

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Vision based, Multi-cue Driver Models for Intelligent Vehicles

Permalink

<https://escholarship.org/uc/item/4v27v981>

Author

Martin, Sujitha Catherine

Publication Date

2016

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Vision based, Multi-cue Driver Models for Intelligent Vehicles

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering
(Intelligent Systems, Robotics, and Control)

by

Sujitha Martin

Committee in charge:

Professor Mohan M. Trivedi, Chair
Professor Garrison Cottrell
Professor David Kriegman
Professor Truong Nguyen
Professor Bhaskar Rao

2016

Copyright
Sujitha Martin, 2016
All rights reserved.

The dissertation of Sujitha Martin is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2016

DEDICATION

To my beloved family.

EPIGRAPH

“Be anxious for nothing, but in everything by prayer and supplication, with thanksgiving, let your requests be made known to God; and the peace of God, which surpasses all understanding, will guard your hearts and minds through Christ Jesus.”

– Philippians 4:6-7, The Holy Bible

TABLE OF CONTENTS

Signature Page	iii
Dedication	iv
Epigraph	v
Table of Contents	vi
List of Figures	ix
List of Tables	xiii
Acknowledgements	xiv
Vita	xvi
Abstract of the Dissertation	xviii
Chapter 1 Introduction	1
Chapter 2 End-to-End, Continuous Gaze Estimation Framework	6
2.1 Introduction	6
2.2 Related Research	7
2.3 Face Analysis: Algorithm Development	9
2.3.1 Face Detection	9
2.3.2 Landmark Estimation	10
2.3.3 Head Pose	11
2.3.4 Eye Cues	12
2.3.5 Multiple Camera Framework	13
2.3.6 Gaze Zone Estimation	15
2.4 Face Analysis: Evaluation on Naturalistic Driving Data	16
2.4.1 Dataset Description	16
2.4.2 Performance Evaluation	18
2.5 Concluding Remarks	22
2.6 Acknowledgments	23
Chapter 3 Gaze Dynamics, Modeling and Behavior Understanding	24
3.1 Introduction	24
3.2 Related Research	26
3.3 Naturalistic Driving Dataset	27
3.4 Modeling Scanpath for Driver Gaze Behavior Recognition	28
3.4.1 Temporal Feature Descriptor	29
3.4.2 Gaze Behavior Modeling	32
3.5 Experimental Design and Analysis	34
3.5.1 Event Description	34
3.5.2 Evaluation on Gaze Modeling	34
3.5.3 Discussion on Gaze Behavior Understanding	36
3.6 Concluding Remarks	37
3.7 Acknowledgments	38

Chapter 4	Driver Modeling by Multi-cue Fusion of Head, Eyes and Hands	39
	4.1 Introduction and Related Works	39
	4.2 Extraction, Representation and Fusion of Spatio-Temporal Features	42
	4.2.1 Head and Eye Features	42
	4.2.2 Hand Analysis	43
	4.2.3 Temporal Modeling and Feature Ranking	44
	4.3 Use Case: Stop-controlled Intersection	46
	4.3.1 Naturalistic Driving Dataset	46
	4.3.2 Event Description	46
	4.3.3 Data Driven Event Analysis	46
	4.4 Concluding Remarks	50
	4.5 Acknowledgments	51
Chapter 5	Naturalistic Driving Studies (NDS) database for algorithm benchmark and development	52
	5.1 Introduction	52
	5.2 Dataset Description	54
	5.2.1 Ground Truth Generation	55
	5.3 Metrics and Performance Evaluation	57
	5.4 Concluding Remarks	58
	5.5 Acknowledgments	60
Chapter 6	Balancing privacy and safety: Protecting driver identity in naturalistic driving video data	61
	6.1 Introduction	61
	6.2 DeIdentification Filter: Definition, Challenges and Related Research	63
	6.2.1 Looking inside the vehicle	63
	6.2.2 Related Studies	64
	6.3 How to Protect Identity and Preserve Gaze?	67
	6.3.1 Foreground Preservation for Gaze Estimation	68
	6.3.2 Background Distortion for Privacy Protection	69
	6.4 Experiment Design, Evaluation and Discussion	70
	6.4.1 Experiment Design	70
	6.4.2 Face Recognition	73
	6.4.3 Gaze Zone Estimation	75
	6.5 Concluding Remarks	79
	6.6 Acknowledgments	79
Chapter 7	Conclusions	80
Appendix A	Continuous Head Movement Estimator: Framework and On-Road Evaluations	83
	A.1 Introduction	83
	A.2 Related Research	85
	A.3 Issues and Challenges in Continuous and Robust Head Movement Analysis	89
	A.4 CoHMEt: Framework and Algorithms	90
	A.4.1 Facial Feature Detection and Tracking	91
	A.4.2 Pose Estimation	95
	A.4.3 Perspective Selection Procedure	96
	A.5 Experimental Evaluations and discussion	97
	A.5.1 Testbed and Dataset	98
	A.5.2 On-road Performance Evaluation	99

A.6	Concluding Remarks	103
A.7	Acknowledgments	104
	Bibliography	105

LIST OF FIGURES

Figure 1.1:	The backbone of intelligent vehicle is an intricately connected set of modules.	2
Figure 2.1:	Process for augmenting training dataset with synthetic data of varying (b) rotation, scale, (c) occlusion, (d) bright spot, (e) brightness, and (f) contrast-limited adaptive histogram equalization.	9
Figure 2.2:	Tracked facial feature/landmarks and their correspondences in 3D face image. Solid red circles are the points utilized for the head pose calculation.	11
Figure 2.3:	Eye ball image formulation: estimating β , gaze-angle with respect to head, from α , θ , d_1 and d_2 [100]	13
Figure 2.4:	Multi-perspective data collected during naturalistic on-road driving. Each row of images shows images are from a particular camera location and each column of images are time-synchronized.	14
Figure 2.5:	Perspective selection approach. Tracking phase utilizes head pose and dynamics to switch between perspectives, while a scoring criterion during a lost track re-initializes with the highest score camera.	15
Figure 2.6:	Illustration of gaze zones of interest and their approximate regions in the vehicle frame. Another gaze zone not illustrated but trained to classify is “Eyes Closed”.	16
Figure 2.7:	Sample instances of drivers looking at various regions: (a) looking forwards, (b) looking left, (c) looking at speedometer, (d) transitioning out of blink state. In realistic, on-road driving scenarios, subtle difference in the appearance of the eyes represent semantically different meanings.	17
Figure 2.8:	Performance evaluation on multiple camera framework with full scale face descriptor (i.e. head pose, horizontal gaze surrogate, vertical gaze surrogate and eye appearance descriptor).	19
Figure 2.9:	Performance evaluation on the subset of a full scale face descriptor (i.e. head pose, horizontal gaze surrogate, vertical gaze surrogate and eye appearance descriptor) with robust switching of multiple cameras.	21
Figure 2.10:	Performance evaluation on the number of gaze zone classes with the full scale face descriptor (i.e. head pose, horizontal gaze surrogate, vertical gaze surrogate and eye appearance descriptor) and with the multiple camera framework.	22
Figure 3.1:	Illustrates fourteen different scanpaths during a 20-second time window centered on lane change event, seven scanpaths during left lane change and seven scanpaths during right lane change event.	30
Figure 3.2:	Glance duration and transition decomposition analysis for (a) right lane change, (b) left lane change, (c) merge and (d) instrument cluster (averaged over multiple runs in the naturalistic driving scenarios).	33
Figure 3.3:	Confusion matrix, where rows are true classes and columns are predicted classes, from two experiments.	36
Figure 3.4:	Illustrates the fitness of the three models (i.e. <i>Left lane change</i> , <i>Right lane</i> , <i>Lane keep</i>) during left and right lane change maneuvers. Mean (solid line) and standard deviation (semitransparent shades) of the three models as applied to the lane change events described in Table 3.2 are shown.	37
Figure 4.1:	Sample instances of (a) Head and eye movements, (b) Hand positioning and (c) external context prior to the start of respective maneuvers (i.e. stop and right/left turns, stop and go straight).	40

Figure 4.2:	A two part figure illustrating an interesting coordination of head, eye and hand movements prior to the driver starting a right turn at the stop-controlled intersection.	41
Figure 4.3:	The left panel depicts the labels assigned to regions around the wheel. These are then used to map hand locations to the corresponding label values. The right panel shows an exemplar hand track sequences encoded using the method described.	43
Figure 4.4:	Geographical location of the stop-controlled intersections where data is extracted from and statistics on the events analyzed.	47
Figure 4.5:	A plot of the histogram of top 25 features with respect to the modality it originates from as a function of time (bottom panel). Exemplar image sequences extracted from different time intervals for each of the 3 maneuvers. Best viewed in color.	48
Figure 4.6:	A plot of relative frequency with which each feature occurs in the top 25 across all modalities and for all time intervals.	49
Figure 4.7:	A plot of the histogram of top 25 features with respect to the modality it originates from as a function of time for three groups of two-class problem: go straight vs. turn left (top row), turn left vs. turn right (middle row), go straight vs. turn right (bottom row)	50
Figure 5.1:	Face detection in general is challenging due to illumination, occlusion and unseen faces. However, detecting faces inside the automobile cabin has advantages such as sparse number of faces in any given image and correlated sets of images from a fixed perspective.	53
Figure 5.2:	Challenges in the dataset: varying illumination (a,b,c), head rotation away from frontal (b,d), occlusion (a,b,c), multiple faces (b). Realistic driving scenarios are prone to such volatile conditions.	55
Figure 5.3:	Example annotations from the VIVA-Face dataset. (Left) shows the ground truth face box derived from the 10 annotated points. (Right) shows an overlay of 3D general face model used to verify the quality of head pose estimation.	56
Figure 5.4:	VIVA-Faces Dataset Statistics: (a) Annotation bounding box size shows a good spread of different face sizes. (b) Number of faces by head pose in yaw rotation angle. Percentage of images where face is occluded with respect to (c) part type and (d) number of parts.	57
Figure 5.5:	Benchmark evaluations for face detection on the VIVA-Face dataset for (left) all 607 faces in 458 images, (middle) 323 non-occluded faces in 289 images, (right) 284 faces with at least one occluded part in 240 images.	59
Figure 6.1:	Illustration of privacy implications in using existing systems on raw camera sensory output to infer driver behavior: posture analysis [107], “keeping hands on the wheel” [104], and head dynamics [95].	62
Figure 6.2:	Illustrating three different de-identification filters, which semantically share the same goal of obscuring driver’s identity and preserving driver’s behavior, but in different degrees. As a decreasing number of driver’s face parts are preserved the less likely it is to identify the driver.	64
Figure 6.3:	Illustration of different combinations of patches around facial landmarks to estimate or predict driver behavior while protecting driver’s identity.	68
Figure 6.4:	Visual demonstration of de-identification by preserving the eyes in the foreground while scrambling or diffusing the context in the background. (a) Scrambling in the transform domain. (b) Anisotropic diffusion preserves edges and lines while smoothing out finer details in the image.	69

