forward stereoscopic vision system is located on the vehicle's roof to capture frontal view information such as dense stereo depth maps at 30 Hz (See Figure 6.2). Details concerning this instrumentation are available in [14]. The sum of our data was recorded with the RoadLAB software system, as shown in Figure 6.3.

6.2 Methodology

This Section describes two models for describing driver gaze visual attention in the forward stereo imaging system. Section 6.2.1 addresses the calibration procedure applied to provide the Point of Gaze (PoG) onto the imaging plane of the forward stereo system. We introduce a Gaussian Process Regression (GPR) that estimates the probability of gaze direction according to driver head pose in Section 6.2.2. Section 6.2.3 describes the tecnique we employ to

locate vanishing points from the stereoscopic imagery.

6.2.1 Projection of PoGs Onto Stereo System

The calibration process brings the eye tracker data into the coordinate system of the forward steroscopic instrumentation. We used a cross-calibration technique developed in our laboratory to transform the 3D driver gaze expressed in the eye tracker reference frame to that of forward stereoscopic vision system [8]. This calibration process is defined as follows:

- *Salient Points Extraction:* A sufficient number of salient points are extracted from the steroscopic imagery (around 20 points provide sufficient data)

- *Depth Estimation:* The driver's eye fixates on preselected salient points for a short period (about 2 seconds). The depth estimate of the salient point, the gaze vector, and the position of the eye center are recorded.

- *Estimation of Rotation and Translation Matrices:* The process estimates the rigid body transformation between the reference frame of the stereoscopic system and the remote eye tracker. The elements composing this transformation are known as extrinsic calibration parameters.

- *Gaze projection onto the imaging plane of stereoscopic system:* The LoG, expressed in eye tracker coordinates, is projected onto the imaging plane of the stereo system using the extrinsic calibration parameters. The PoG is determined as the location where the LoG intersects with a valid depth estimate within the reference frame of the stereo vision system.

6.2.2 Gaussian Process Regression

Technically, direct use of gaze is complicated by the fact that eyes may exhibit rapid saccadic movements resulting in difficulties for assessing the correct image area corresponding to a driver's visual attention. Our research lab proposed another model to alleviate this problem by approximating the 3D gaze from the 3D head pose, as the head does not experience saccadic movements.

In our recent research, instead of directly estimating the gaze, which depends on the driver's visual cognitive tasks, we introduced a stochastic model for representing driver visual attention. This model inherits the advantage of the Gaussian Process Regression (GPR) technique to estimate the probability of the driver's gaze direction according to head pose over the imaging plane of the stereo system. It establishes a confidence area within which the driver gaze is most likely contained. We have shown that drivers concentrate most of their attention on the 95% confidence interval region estimated from the head pose. We refer the reader to [**?**] for details on the GPR technique.

6.2.3 Vanishing Points

A vanishing point is a point on the image plane where the two-dimensional perspective projections of mutually parallel lines in three-dimensional space appear to converge. The vanishing point plays an essential role in the prediction of driver eye fixations. The vanishing point is considered as guidance for predicting driver intent, as drivers mostly gaze at traffic objects near the vanishing point.

Available methods to detect the vanishing point are mainly edge, region, or texture-based models. Edge-based models are adequate when edge boundaries and lane markings are available within the driving scene. Region-based