Figure 6.5:	Illustration of a de-identification method where region around the eyes are preserved and the background is replaced with black pixels. While pixel replacement for the background ensures more privacy, it removes context information often helpful in determining driver's gaze.	70
Figure 6.6:	Illustrates (a) the three gaze zone regions of interest: Left, Front, Right and (b) the five gaze zone regions of interest: Left, Front, Right, Rear Mirror, and Inside.	71
Figure 6.7:	Layout of the face recognition testing toolbox for user study. Given a de-identified image, participants choose one of the 12 candidates that best represents the driver in the de-identified image.	72
Figure 6.8:	Layout of gaze zone estimation tool box for user study. Given a de-identified image, participants choose one of following categories that best represents the driver's gaze: Left, Front, Right, Rear Mirror, Down, Unknown.	73
Figure 6.9:	Confusion matrix for the five gaze zone classification by participants of de-identified images with (a) one eye and (b) two eyes. On the average, gaze zone estimation is accurate 65% and 71% for de-identification with one-eye and with two-eyes, respectively.	76
Figure 6.10:	Five gaze zone performance of de-identified images with (a) one eye and (b) two eyes. Each element in this matrix of images is a cropped image of one of the highest participant response in respective categories.	77
Figure 6.11:	Illustrates multiple instances where typical confusion between gaze zones could occur. Each collage of images is comprised of three images: bottom is the raw image, top left and top right are cropped images of de-identification with one-eye and two-eyes, respectively.	77
Figure 6.12:	Three gaze zone performance of de-identified images with (a) one eye and (b) two eyes. Accuracy for Right gaze zone goes up significantly when Rear-Mirror gaze zone is considered to be part of Right gaze zone. Average accuracy for (a) one eye is 0.79 and (b) two eyes is 0.88.	78
Figure 6.13:	A sequence of de-identified images with respective raw images from a video sequence of the driver glancing at the rear view mirror. Gaze estimation of the de-identified images can be more accurate when provided the sequence leading up to it and possibly following it.	78
Figure A.1:	Head movements during a merge event. The 3D model of a head illustrates observed facial feature from a fixed camera perspective and self-occlusion induced by large head movements.	84
Figure A.2:	Multi-perspective data collected during naturalistic on-road driving. Each row of images shows images are from a particular camera location and each column of images are time-synchronized. Notice, challenges (e.g. external/self-occlusion, shadows, illumination change) present in real world data. . .	90
Figure A.3:	Illustration of the online learning process of estimating restricted face region in the image plane.	93
Figure A.4:	Process of reducing search space for video analysis using mixture of pictorial structures. Part space is reduced by constraining to region around the face location and mixture space is reduced by searching over neighboring mixture components around the estimated component from the previous frame. . . .	95
Figure A.5:	Tracked facial feature/landmarks and their correspondences in 3D face image. Solid red circles are the points utilized for the head pose calculation.	96
Figure A.6:	Perspective selection approach. Tracking phase utilizes head pose and dynamics to switch between perspectives, while a scoring criterion during a lost track re-initializes with the highest score camera.	97

Figure A.7: Illustration of multiple perspective framework on a segment taken from a subject’s naturalist on-road driving experiment. The plot shows the head scan by the driver from left to the right mirror starting from front pose. The evolution of the perspective selection is presented.	98
Figure A.8: LISA-A experimental testbed equipped with and capable of time synchronized capture of camera array and multiple Inertial Measurement Units (IMUS) [98].	99
Figure A.9: Distribution of head pose values in yaw rotation angle for an entire test drive (a) and accumulation of selected events (b) in the same test drive. Clearly, data from chosen events shows a more even distribution across yaw when compared to the entire drive.	100
Figure A.10: Shows the setup of the single-camera view, 2-camera view and 3-camera view as discussed and compared for performance evaluation of the multi-view framework.	101
Figure A.11: Error distribution with respect to the true head pose in yaw. The graphs reflects the first three error quartiles for single perspective (1st column), 2-camera perspective (2nd column) and 3-camera perspective (3rd column) using CLM+POS (1st row) and MPS+POS (2nd row)	102
Figure A.12: Quality of head pose estimation from individual camera view with respect to head orientations in the yaw angle. A useful means of configuring camera positions to maximize operational range of the overall system.	103

LIST OF TABLES

Table 2.1:	Naturalistic driving dataset description for evaluation of the end-to-end continuous gaze zone estimator	17
Table 3.1:	SAE identifies six levels of automation from “no automation” to “full automation” in order to provide common terminology for automated driving (for full table with definitions see [88]).	25
Table 3.2:	Dataset description of naturalistic driving for gaze modeling and behavior classification	28
Table 3.3:	The recall and precision of lane change prediction (averaged over multiple runs in naturalistic driving scenarios, 88 min each) via gaze behavior modeling using multivariate Gaussian.	35
Table 4.1:	Candidate features for temporal sequences	45
Table 5.1:	Comparison of Selected Studies in Literature on Face Analysis Methodologies (with emphasis on face detection) as Applied to Naturalistic Driving Data . .	54
Table 5.2:	Evaluation results from benchmark and submissions on yaw angle estimation of the head pose. The evaluation is split by occlusion levels: L0 (all 607 faces in 458 images), L1 (323 non-occluded faces in 289 images), L2 (284 faces with at least one occluded part in 240 images).	60
Table 6.1:	Comparison of selected studies in literature of de-identification of faces or people.	67
Table 6.2:	Face Recognition User Study: Evaluation of Participants’ Response	74
Table 6.3:	Gaze Zone Estimation User Study: Evaluation of Participants’ response . . .	76
Table A.1:	Selected studies on vision based head pose and dynamics estimation systems which are already tested or have potential to work in automobile environment.	86
Table A.2:	A list of events considered for evaluation, and its respective count and number of frames.	100
Table A.3:	On-road performance evaluations of the proposed CoHMET.	102

ACKNOWLEDGEMENTS

My deepest and sincerest thanks to my advisor Professor Mohan Trivedi. Through the years under your mentorship and guidance, I have learned invaluable lessons as a researcher. I thank him for all the opportunities, which at the time may have seemed impossible or daunting, but it has shaped me to be the proud researcher that I am. I'm always thankful for your encouraging words and treating me like a family member. I'll always aspire to part the kindness, patience and support you have showed me, to those around me.

I thank the committee members, Professor Garrison Cottrell, Professor David Kriegman, Professor Truong Nguyen and Professor Bhaskar Rao. Their teaching, questions, and guidance greatly influenced this body of research.

Words cannot express my thanks and gratitude to my family, dad, mom, Jee, Cesar, Adorae and Arielle. They have been with me through the best and lowest times. You have been my pillar of support, my inspiration, my joy, my happiness. I aspire to be better thanks to you.

I thank my labmates for their guidance, collaboration, and assistance over the years. In particular, I'd like to acknowledge Dr. Cuong Tran, Dr. Ashish Tawari, Mr. Eshed Ohn-Bar, Mr. Kevan Yuen, Dr. Ravi Satzoda, Ms. Jade Kwan, Mr. Akshay Ranges, Mr. Rakesh Nattoji, Mr. Sourabh Vora, Mr. Larry Ly, Mr. Sean Lee, Mr. Frankie Liu, Dr. Sayanan Sivaraman and Dr. Shiko Cheng for their help and contributions.

Thanks to all the friends I've met over the years, for all the good times we've had.

I thank the University of California Discovery Grant, Toyota Collaborative Safety Research Center, Audi AG, and Fujitsu Laboratories of America for funding this research.

Publication acknowledgments:

Chapter 2 is a partial reprint of materials published in IEEE Intelligent Transportation Systems Conference (2013), by Sujitha Martin, Ashish Tawari and Mohan M. Trivedi, in the IEEE Transactions on Intelligent Transportation Systems (2014), by Ashish Tawari, Sujitha Martin and Mohan M. Trivedi, and in IEEE International Conference on Pattern Recognition (2016), by Kevan Yuen, Sujitha Martin and Mohan M. Trivedi. The dissertation author was one of the primary investigator and author of these papers.

Chapter 3 is a partial reprint of material currently being prepared for submission for publication to the IEEE Transactions on Intelligent Vehicles, by Sujitha Martin, Sourabh Vora, Kevan Yuen and Mohan M. Trivedi. The dissertation author was the primary investigator and author of this paper.

Chapter 4 is in full a reprint of material that is published in the IEEE Intelligent Vehicles Symposium (2016), by Sujitha Martin, Akshay Ranges, Eshed Ohn-Bar and Mohan M. Trivedi, and a partial reprint of material published in the IAPR International Conference on Pattern Recognition (2016), by Sujitha Martin, Akshay Ranges, Eshed Ohn-Bar and Mohan M. Trivedi. The dissertation author was the primary investigator and author of these papers.

Chapter 5 is in full a reprint of material that is published in the IEEE Intelligent Vehicles Symposium (2016), by Sujitha Martin, Kevan Yuen and Mohan M. Trivedi. The dissertation author was the primary investigator and author of these papers.

Chapter 6 is in full a reprint of material that is published in the IEEE Transactions on Intelligent Transportation Systems (2014), by Sujitha Martin, Ashish Tawari and Mohan M. Trivedi. The dissertation author was the primary investigator and author of this paper.