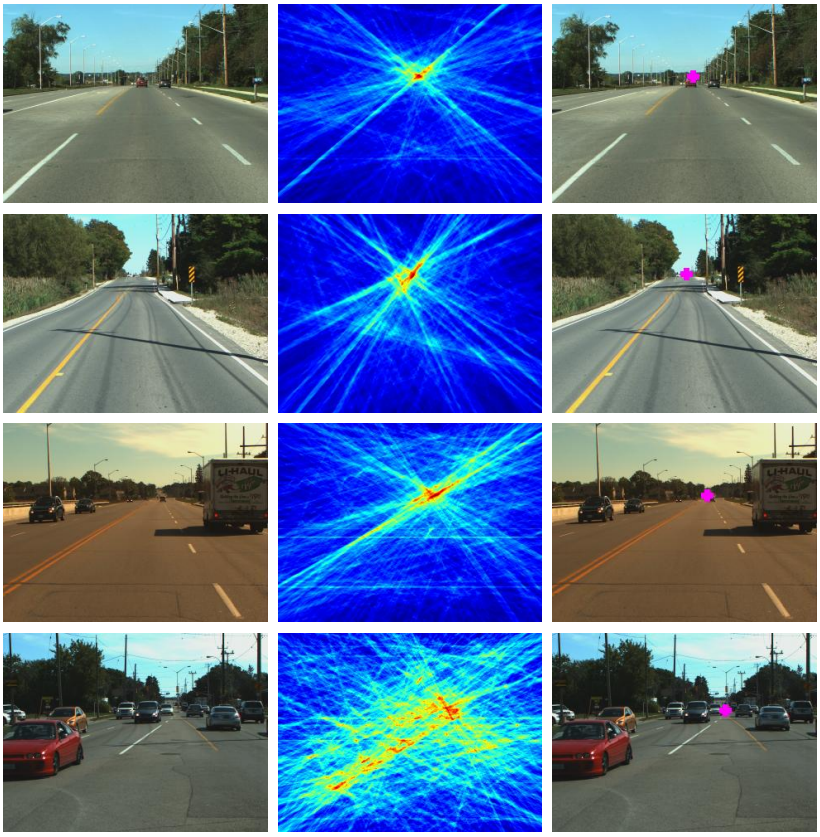


Figure 6.4: *Examples of vanishing-points (from left to right:) input frames, voting map, and detected vanishing points.*

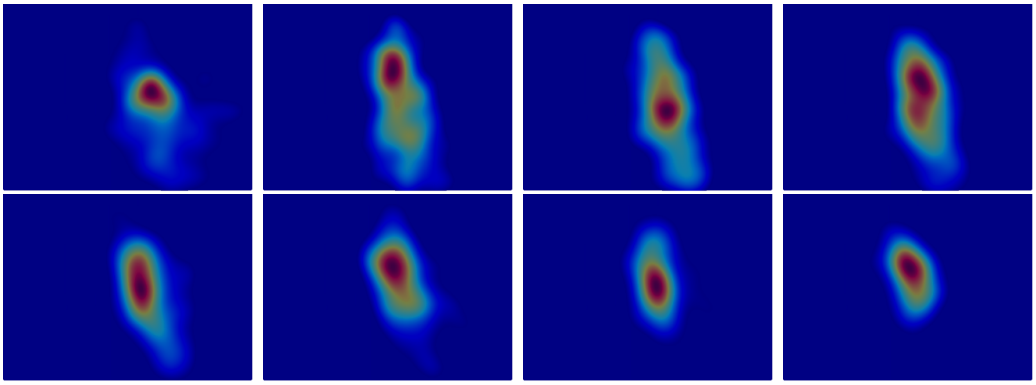


Figure 6.5: *Driver attention versus vanishing point with respect to speed.* **a) to h):** *As the speed increases, the driver gaze converges to the vanishing point.*

Table 6.1: Description of Data Used For Analyze of Drivers Gaze and Vanishing Point according to vehicle speed: A (0≤Speed <10), B (10≤Speed <20), C (20≤Speed <30), D (30≤Speed <40), E (40≤Speed <50), F (50≤Speed <60), G (60≤Speed <70), and H(Speed≥70)

Seq#	A	B	C	D	E	F	G	H
Seq. 2	11530	2693	3181	4426	4475	3930	4371	2350
Seq. 8	8515	2556	2959	3297	3594	3679	2157	2543
Seq. 9	7756	2544	3263	4197	3131	3148	3169	2166
Seq. 10	7199	1538	2068	3912	4665	4200	3042	1211
Seq. 11	8008	1714	2425	3373	3417	3330	2954	887
Seq. 13	11545	1956	2098	2248	2447	2711	3528	2605
Seq. 14	4495	1123	1311	1986	2285	2442	1204	1448
Seq. 16	9056	2085	2440	3046	2874	3321	1241	1628