Appendix A is in full a reprint of material that is published in the IEEE Transactions on Intelligent Transportation Systems (2014), by Ashish Tawari, Sujitha Martin and Mohan M. Trivedi. The dissertation author was one of the primary investigator and author of this paper.

VITA

2010	B. S. in Electrical Engineering, California Institute of Technology
2012	M. S. in Electrical Engineering (Intelligent Systems, Robotics, and Control), University of California, San Diego
2010-2016	Graduate Student Researcher, University of California, San Diego
2016	Ph. D. in Electrical Engineering (Intelligent Systems, Robotics, and Control), University of California, San Diego

PUBLICATIONS

Sujitha Martin, Akshay Rangesh, Eshed Ohn-Bar and Mohan M. Trivedi, "Preparatory Coordination of Head, Eyes and Hands: Experimental Study at Intersections," *International Conference on Pattern Recognition*, 2016.

Kevan Yuen, Sujitha Martin and Mohan M. Trivedi, "On Looking at Faces in an Automobile: Issues, Algorithms and Evaluation on Naturalistic Driving Dataset," *International Conference on Pattern Recognition*, 2016.

Borhan Vasli, Sujitha Martin Mohan M. Trivedi, "On Driver Gaze Estimation: Explorations and Fusion of Geometric and Data Driven Approaches," *IEEE Intelligent Transportation Systems Conference*, 2016.

Kevan Yuen, Sujitha Martin and Mohan M. Trivedi, "Looking at Faces in a Vehicle: A Deep CNN Based Approach and Evaluation," *IEEE Intelligent Transportation Systems Conference*, 2016.

Sujitha Martin, Kevan Yuen and Mohan M. Trivedi, "Vision for Intelligent Vehicles and Applications (VIVA): Face Detection and Head Pose Challenge," *IEEE Intelligent Vehicles Symposium*, 2016.

Sujitha Martin, Akshay Rangesh, Eshed Ohn-Bar and Mohan M. Trivedi, "The Rythms of Head, Eyes, and Hands at Intersections," *IEEE Intelligent Vehicles Symposium*, 2016.

Eshed Ohn-Bar, Ashish Tawari, Sujitha Martin and Mohan M. Trivedi, "On Surveillance for Safety Critical Events: In-Vehicle Video Networks for Predictive Driver Assistance Systems," *Computer Vision and Image Understanding*, 2015.

Sujitha Martin, Eshed Ohn-Bar and Mohan M. Trivedi, "Automatic Critical Event Extraction and Semantic Interpretation by Looking-Inside," *IEEE Conference on Intelligent Transportation Systems*, 2015.

Sujitha Martin, Ashish Tawari and Mohan M. Trivedi, "Towards Privacy Protecting Safety Systems for Naturalistic Driving Videos," *IEEE Transactions on Intelligent Transportation Systems*, 2014

Ashish Tawari, Sujitha Martin and Mohan M. Trivedi, "Continuous Head Movement Estimator (CoHMET) for Driver Assistance: Issues, Algorithms and On-Road Evaluations," *IEEE Transactions on Intelligent Transportation Systems*, 2014.

Ashish Tawari, Andreas Mgelmoose, Sujitha Martin, Thomas Moeslund and Mohan M. Trivedi, "Attention Estimation by Simultaneous Analysis of Viewer and View," *IEEE Intelligent Transportation Systems Conference*, 2014.

Sujitha Martin, Ashish Tawari and Mohan Trivedi, “Balancing privacy and safety: protecting driver identity in naturalistic driving video data,” *ACM SIGCHI International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 2014.

Eshed Ohn-Bar, Sujitha Martin, Ashish Tawari, and Mohan M. Trivedi, “Head, Eye, and Hand Patterns for Driver Activity Recognition,” *International Conference on Pattern Recognition*, August 2014.

Eshed Ohn-Bar, Ashish Tawari, Sujitha Martin, and Mohan M. Trivedi, “Vision on Wheels: Looking at Driver, Vehicle, and Surround for On-Road Maneuver Analysis,” *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014.

Sujitha Martin, Eshed Ohn-Bar, Ashish Tawari and Mohan M. Trivedi, “Understanding Head and Hand Activities and Coordination in Naturalistic Driving Videos,” *IEEE Intelligent Vehicles Symposium*, 2014.

Eshed Ohn-Bar, Ashish Tawari, Sujitha Martin and Mohan M. Trivedi, “Predicting Driver Maneuvers by Learning Holistic Features,” *IEEE Intelligent Vehicles Symposium*, 2014.

Sujitha Martin, Ashish Tawari and Mohan M. Trivedi, “Monitoring Head Dynamics for Driver Assistance: A Multiple Perspective Approach,” *IEEE Intelligent Transportation Systems Conference*, 2013.

Ravi Kumar Satzoda, Sujitha Martin, Minh Van Ly, Pujitha Gunaratne and Mohan M. Trivedi, “Towards Automated Drive Analysis: A Multimodal Synergistic Approach,” *IEEE Intelligent Transportation Systems Conference*, 2013.

Minh Van Ly, Sujitha Martin and Mohan M. Trivedi, “Driver Classification and Driver Style Recognition Using Inertial Sensors,” *IEEE Intelligent Vehicles Symposium*, 2013.

Sujitha Martin, Cuong Tran, Mohan M. Trivedi, “Optical flow based Head Movement and Gesture Analyzer (OHMeGA),” *International Conference on Pattern Recognition*, 2012.

Sujitha Martin, Ashish Tawari, Erik Murphy-Chutorian, Shinko Y. Cheng, Mohan Trivedi, “On the Design and Evaluation of Robust Head Pose for Visual User Interfaces: Algorithms, Databases, and Comparisons,” *ACM SIGCHI International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 2012.

Sujitha Martin, Cuong Tran, Ashish Tawari, Jade Kwan, and Mohan M. Trivedi, “Optical Flow Based Head Movement and Gesture Analysis in Automotive Environment,” *IEEE Intelligent Transportation Systems Conference*, 2012.

ABSTRACT OF THE DISSERTATION

Vision based, Multi-cue Driver Models for Intelligent Vehicles

by

Sujitha Martin

Doctor of Philosophy in Electrical Engineering
(Intelligent Systems, Robotics, and Control)

University of California, San Diego, 2016

Professor Mohan M. Trivedi, Chair

This dissertation seeks to enable intelligent vehicles to see, to predict intentions, to understand and to model the state of driver.

We developed a state of the art vision based non-contact gaze estimation framework by carefully designing submodules which will build up to achieve continuous and robust estimation. Key modules in this system include, face detection using deep convolutional neural networks, landmark estimation from cascaded regression models, head pose from geometrical correspondence mapping from 2-D points in the image plane to 3-D points in the head model, horizontal gaze surrogate based on geometrical formulation of the eye ball and iris position, vertical gaze surrogate based on openness of the upper eye lids and appearance descriptor, and finally, a 9-class gaze zone estimation from naturalistic driving data driven random forest algorithm.

We developed a framework to model driver's gaze behavior by representing the scanpath over a time period using glance durations and transition frequencies. As a use case, we explore the

driver’s scanpath patterns during maneuvers executed in freeway driving, namely, left lane change maneuver, right lane change maneuver and lane keep. It is shown that condensing temporal scanpath into glance durations and glance transition frequencies leads to recurring patterns based on driver activities. Furthermore, modeling these patterns show predictive powers in maneuver detection up to a few seconds a priori and show a promise for developing gaze guidance during take over requests in highly automated vehicles.

We introduce a framework to model the spatio-temporal movements of head, eyes and hands given naturalistic driving data of looking-in at the driver for any events or tasks performed of interest. As a use case, we explore the temporal coordination of the modalities on data of drivers executing maneuvers at stop-controlled intersections; the maneuvers executed are go straight, turn left and turn right. In sequentially increasing time windows, by training classifiers which have the ability to provide discriminative quality of its input variable, the experimental study at intersections shows which type of, when and how long distinguishable preparatory movements occur in the range of a few milliseconds to a few seconds.

We introduce one part of the Vision for Intelligent Vehicles and Applications (VIVA) challenge, namely, the VIVA-face challenge. VIVA is a platform designed to share naturalistic driving data with the community in order to: present issues and challenges in vision from real-world driving conditions, benchmark existing vision approaches using proper metrics and progress the development of future vision algorithms. With a special focus on challenges from looking inside at the drivers face, we provide information on how the data is acquired and annotated, and how methods are benchmarked, compared and shared on leaderboards.

Finally, we propose de-identification filters for protecting the privacy of drivers while preserving sufficient details to infer driver behavior, such as the gaze direction, in naturalistic driving videos. We implement and compare de-identification filters, which are made up of a combination of preserving eye regions and distorting the background, to show promising results. With such filters, researchers may be more inclined to publicly share deidentified naturalistic driving data. The research community can then tremendously benefit from large amounts of naturalistic driving data and focus on the analysis of human factors in the design and evaluation of intelligent vehicles.

Chapter 1

Introduction

Intelligent vehicles of the future are that which, having a holistic perception (i.e. inside, outside and of the vehicle) and understanding of the driving environment, make it possible for occupants to go from point A to point B safely, comfortably and in a timely manner. This may happen with the human driver in full control and getting active assistance from the robot, or the robot is in partial or full control and human drivers are passive observers ready to take over as deemed necessary by the machine or humans. Therefore, the future of intelligent vehicles lies in the collaboration of two intelligent systems, one robot and another human [74, 9]. Specifically, intelligent vehicle is an entity composed of intricately connected modules (see Figure 1.1) as listed here:

- **Holistic Sensing:** The driving environment is made up of many components including the driver, ego-vehicle, surround-vehicle, pedestrians, traffic lights, traffic signs, roads, vegetation, curbs, etc. For an intelligent vehicle to traverse safely through the dynamic environment, sensors are required to perceive the components and their characteristics; sensors such as camera array, radars, lidars, GPS, CAN bus. etc. Intelligent vehicles need to be instrumented with a suite of such sensors to allow holistic sensing, where holistic sensing represents sensing the state within the vehicle, around the vehicle and of the vehicle simultaneously and synchronously.
- **Semantic level Perception:** This stage derives semantically meaningful information from the output of sensors. For example, when a camera sensor outputs an image, a computer sees a matrix of numbers while humans see objects, their orientations, their interactions with other objects and may even predict the state of the object in near future, etc. Therefore, a critical milestone for intelligent vehicles is to extract semantic information from raw sensory data. Semantic information such as head pose, eye movement, hand position, foot gesture, headway distance, lateral position with respect to lane, etc. Extracting such semantic

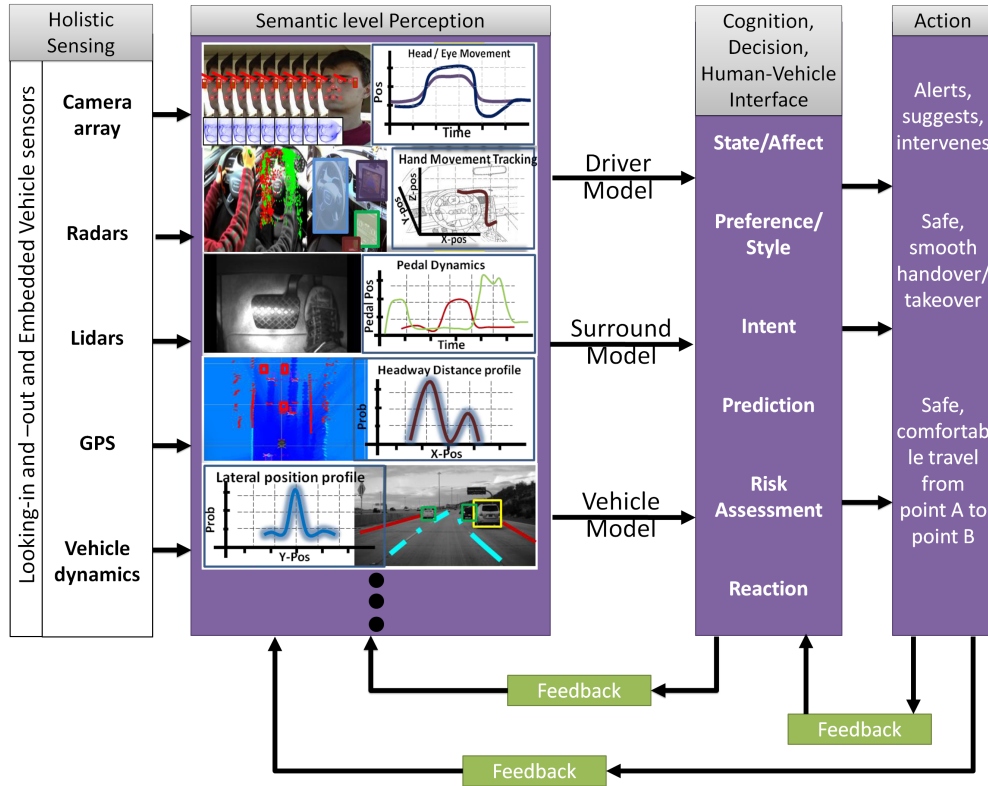


Figure 1.1: The backbone of intelligent vehicle is an intricately connected set of modules.

information requires developing computer vision techniques, machine learning approaches, large datasets with annotations and calibration methods, to name a few.

- Models:** Models are required to represent the state of an entity. There are three broad models of interest, driver models, surround models and vehicle models. For example, what kind of driver models will be useful? It will be useful to model driver's attention to the forward driving direction, level of alertness/fatigue, engagement in non-driving secondary tasks, "normal"/safe driving behavior, driving style, etc. For the surround environment, it will be useful to model trajectories of surrounding vehicles/pedestrians, lane deviations, traffic conditions, etc. Finally, for representing the state of the vehicle, it will be useful to model tire frictions, brake response, seat belt tension, air bag deployment, etc.
- Cognition, Decision, Human-Vehicle Interface:** Using models representing the state of the driver, surround and ego-vehicle, this stage encompasses everything short of taking an action. It includes predicting the action of driver and elements in the surrounding environment. For example, predicting whether a pedestrian will cross traffic, predicting whether a surrounding vehicle will change lanes, predicting whether the ego vehicle will make a turn, etc. Other modules include intent, risk assessment, reaction, state/affect,

preference/style, to name a few.

- **Action:** Whether the driver is control, some level of autonomy is used or in the transition state between driver and system, when there is an impending danger, an intelligent vehicle should take appropriate action in collaboration with the humans. For example, an intelligent vehicle should alert drivers to unseen objects (e.g. pedestrians, traffic signs), give suggestions (e.g. how to merge), and at times intervene (e.g. auto braking). Furthermore, even in the absence of impending danger, actions such safe, smooth handover/takeover and safe, comfortable travel from point A to point B is some among many actions necessary in intelligent vehicle.
- **Feedback:** Every intelligent system requires a feedback. For example, when the intelligent vehicle takes an action, how did the driver react? The driver’s reaction will factor into future actions and intentions of drivers. Another use of feedback is in managing computational resources in the semantic level perception based on risk assessment from the cognitive/decision module.

This dissertation focuses on looking-inside the vehicle at the driver, especially the driver’s face, and in so doing has contributions in the holistic sensing, semantic level perception, driver modeling and cognitive/decision logic modules. With respect to holistic sensing, we present a distributed camera framework for head movement analysis, with emphasis on the ability to robustly and continuously operate even during large head movements. The proposed system tracks facial features and analyzes their geometric configuration to estimate the head pose using a 3-D model. For the semantic level perception modules, we developed a state of the art vision based non-contact gaze estimation framework by carefully designing submodules which will build up to achieve continuous and robust estimation. Key modules in this system include, face detection using deep convolutional neural networks, landmark estimation from cascaded regression models, head pose from geometrical correspondence mapping from 2D points in the image plane to 3D points in the head model, horizontal gaze surrogate based on geometrical formulation of the eye ball and iris position, vertical gaze surrogate based on openness of the upper eye lids and appearance descriptor, and finally, a 9-class gaze zone estimation from naturalistic driving data driven random forest algorithm.

With respect to our contributions in the driver modeling and cognitive/decision logic module, we developed a framework to model driver’s gaze behavior by representing the scanpath over a time period using glance durations and transition frequency. As a use case, we explore the driver’s scanpath patterns during maneuvers executed in freeway driving, namely, left lane change maneuver, right lane change maneuver and lane keep. It is shown that condensing temporal scanpath into glance durations and glance transition frequencies leads to recurring patterns based on driver activities. Furthermore, modeling these patterns show predictive powers in maneuver

detection up to a few seconds a priori and show a promise for developing gaze guidance during take over requests in highly automated vehicles.

Furthermore, we introduce a framework to model the spatio-temporal movements of head, eyes and hands given naturalistic driving data of looking-in at the driver for any events or tasks performed of interest. As a use case, we explore the temporal coordination of the modalities on data of drivers executing maneuvers at stop-controlled intersections; the maneuvers executed are go straight, turn left and turn right. In sequentially increasing time windows, by training classifiers which have the ability to provide discriminative quality of its input variable, the experimental study at intersections shows which type of, when and how long distinguishable preparatory movements occur in the range of a few milliseconds to a few seconds.

All of these above described contributions and components rely heavily on vision based techniques. The question is, how well do these vision techniques work in order to be used in time and safety critical driving situations? Therefore, we introduce one part of the Vision for Intelligent Vehicles and Applications (VIVA) challenge, the VIVA-face challenge. VIVA is a platform designed to share naturalistic driving data with the community in order to: present issues and challenges in vision from real-world driving conditions, benchmark existing vision approaches using proper metrics and progress the development of future vision algorithms. With a special focus on challenges from looking inside at the drivers face, we provide information on how the data is acquired and annotated, and how methods are compared and shared using leaderboards.

Finally, we propose de-identification filters for protecting the privacy of drivers while preserving sufficient details to infer driver behavior, such as the gaze direction, in naturalistic driving videos. We implement and compare de-identification filters, which are made up of a combination of preserving eye regions and distorting the background, to show promising results. With such filters, researchers may be more inclined to publicly share deidentified naturalistic driving data. The research community can then tremendously benefit from large amounts of naturalistic driving data and focus on the analysis of human factors in the design and evaluation of intelligent vehicles.

Among the contributions of this dissertation:

- A state-of-the art machine vision based non-contact, end-to-end, continuous, gaze estimation framework and extensive evaluation on naturalistic driving data. (Chapter 2).
- Gaze pattern modeling for inferring driver behavior and predicting actions using large scale naturalistic driving data. (Chapter 3)
- A multi-cue, data driven framework to model the spatio-temporal movements of head, eyes and hands using computer vision. (Chapter 4)
- Naturalistic Driving Studies (NDS) database for vision algorithm benchmarking and devel-

opment. (Chapter 5)

- De-Identification filters for protecting the privacy of drivers while preserving sufficient details to infer driver behavior in naturalistic driving videos. (Chapter 6)
- Systematic cost-benefit analysis of various camera configurations of a distributed multiple camera framework to maximize system performance in terms of continuous head pose estimation. (Appendix)

Chapter 2

End-to-End, Continuous Gaze Estimation Framework

2.1 Introduction

A visually demanding driving environment, where elements surrounding a driver are constantly and rapidly changing, it is important for an intelligent vehicle to know where or at what the driver is looking. Vision-based systems are commonly used for deriving driver's gaze as they provide a non-contact and noninvasive solution. In literature, the driver's gaze has been approximated from head pose alone to head plus eye cues. At least four surrogate ways to represent the driver's gaze continuously are considered: head pose, horizontal eye-gaze surrogate, vertical eye-gaze surrogate and gaze zones. The last representation, gaze zone, builds upon the other modes of representation for a higher semantic level perception of where the driver is looking (e.g. forward, rear-view mirror, center stack). In later chapters, different subsets for representing the driver's gaze will be used for modeling the state of the driver. In particular, the next chapter will use the driver's gaze estimation approach for gaze behavior modeling and understanding, where as the chapter after that about multi-cue fusion with hand analysis will use a lesser semantic representative cues about the driver's gaze.