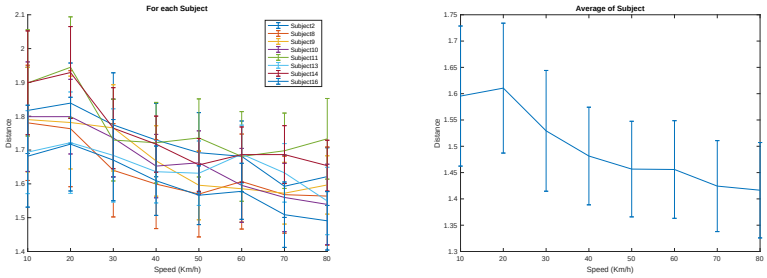


Figure 6.6: **Model A (Left):** *Average and variance of distance from driver gaze fixation to vanishing point versus vehicle speed for each driver.* **(right):** *Average of all drivers.*

methods divide the driving view into path and non-path according to low-level features (color, intensity, etc). These two types of models are suitable for structured roads. They experience difficulty with scenery involving unstructured or complex features.

Because the RoadLab dataset includes both structured and unstructured imaging elements, we adopted a texture-based model proposed by [15]. Their model is based on Gabor filters to estimate the local orientation of pixels. Figure 6.4 shows a sample of RoadLab frames with detected vanishing points.

6.3 Analysis of Driver Attention

In this Section, we describe the preprocessing we applied to the RoadLAB dataset and provide our analysis of the results we obtained.

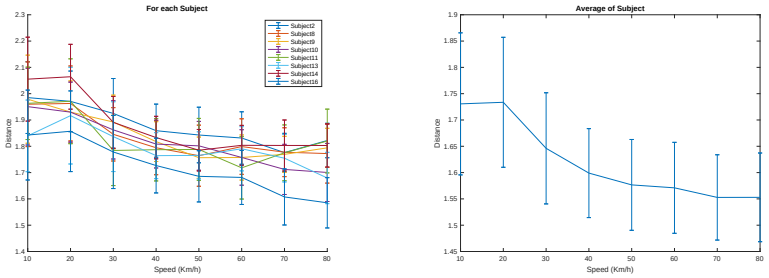


Figure 6.7: **Model B (Left):** *Average and variance of distance from driver gaze fixation to vanishing point versus vehicle speed for each driver.* **(right):** *Average of all drivers.*

6.3.1 Data Preparation

Our experimental vehicle relies on sensors and cameras to track its driver's behavioral features. The RoadLab software provided a confidence measure on the quality of its estimations of head pose and gaze. The head pose confidence measure ranged from 0 to 2, while the gaze quality metric ranged from 0 to 3. We considered the head pose and gaze as reliable elements when these metrics had a minumum value of 1 or, and 2 or higher for the gaze. The PoG that passed the quality metric thresholds were projected onto the vehicle forward stereo system for the 5 preceding consecutive frames. Table 6.1 provides the number of frames selected from test drivers according to vehicular speed.

6.3.2 Speed and Visual Attention Analysis

Our results show that drivers generally tend to concentrate their gaze on vanishing points created by the motion of the vehicle. Figure 6.5 illustrates the fact that the frequency of driver gaze fixations near the vanishing point is considerably higher than that of fixations on other image regions. This indicates that driver attention is more likely to fixate on traffic objects near the vanishing point. Also, Figure 6.5 illustrates how the gaze position changes at different vehicle speeds (for one particular driving sequence). When the vehicle speed smoothly increases from below 10 km/h to over 70 km/h, the gaze position rapidly converges to the vanishing point.