This chapter will describe modules which build up to achieving a continuous and (near) robust gaze zone estimation. Key modules in this system include, face detection using deep convolutional neural networks, landmark estimation from cascaded regression models, head pose from relative configuration of 2-D points in the image plane to 3-D points in the head model, horizontal gaze surrogate based on geometrical formulation of the eye ball and iris position, vertical gaze surrogate based on openness of the upper eye lids and appearance descriptor, and finally, a 9-class gaze zone estimation from naturalistic driving data driven random forest algorithm.

In developing the gaze estimation module and all the submodules within, there is a need for robustness and continuous monitoring of the driver state in time and safety critical, and constantly changing driving environment. However, the real-world, on-road driving environment presents a lot of challenging vision problems when observing the driver’s face. Therefore, specifically, any driver face analysis system, including a gaze estimator, for in vehicular application and development should have the following capabilities:

- **Automatic:** There should be no manual initialization, and the system should operate without any human intervention. This criterion precludes the use of pure-tracking approaches that measure the driver’s state relative to some initial configuration.
- **Fast:** The system must be able to derive driver state while driving, with real-time operation.
- **Wide operational range:** The system should be able to accurately and robustly handle spatially large and varying speed of head movements.
- **Lighting invariant:** The system must work in varying lighting conditions (e.g. sunny, cloudy).
- **Person invariant:** The system must work across different drivers.
- **Occlusion tolerant:** The system should work in the presence of typical partially occluding objects (e.g. eyewear, hats) or actions (e.g. hand movements).

Each of the submodules and the gaze estimator as a whole, as described in the following sections, are designed with the above listed capabilities in mind. The sensitivity of each module to the above listed criteria is either explicitly or implicitly discussed in respective sections. In particular, the sensitivity of the gaze estimator to each of the above mentioned capabilities are specifically addressed in the evaluation section towards the end of this chapter.

The performance evaluation in this chapter is given for the highest semantic level perception module only, which is the gaze zone estimator, and we refer the readers to appropriate publications, as mentioned in following sections, for performance evaluations of the submodules.

2.2 Related Research

Gaze zone estimation can be accomplished in one of two ways: first given an input image or video sequence, learn a black box (e.g. deep CNN) which outputs the gaze zone and second is in carefully designing the contents of the black box. The latter is the focus of our research for at least two reasons. First, with a black box it is hard to control what the system is learning especially with small dataset, and large datasets of sufficient variations for gaze zone estimation is difficult to not only to obtain but also to annotate. Second, when designing each of the semantic

modules leading up to the gaze zone estimator, these intermediate representations of gaze can be used for other studies. Studies such as lane change intent prediction [24] and overtake versus brake prediction [75] relies on a time window of head pose estimation, driver activity recognition [30] relies on both head pose and vertical eye gaze estimation, pedestrian recognition detection [96] relies on horizontal eye gaze, etc.

However, while many works have used intermediary signals for predictions and activity recognition, this does not diminish the importance of higher level semantic gaze zone labels. In a study on the effects of performing secondary tasks in a highly automated driving simulator [59], it was found that the frequency and duration of mirror-checking reduced during secondary task performance versus normal, baseline driving. Alternatively, Ahlstrom et al. [3] developed a rule based 2-second ‘attention buffer’ framework which depleted when the driver looked away from the field relevant to driving (FRD); and it starts filling up when the gaze direction is redirected toward FRD. In such a framework, parameters such as rate at which the depletion occurs and whether the depletion should be delayed, etc. require higher level semantic labels as described in this chapter.

Therefore, the end-to-end continuous gaze zone estimator described in this chapter is made up of many submodules. Some of these modules, such as the face detector [121], continuous head pose estimator [95, 65], and vertical and horizontal eye gaze surrogate [100], have been developed and tested for robustness to the challenging driving environment. Other modules, such as the facial landmark estimator [12, 117], have been incorporated without the full fledged evaluation on driving data but it’s quantitative performance in non-driving datasets and visually inspected performance in driving datasets are promising for in-vehicular application. Currently, there is on going research on developing a deep CNN based facial landmark estimator to not only handle the harsh lighting conditions and occlusion seen in driving conditions but also to provide information on which landmarks, if at all, are occluded [120].

In literature, works on gaze zone estimation is relatively new and of those, there are two categories: first is the geometric method and second is learning based method. The work presented in [3] estimates gaze zones based on geometric methods where a 3-D model car is divided into different zones and 3D gaze tracking is used to classify into gaze zones; however, no evaluations on gaze zone level is given in [3]. Another geometric based method is presented in [111], but the number of gaze zones estimated is very limited (i.e. on-road versus off-road) and evaluations are conducted in stationary vehicles. In terms of learning based methods, there are two prevalent works. Work by Tawari et al. [100] has the most similarity to the work presented in this chapter in terms of the features selected (e.g. head pose, vertical gaze surrogate and horizontal gaze surrogate), classifier used (i.e. random forest) and evaluation on naturalistic driving data. The differences are this work introduces another feature to better represent the eyes (i.e. appearance descriptor), has an increased number of gaze zones and performs evaluations on a larger dataset. Another learning based method is the work presented by Fridman et al. [40, 39]

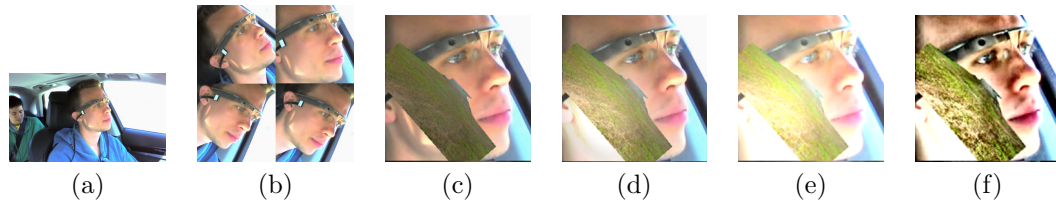


Figure 2.1: Process for augmenting training dataset with synthetic data of varying (b) rotation, scale, (c) occlusion, (d) bright spot, (e) brightness, and (f) contrast-limited adaptive histogram equalization. Since the images from AFLW cannot be published in this article, (a) an image from the VIVA-Face test set is used to illustrate the process.

where the evaluations are commendably done on a large dataset but the design of the features to represent the state of the head and eyes are what is causing their classifier to overfit to user based models and causes sharp decrease in performance for global based models; furthermore, the largest number of gaze zones defined and trained for is six.

To sum up the differences to previous works in literature, the gaze estimator described in this chapter is purely learning based, is developed to classify gaze into nine zones, is designed to better represent the state of the eye and is evaluated on naturalistic driving data.

2.3 Face Analysis: Algorithm Development

2.3.1 Face Detection

The study and formulation of face detection systems has populated literature for many decades. Largely, the objective of face detection systems have been to find faces in arbitrary images with minimal to no leverage on contextual information.

In the past decade, Viola-Jones’s adaboost casacade with haar features [112] and deformable parts model [122] have dominated the field of face detection. The Adaboost classifier has especially been a favorite in the community of intelligent vehicles not only because of its real-time processing speed but also because of its implementation in OpenCV which allows for easy training and testing. Works as early as [10] and as late as [63] have used the boosted cascade with haar feature in order to monitor driver vigilance and generate drive reports, respectively. Viola-Jone’s face detector has also been a key stepping stone in estimating driver’s head pose [72], predicting driver’s intent to change lane [24], monitoring driver’s alertness [78] and estimating driver’s gaze [100, 111]. Evaluations of this face detector for faces from naturalistic driving in these selected publications, however, is negligible

Benchmarking competitions like FFDB [53] and VIVA-face-off [64] give more in-depth and comparative results on state-of-the-art face detection schemes on general and driver-specific domain, respectively. These benchmarks show that Viola-Jones face detector is not sufficient for the robust performance required in driver assistance. Deformable parts model, comparatively, shows better performance, but at the expense of computational time (i.e. the conventional DPM

executes in the order of seconds).

In the computer vision field, there has been a rise in the use of deep convolutional neural networks (DCNN) for training face detectors with top performing results such as DDFD [32], Faceness [118], and CascadeCNN [58]. In DDFD, they fine-tune AlexNet [56] to train a face detector by generating a score image and extracting detected faces. Faceness trains multiple DCNN components tuned for individual components of the face (e.g. eyes, nose, mouth, etc.) and combines the output responses to detect a face. CascadeCNN focuses on cascading multiple CNNs for faster run times by rejecting background areas in the earlier stages of the detector. While many studies have developed CNN based face detection for general application, this study emphasizes the need for learning a model which is keen on challenges and advantages unique to in-vehicular space.

This section gives a brief insight and description of our deep convolutional neural network using the AlexNet network structure to detect faces (similar to DDFD). For the in vehicle domain for driving safety where there are harsh lighting conditions and occlusions, it is critical that the system is able to continuously monitor the driver’s face and behavior even under these situations. Therefore, one of the key contributions of this module is heavily augmenting the training samples to include more examples of faces under harsh lighting and occlusion to explore improvements for handling these scenarios. More specifically, the training data augmentation includes window sampling, superimposing objects from the SUN2012 dataset to occlude different parts of the face, random bright spots on the face to mimic the effects of sunlight, brightness variation and contrast-limited adaptive histogram equalization (CLAHE). Illustrative examples of the augmented dataset is in given Figure 2.1 and more details on the algorithm development and training dataset description can be found in [121].

2.3.2 Landmark Estimation

A total of 68 landmarks as defined originally in the CMU Multi-Pie dataset [45] and 2 iris locations [100] are estimated in the current framework. In this module, the landmarks are estimated using a cascade of regression models as described in [12, 117] with more details for iris localization given in [100]. The idea is given an initial estimate of the facial landmark locations, say p_0 , which is generally a mean shape, and a learned sequence of regression models (R_1, \dots, R_K), facial landmark location at the k th iteration is computed as follows:

$$p_k = p_{k-1} + R_k * F(p_{k-1}, M)$$

where $k \in [1, K]$, M represents the image on which the landmarks are being estimated, and $F(p_{k-1}, M)$ is a vector of features extracted at landmark positions found at the $(k - 1)$ iteration on image M .

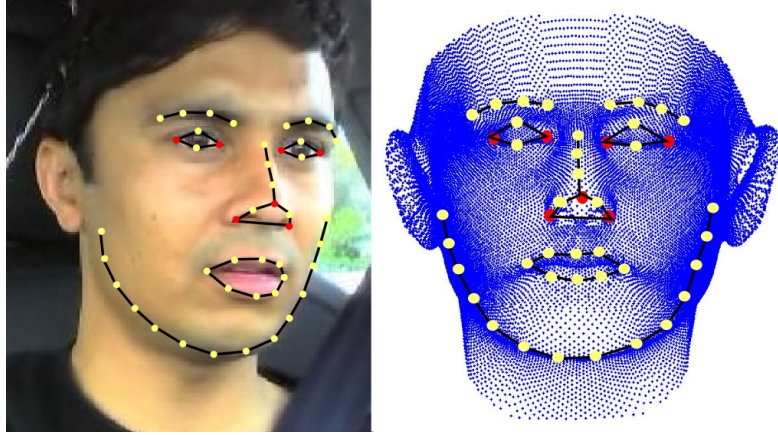


Figure 2.2: Tracked facial feature/landmarks and their correspondences in 3D face image. Solid red circles are the points utilized for the head pose calculation.

2.3.3 Head Pose

Given a 3-D model of an object, the pose from orthography and scaling (POS) [20] finds the position and orientation of the camera coordinate with respect to the object reference frame. It minimizes the reprojection error using a weak perspective transform. Given a point on a 3-D model, e.g., M_i , and its measured projection in the image plane, e.g., $p_i = (x_i, y_i)$, POS solves the following linear system of equations:

$$M_0 M_i \cdot \alpha \mathbf{i} = x_0 x_i \quad i = 1 \cdots N_c$$

$$M_0 M_i \cdot \alpha \mathbf{j} = y_0 y_i \quad i = 1 \cdots N_c$$

where $M_0 M_i$ represents the vector from the reference point on the 3D model M_0 to M_i , α is the scale factor associated with the weak perspective projection, and N_c is the number of $3D - 2D$ point correspondences. Vectors \mathbf{i} and \mathbf{j} form the first two rows of the rotation matrix, and the third row is given by vector $\mathbf{k} = \mathbf{i} \times \mathbf{j}$, which is a cross product. Note, however, that, although \mathbf{k} is perpendicular to \mathbf{i} and \mathbf{j} , vectors \mathbf{i} and \mathbf{j} are not necessarily perpendicular due to noisy $3D - 2D$ point correspondences. Therefore, the rotation matrix is projected into the $SO(3)$ space by normalizing the magnitude of the eigenvalues.

To solve this system of equations, POS requires at least four points of correspondences in general positions. The cascaded regression algorithm (as described earlier), however, are trained to output more than four fiducial points. In our current implementation, we use the following fiducial points, i.e., four eye corners, two nose corners, and a nose tip, as they are comparatively less deformable. Figure 2.2 shows these points (the red solid circle) on a test image and its corresponding points on the 3-D mean face model.

From the obtained rotation matrix, pitch (ψ), yaw (θ) and roll (ϕ) rotation angles in the Euler coordinate system are computed as follows [93]. We assume that

$$\mathbf{R} = \mathbf{R}_z(\phi)\mathbf{R}_y(\theta)\mathbf{R}_x(\psi) = \begin{bmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{bmatrix}$$

and compute

$$\theta = -\sin^{-1}(R_{31})$$

$$\psi = \tan^{-1}(R_{32}/R_{33})$$

$$\phi = \tan^{-1}(R_{21}/R_{11})$$

2.3.4 Eye Cues

Given landmarks around the eye, including iris centers, we compute two gaze-surrogate measurements, one to account for horizontal eye gaze movement and another for vertical eye gaze movement, and one appearance measurement. The first two measurements are more quantitative and objective, while the latter is less quantitative and more subjective at this stage. For reasons more clearly illustrated and described in the gaze zone estimation subsection, horizontal gaze surrogate and vertical gaze surrogate are insufficient descriptors to represent the state of the eyes and therefore insufficient to distinguish between certain zones of interest (e.g. speedometer vs. closing eyes).

Horizontal gaze-direction computation is explained in a work by Tawari et al. [100], but we describe mathematical details here for the sake of completion and the readers convenience. The horizontal gaze-direction β with respect to head, see Figure 2.3, is estimated as a function of α , angle subtended by an eye in horizontal direction, head-pose (yaw) angle θ with respect to the image plane, and $\frac{d_1}{d_2}$, the ratio of the distances of iris center from the detected corner of the eyes in the image plane. The equations below show the calculation steps:

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

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

Vertical gaze-direction with respect to head is inferred from the eye area. The eye area is computed using only upper-eyelid contour. Unlike the horizontal gaze-direction measure, this way of computing vertical gaze is sensitive to head pose; meaning eye area computed when driver

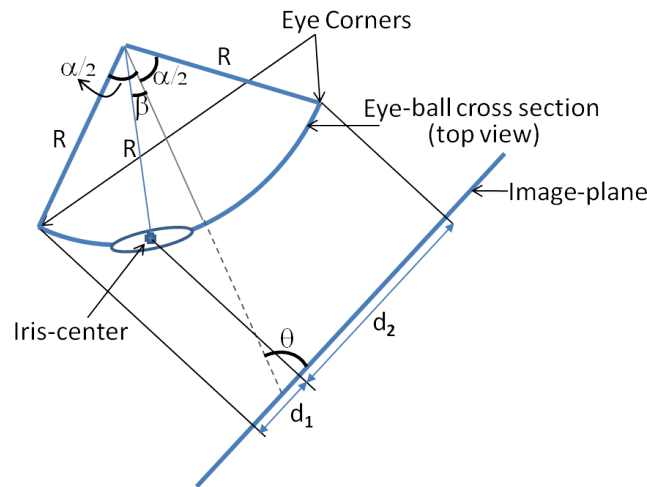


Figure 2.3: Eye ball image formulation: estimating β , gaze-angle with respect to head, from α , θ , d_1 and d_2 [100]

has his eyes fully open but the head is turned away from camera is similar in value to when driver is facing toward camera but closing his or her eyes. Additionally, by using raw upper eye lid area, this measure is sensitive to driver ethnicity and driver seating position.

The last measurement representing eyes cues is the appearance. In this work, appearance of the eye is represented by computing HoG (Histogram of Gradients) in a 2-by-2 patch around the eye. This descriptor can not only capture the appearance of the eye, which cannot be represented by the horizontal and vertical gaze surrogates described above, it has the potential to substitute the gaze surrogate measures. However, if the training data is not sufficiently diverse, it may have the adverse affect of learning unintended attributes (e.g. driver specific). The evaluation section covers the importance of each of the head and eyes cues in detail.

2.3.5 Multiple Camera Framework

Many existing state of the art face analysis systems, explicitly or implicitly, rely on a portion of the face to be visible in the image plane to estimate head pose. This means that even during large head movements, algorithms require the visibility of facial features to continuously track the state of the head. With a single perspective of the driver's head, however, large spatial head movements induce self-occlusions of facial features as illustrated in the first two columns of Figure 2.4. In Figure 2.4, each row of images are taken from a different camera perspective and each column of images are time-synchronized. Clearly, the availability of multiple perspectives decreases the severity of self-occlusions at any instant in time which translates to an increase in the robustness of continuous head tracking.

Occlusions of facial features can also occur due to external objects (e.g. hand movements near the face region, sunglasses). Depending on camera perspective, hand movements on steering



Figure 2.4: Multi-perspective data collected during naturalistic on-road driving. Each row of images shows images are from a particular camera location and each column of images are time-synchronized. Locations of the camera: Camera 1 is near the left A-pillar, Camera 2 is close to the dashboard, and Camera 3 is near the rear-view mirror. Notice, challenges (e.g. external-/self-occlusion, shadows, illumination change) present in real world data.

wheel during vehicle maneuvers, to adjust sunshade, to point, etc. can cause occlusion. The middle two columns in Figure 2.4 show examples of the latter two scenarios with hand movements. Effects of lighting conditions are also highly dependent on the camera location. In Figure 2.4, the last two columns illustrate the effects of lighting conditions. Therefore, a multiple perspective approach with suitable camera placements, can mitigate the adverse effect of any one camera perspective being unreliable to track the head.

For this, we propose a distributed camera framework inside the vehicle cockpit. The CoHMEt (Continuous Head Movement Estimator) framework treats each camera perspective independently, and a perspective selection procedure provides the chosen perspective by analyzing temporal dynamics and the current quality of the estimated head pose in each perspective. The block diagram in Figure 2.5 illustrates this process for a general setup of N cameras, where the cameras are numbered in the increasing order from the leftmost position in the distributed camera array setup. In the proposed system, we utilize two cameras, which are positioned one each to the left and right of the driver. The system is initialized with the right camera and during the tracking phase, transitions from one perspective to another is allowed based on the operating range $(\Omega_-^{N_c}, \Omega_+^{N_c})$ of the selected camera and yaw movement direction. When tracking is lost, due to either loss of facial point detection or rejection of the estimated points, re-initialization is performed using a scoring criteria. More detail on CoHMEt and systematic cost-benefit analysis of various camera configurations to maximize system performance in terms of continuous head pose estimation can be found in Appendix A.