We estimated driver visual attention with two different models for gaze direction: model A which estimates the probability of driver gaze direction according to head pose, and model B which directly uses the 3D driver gaze in absolute coordinates. We measured the logarithmic distance of gazes from vanishing points and calculated the averages and variances of these distances for a range of vehicle speeds. As observed in Figures 6.6 and 6.7 the distance average of gaze fixations and vanishing points decreases significantly with an increase in vehicle speed. These results show that the drivers were more focused on vanishing points at high the vehicle speeds. The variance of gaze fixations at high vehicluar speeds is significantly lower than that observed at lower speeds.

The human visual system is limited in the quantity of information it is able to process per time unit, and compensates by decreasing its visual field when the mass of elements to process in the spatial or temporal context increases. In driving circumstances, this generally occurs at high speeds, as the amount of available information per unit of time increases proportionally.

6.4 Conclusions

The literature shows that the vanishing point is a helpful clue in driving and other visual tasks. We analyzed driver gaze behavior in relation to vanishing points with respect to increasing vehicular speeds with the RoadLab dataset obtained from an instrumented vehicle. This research investigated two models for driver gaze estimation. The first model estimated 3D point of gaze in absolute coordinate, while the second model used a probabilistic process to estimate the probability of driver gaze direction based on the head pose. The results clearly indicate that vanishing points attract driver gaze with increasing force at high vehicle speeds for both models. These results can be considered a measure of driver distraction when the driver gaze deviates from the vanishing point in different vehicle speeds.

Bibliography

[1] Sivak M. The information that drivers use: is it indeed 90% visual? Perception. 1996;25(9):1081–1089.

[2] Khairdoost N, Shirpour M, Bauer MA, Beauchemin SS. Real-Time Maneuver Prediction Using LSTM. IEEE Transactions on Intelligent Vehicles. 2020;5(4):714–724.

[3] Jha S, Busso C. Probabilistic estimation of the gaze region of the driver using dense classification. In: International Conference on Intelligent Transportation Systems. IEEE; 2018. p. 697–702.

[4] Shirpour M, Beauchemin SS, Bauer MA. A Probabilistic Model for Visual

Driver Gaze Approximation from Head Pose Estimation. In: Connected and Automated Vehicles Symposium. IEEE; 2020. p. 1–6.

[5] Shirpour M, Beauchemin SS, Bauer MA. Driver's Eye Fixation Prediction by Deep Neural Network. In: Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4 VISAPP: VISAPP,. INSTICC. SciTePress; 2021. p. 67–75.

[6] Shirpour M, Khairdoost N, Bauer MA, Beauchemin SS. Traffic Object Detection and Recognition: A Survey and an Approach Based-on the Attentional Visual Field of Driver. IEEE Transactions on Intelligent Vehicles (in press);.

[7] Zabihi SM, Beauchemin SS, De Medeiros EAM, Bauer MA. Frame-rate vehicle detection within the attentional visual area of drivers. In: Intelligent Vehicles Symposium. IEEE; 2014. p. 146–150.

[8] Kowsari T, Beauchemin SS, Bauer MA, Laurendeau D, Teasdale N. Multi-depth cross-calibration of remote eye gaze trackers and stereoscopic scene systems. In: Intelligent Vehicles Symposium. IEEE; 2014. p. 1245–1250.

[9] Zabihi SM, Beauchemin SS, Bauer MA. Real-time driving manoeuvre prediction using IO-HMM and driver cephalo-ocular behaviour. In: Intelligent Vehicles Symposium. IEEE; 2017. p. 875–880.

[10] Deng T, Yang K, Li Y, Yan H. Where does the driver look? Top-down-based saliency detection in a traffic driving environment. IEEE Transactions on Intelligent Transportation Systems. 2016;17(7):2051–2062.

[11] Wolfe B, Dobres J, Rosenholtz R, Reimer B. More than the Useful Field: Considering peripheral vision in driving. Applied ergonomics. 2017;65:316–325.