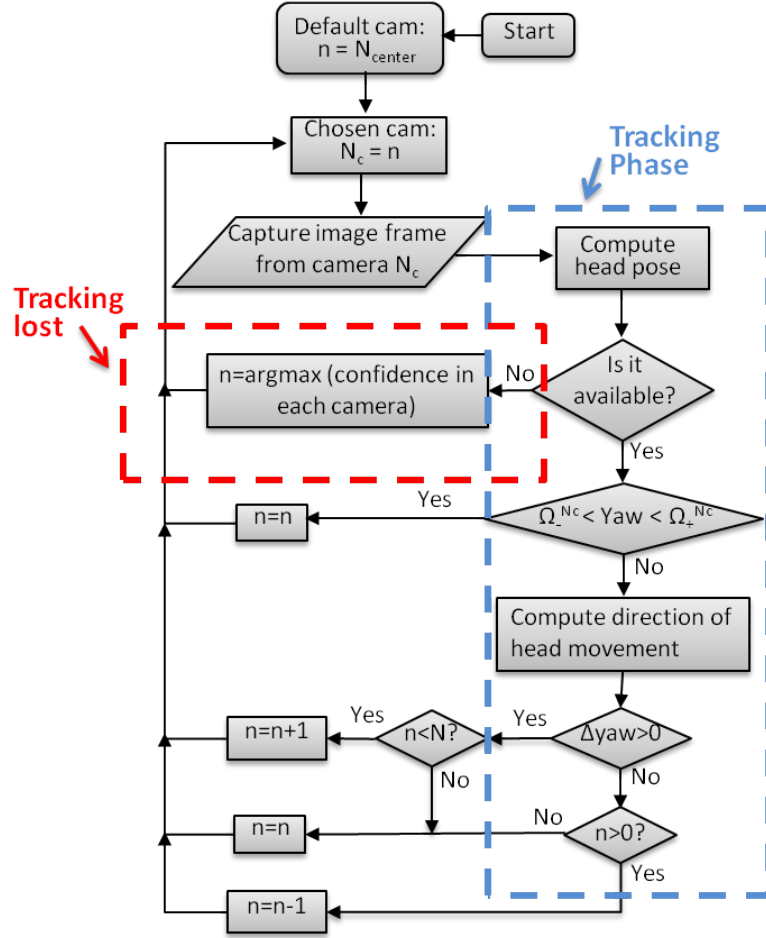


Figure 2.5: Perspective selection approach. Tracking phase utilizes head pose and dynamics to switch between perspectives, while a scoring criterion during a lost track re-initializes with the highest score camera.

2.3.6 Gaze Zone Estimation

In literature, driver’s gaze has been represented and interpreted in many ways. For example, driver’s gaze can be approximated using head pose and eye gaze surrogates features, as described in respective sections above. In this section, however, we develop an approach to give a more semantic understanding of where or at what the driver is looking. Eight semantic gaze-zones of interest are, *far left*, *left*, *front*, *speedometer*, *rear view*, *center stack*, *front right* and *right*, as illustrated in Figure 2.6. Another class of interest, but not illustrated in the figure, is the state of eyes closed. Therefore, in total, the gaze estimator in this study will learn to classify between 9-classes of interest.

Consider a set of feature vectors $\vec{F} = \{\vec{f}_1, \vec{f}_2, \dots, \vec{f}_N\}$, and their corresponding class labels $Y = \{y_1, y_2, \dots, y_N\}$, for N sample instances. In the context of our study, the class labels are one of the nine gaze zones of interest, as illustrated in Figure 2.6, and the feature vector is

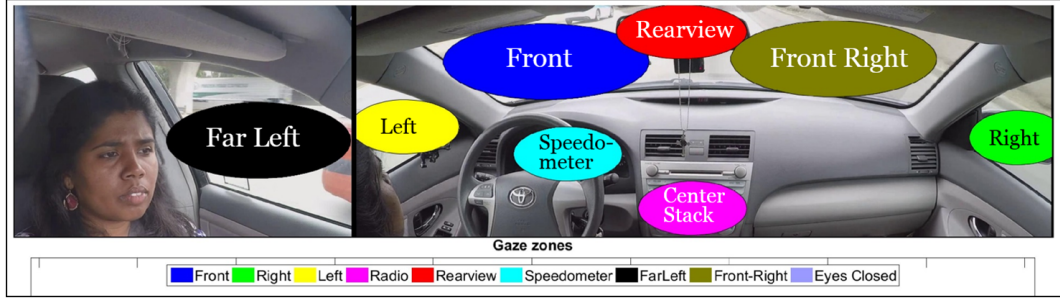


Figure 2.6: Illustration of gaze zones of interest and their approximate regions in the vehicle frame. Another gaze zone not illustrated but trained to classify is “Eyes Closed”.

a subset of head and eyes cues described in the previous sections. Given \vec{F} and Y , we train a random forest (RF) with an ensemble of 1000 decision trees on the entire corpus. The maximum depth of each tree is restricted to 10 to prevent overfitting. The advantage in choosing RFs over other feature selection techniques is three fold: First, we sidestep the entire process of tuning hyper-parameters through cross-validation. Second, RFs make no assumption about the linear separability of the data. Third, RFs are capable of handling different data formats like integers, floats or labels. Thus, no standardization or regularization is necessary.

The creativity and sensitivity of this framework is in the design of the feature vector \vec{f} . If head pose in pitch, yaw, and roll alone are represented in the feature vector, all instances illustrated in Figure 2.7 would be classified as *front*. If head pose plus horizontal gaze surrogate make up the feature vector, instances in Figure 2.7a and Figure 2.7b will be properly classified as *front* and *left*, respectively, but instance in Figure 2.7c is equally likely to be classified as any of the following classes: *front*, *center stack*, *front right*, *speedometer*. An addition of the vertical gaze surrogate to the feature vector will correctly classify the instance in Figure 2.7c to be *speedometer*. However, head pose, horizontal gaze surrogate and vertical gaze surrogate are insufficient descriptors to classify the instance in Figure 2.7d as *eyes closed* instead of *speedometer*. Therefore, the appearance of the eyes, as represented by HoG descriptor, is appended in the feature vector, \vec{f} . While the best performance of the gaze estimator is when all features are reliably extractable (e.g. no occlusion of the eyes, proper estimation of the landmarks), the framework has flexibility to balance the availability of information with the resolution of gaze zones.

2.4 Face Analysis: Evaluation on Naturalistic Driving Data

2.4.1 Dataset Description

Data are collected from naturalistic and on-road driving using two different subject-owned vehicles. The vehicle is instrumented with four cameras, two of which are mounted facing



Figure 2.7: Sample instances of drivers looking at various regions: (a) looking forwards, (b) looking left, (c) looking at speedometer, (d) transitioning out of blink state. In realistic, on-road driving scenarios, subtle difference in the appearance of the eyes represent semantically different meanings.

Table 2.1: Naturalistic driving dataset description for evaluation of the end-to-end continuous gaze zone estimator

Gaze-zones	Trip No.			
	1	2	3	All
Front	5183	3280	1710	10712
Right	426	419	246	1091
Left	645	479	644	1768
Center Stack	392	214	288	894
Rear view	644	392	344	1400
Speedometer	448	480	253	1181
Far Left	423	282	66	771
Front Right	722	421	109	1252
Eyes closed	335	369	207	911
All	9238	6876	3867	19987

the driver, i.e., one camera near the A pillar on the side window, and one camera near the rear-view mirror; the other two cameras are one for looking out and one for looking at the hands. The camera suite is time synchronized and capture their respective views in a color video stream at 30 frames/s and at a resolution of 1980×1240 pixels. Data for evaluation is meticulously gathered from three different drives. Annotation of ground truth of which zone the driver is looking at is done by visually inspecting the temporal sequence of inside and outside around the instance under consideration. Table 2.1 shows distribution of annotated instances for each of the gaze zones of interest with respect to each drive and overall. The evaluations reported in the following section will be on this dataset, where the portions for training and testing are given in more detail below depending on the criterion for testing.

2.4.2 Performance Evaluation

Evaluation for each of the following experiments are given in three forms. First is the 9-class confusion matrix where each row represents true gaze and each column represents estimated gaze. The last column in the confusion matrix represents the *unknown* class, which occurs when submodules (i.e., face detection, landmark estimation) do not produce a reliable output. Other two forms of evaluation metrics are unweighted accuracy and weighted accuracy:

$$\text{Unweighted Accuracy} = \frac{\sum_{i \in \{\text{gaze zones}\}} (\text{True Positive})_i}{\sum_{i \in \{\text{gaze zones}\}} (\text{Total population})_i}$$

$$\text{Weighted Accuracy} = \frac{1}{\text{No. of Gaze Zones}} \times \left\{ \sum_{i \in \{\text{gaze zones}\}} \frac{(\text{True Positive})_i}{(\text{Total Population})_i} \right\}$$

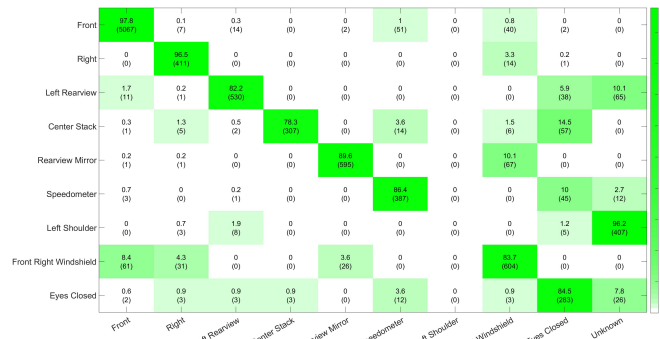
In the following sections, discussions are based on individual class accuracies and weighted accuracy, and other accuracies are reported for the sake of completion. Weighted accuracy is preferred over unweighted accuracy in the discussion due to the unbalance nature of the dataset.

Multiple Camera Framework

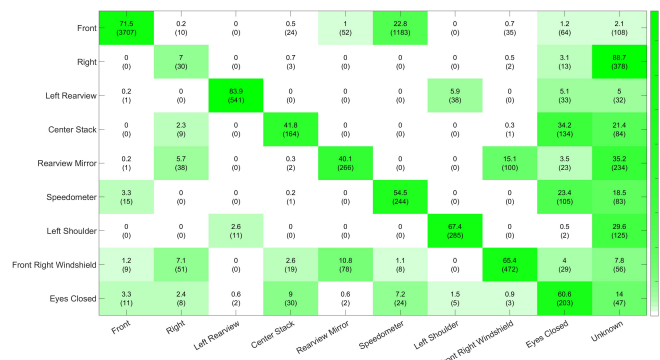
One of the criteria initially stated as a desired property of a robust and continuous gaze estimator, is its wide operational range. In this section, performance evaluation of our end-to-end gaze estimator is conducted to show the indispensable nature of the multiple camera framework as well as sensitivity of the system to camera placement/perspective. Our multiple camera framework considers two perspectives, as described in the data description section, one observing the driver from the left and another from the right.