[12] Engbert R, Longtin A, Kliegl R. A dynamical model of saccade generation in reading based on spatially distributed lexical processing. Vision Research. 2002;42(5):621–636.

[13] Kennedy A. Parafoveal processing in word recognition. The Quarterly Journal of Experimental Psychology A. 2000;53(2):429–455.

[14] Beauchemin SS, Bauer MA, Kowsari T, Cho J. Portable and scalable vision-based vehicular instrumentation for the analysis of driver intentionality. IEEE Transactions on Instrumentation and Measurement. 2011;61(2):391–401.

[15] Moghadam P, Starzyk JA, Wijesoma WS. Fast vanishing-point detection in unstructured environments. IEEE Transactions on Image Processing. 2011;21(1):425–430.

Chapter 7

Conclusion and Future Work

In this research, we have demonstrated that maneuver prediction is possible a few seconds ahead of time. Therefore such functionalities would allow an ADAS to determine whether the next most probable maneuver is safe or unsafe.

In Chapter 2 we developed a prediction model using LSTM that anticipates 5 types of driver maneuvers. Our model used both vehicular dynamics and driver cephalo-ocular behavior as a basis for maneuver prediction. Quantitative results in this contribution demonstrated the superiority of deep learning techniques over traditional machine learning for the purpose of real-time driver maneuver prediction.

Identifying objects that drivers visually atttend to potentially reveals the objects of driver visual attention. Chapter 3 provides a vision-based framework that detects and recognizes traffic objects inside and outside the driver's attentional visual area. This approach uses the driver 3D absolute gaze point obtained through the combined use of a front-view stereo imaging system and a non-contact 3D gaze tracker. We built a model from a combination of multi-

scale HOG-SVM and Faster R-CNN-based models. The recognition stage is performed by employing a ResNet-101. This contribution empirically demonstrates that the identification of objects drivers visually attend to is indeed feasible in a real-time fasion. Conversely, it becomes equally feasible to identify objects drivers may incorrectly not visually attend to.

Generally, the driver moves both the head and eyes to obtain a fixation. In Chapters 4, we provided techniques to obtain confidence intervals within which driver gaze may fall into, using head pose instead of explicit gaze direction. These results may simplify the on-board equipment required for gaze estimation within the immediate environment of the vehicle.

In Chapter 5, we proposed convolution neural networks to predict saliency regions in the driving environment and used the estimated driver gaze direction heat map as estimated in Chapter 4 to obtain the intersections of most probable gaze direction and location of salient objects. These results may be used to ascertain if a driver is gazing at salient traffic objects, which may be of importance in assessing a driver's competence in safely performaing the task of driving a vehicle in real-time.

In Chapter 6 we analyzed driver gaze behavior with respect to driving speed and vehicular motion-induced vanishing points. We were able to demonstrate that drivers visual attention tend to shift towards these vanishing points with a probability that increased with vehicular speed. This result extends our knowledge of driver visual behavior in a general sense.

Our contributions are summarized as follows:

1. Proposing a real-time model to predict driver maneuvers

2. Presenting a framework to detect and recognize traffic objects inside a driver's attentional field.

3. Collecting and labeling a large dataset for different traffic objects

4. Proposing a stochastic method to identify forward image regions attracting the visual attention of drivers

5. Proposing a deep neural network for the prediction of drivers eye fixations

6. Analysing driver gaze behavior with respect to vanishing points and vehicle speed

7.1 Future Work

Research on modeling driver intent is a recent endeavour with the potential for notable results in the near future. Here are five possible research areas that could be undertaken directly:

1. The predictive model for driver gaze direction could be used as an input feature in the driving maneuver prediction model.

2. The driver gaze prediction model coupled with the identification of salient objects could be used to assess if a driver's visual attention is attending to relevant traffic objects, given the most likely next maneuver.

3. While the instrumentation represents a successful proof of concept, it was noted that wider viewing angles for the stereo cameras and eye-trackers using more than two cameras (to compensate for head rotations) would allow us to track the 3D driver gaze into the surroundings in a more comprehensive manner.

4. The physical limitations of the instrumentation prevented its use at night and in adverse weather conditions. Such limitations could be removed entirely by a judicious choice of hardware, enabling the study of driver intent in diverse conditions.

5. Features play a critical role in maneuver prediction system. Moreover, using the SHAP (Shapley Additive exPlanations) method, we can determine which of the features is of higher importance to the prediction systems. The importance of these features has not been taken into account in the current research.

Vita

NAME:	Mohsen Shirpour
POST-SECONDARY EDUCATION AND DEGREES:	University of Western Ontario London, Canada 2017-2020 PhD in Computer Science.
	Sharif University Tehran, Iran 2012-2014 M.Sc. in Computer Eng.
	Shahid Bahonar University Kerman, Iran 2008-2012 B.Sc. in Computer Science.
AWARDS:	WGRS scholarship for 2017-2020, University of Western Ontario First rank, according to the total GPA among B.S. students of Computer Science.
RELATED WORK EXPERIENCE:	Research Assistant Sharif University 2014 - 2017 Teaching Assistant University of Western Ontario 2017 - 2020 Research Assistant University of Western Ontario 2017 - 2020

Bibliography

[1] K Aghajani, M Shirpour, and MT Manzuri. Structural image representation for image registration. In *2015 The International Symposium on Artificial Intelligence and Signal Processing (AISP)*, pages 95–100. IEEE, 2015.

[2] K Aghajani, R Yousefpour, M Shirpour, and MT Manzuri. Intensity based image registration by minimizing the complexity of weighted subtraction under illumination changes. *Biomedical Signal Processing and Control*, 25:35–45, 2016.

[3] N. Khairdoost, M. Shirpour, M. A. Bauer, and S. S. Beauchemin. Real-time driver maneuver prediction using lstm. *IEEE Transactions on Intelligent Vehicles*, 5(4):714–724, 2020.

[4] J Motley, AJ Nelson, L Watson, LS Poeta, DH Seston, MF Heidari, and M Shirpour. Improving digitally stitched x-rays and interpretational standards for field paleoradiography.

[5] J Rahnama, M Shirpour, and MT Manzuri. Single image super resolution by adaptive k-means clustering. In *2017 10th Iranian Conference on Machine Vision and Image Processing (MVIP)*, pages 209–214. IEEE, 2017.

[6] H Sarikhani, E Abdollahian, M Shirpour, A Javaheri, and MT Manzuri. A robust and invariant keypoint extraction algorithm in brain mr images. In *International Symposium on Artificial Intelligence and Signal Processing*, pages 121–130. Springer, 2013.

[7] M Shirpour. 3d human pose estimation from single images using dual-wing harmonium model. *International Journal of Applied Pattern Recognition*, 2(3):199–212, 2015.

[8] M Shirpour, K Aghajani, and MT Manzuri-Shalmani. A new similarity measure for intensity-based image registration. In *2014 4th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 227–232. IEEE, 2014.

[9] M. Shirpour, S. S. Beauchemin, and M. A. Bauer. A probabilistic model for visual driver gaze approximation from head pose estimation. In *2020 IEEE 3rd Connected and Automated Vehicles Symposium (CAVS)*, pages 1–6, 2020.

[10] M Shirpour, SS Beauchemin, and MA Bauer. Driver's eye fixation prediction by deep neural network. 2021.

[11] M Shirpour, SS Beauchemin, and MA Bauer. What does visual gaze attend to during driving? In *VEHITS*, 2021.

[12] M Shirpour, N Khairdoost, MA Bauer, and SS Beauchemin. Traffic object detection and recognition: A survey and an approach based-on the attentional visual field of driver. *IEEE Transactions on Intelligent Vehicles (in press)*.