In this section, we conduct three experiments to estimate drivers gaze zone. All experiments are trained on data taken from Trip no. 2 and tested on Trip no. 1. In the first experiment, the classifier is trained and tested on Right Perspective (i.e., from camera placed right of the driver). In the second experiment, the classifier is trained and tested on Left Perspective (i.e., from camera placed left of the driver). In the third experiment, classifiers are trained for both perspectives but during testing the perspective selection algorithm determines the optimal perspective and therefore determines the class based on classifier trained for that perspective. Confusion matrix for each of these experiments is shown in Figure 2.8.

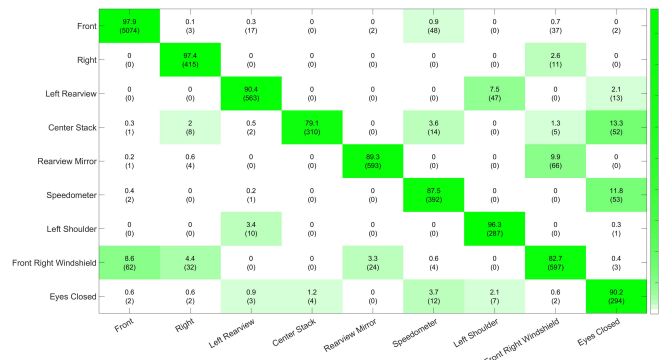
From the weighted accuracy alone, it shows that the placement of the cameras has a significant effect on the overall accuracy. This is likely because the Right Perspective is more spatially centered for majority of the gaze zones. On the other hand, switching between perspectives at optimal times derived by the perspective selection algorithm gives the highest weighted accuracy of 94%, which is a 5% improvement from Right Perspective alone and a 30% improve-



(a) Right Perspective only gives weighted accuracy of 78% and unweighted accuracy of 89%.



(b) Left Perspective only gives weighted accuracy of 55% and unweighted accuracy of 64%.



(c) Switching between both perspectives gives weighted accuracy of 90% and unweighted accuracy of 94%.

Figure 2.8: Performance evaluation on multiple camera framework with full scale face descriptor (i.e. head pose, horizontal gaze surrogate, vertical gaze surrogate and eye appearance descriptor).

ment from Left Perspective alone. Comparing the confusion matrix of the best perspective (i.e., Right Perspective) to switching between two perspectives, the improvement in accuracy comes from gaze zones furthest from the camera perspective, *Left* and *Far Left*. Full transcript of

the Continuous Head Movement Estimator (CoHMEt) with extensive evaluation on naturalistic driving data can be found in Appendix A.

Subset of Face Descriptors

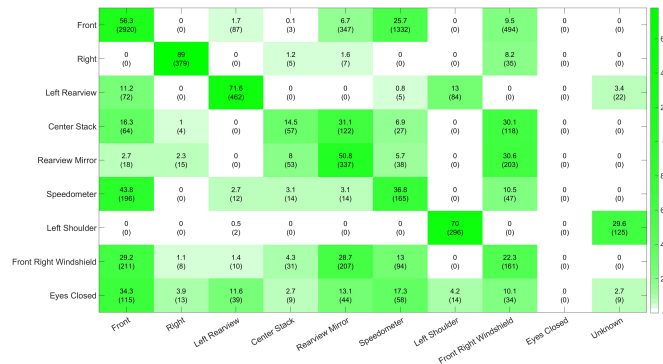
Another one of the criteria initially stated as a desired property of a robust and continuous gaze estimator, is the operability of the system under varying lighting conditions and occlusions. In this section, performance evaluation of our end-to-end gaze estimator is conducted to show effects on performance based on availability of information. For instance, what is the accuracy when eyes are occluded (e.g. sunglasses) and only head pose is available?

Therefore, in this section, we conduct three experiments to test the sensitivity of the system as a function of available information. All experiments are trained on data taken from Trip no. 2 and tested on Trip no. 1. First experiment is with head pose information alone, second experiment is with both head and partial eye cues (horizontal gaze-direction and vertical gaze direction), and third experiment is with both head and full eye cues (horizontal gaze-direction and appearance of the eye).

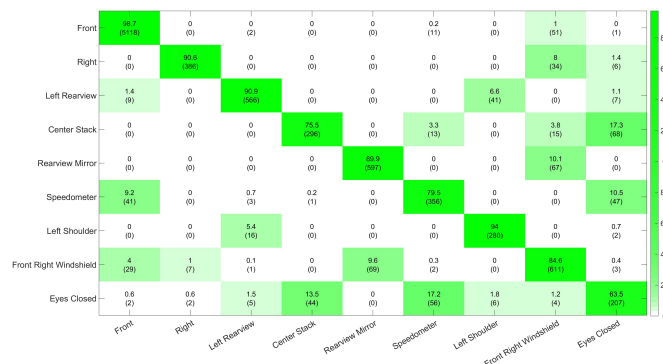
From weighted accuracy alone, it shows that there is a significant improvement when using any information representative of the eyes in addition to the head cues; an almost 40% improvement when augmenting with gaze surrogate features and an additional 5% improvement when augmenting with appearance descriptor. This observation can be further explored in the confusion matrix (see Figure 2.9) where neighboring gaze zones requiring more eye movements than head movements are often confused (e.g. front vs. speedometer, front right windshield vs. rear view mirror). When comparing the confusion matrix for including partial (Figure 2.9b) versus full eye descriptor cues (Figure 2.9c), predictably the improvements are in gaze zones which are spatially situated inside the vehicle (i.e. *Speedometer*, *Center Stack* and *Eyes Closed*).

Number and Spatial Distribution of Gaze Zones

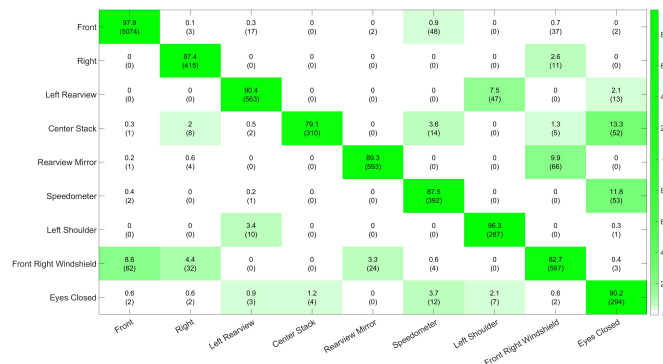
In this section, the performance evaluation deals with accuracies as a function of resolution in gaze zones. The intention of exploring different resolutions in gaze zones is based on the observation that different gaze related application require different information. So far, this chapter has considered estimating gaze-zone as one of 9-classes and has shown to achieve a weighted accuracy of 90%. What if the sole purpose of this gaze estimator was to determine whether the driver is looking *on-road* or *off-road*. For such instances, our end-to-end gaze estimator achieves a weighted accuracy of 98% (see Figure 2.10a). Here on-road encompasses *front* and *front right windshield* and off-road contains the other 7 classes. Another organization of gaze-zones is the spatial locations with respect to the car: *front*, *left*, *right*, *back* and *inside*. Such a distribution of gaze zone results in a weighted accuracy of 96% (see Figure 2.10b). Therefore, depending on system requirement, the number and spatial distribution of the gaze zones can be adjusted for



(a) Head pose only gives weighted accuracy of 46% and unweighted accuracy of 52%.



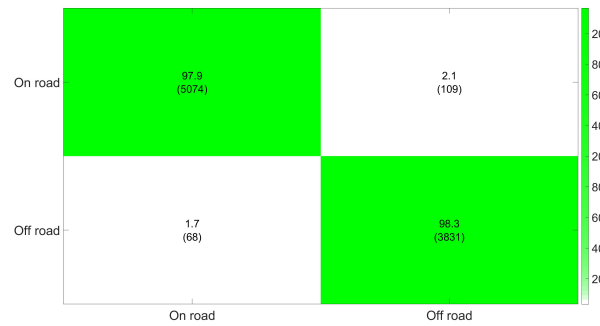
(b) Head pose plus horizontal and vertical gaze surrogate gives weighted accuracy of 85% and unweighted accuracy of 93%.



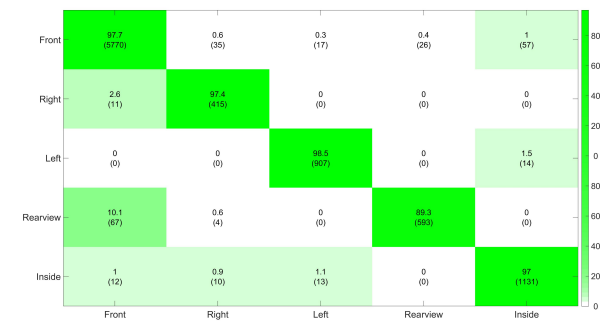
(c) Head pose plus horizontal gaze surrogate plus appearance descriptor gives weighted accuracy of 90% and unweighted accuracy of 94%.

Figure 2.9: Performance evaluation on the subset of a full scale face descriptor (i.e. head pose, horizontal gaze surrogate, vertical gaze surrogate and eye appearance descriptor) with robust switching of multiple cameras.

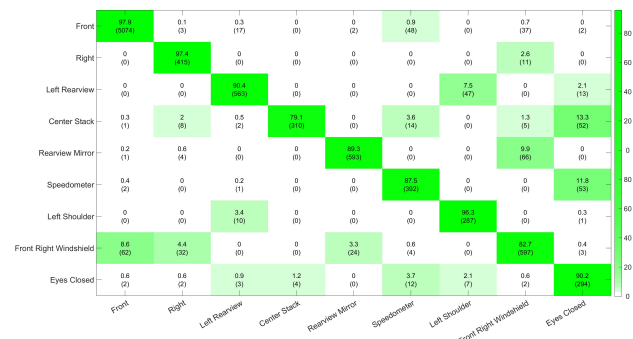
higher accuracy at the expense of lower gaze resolution or higher gaze resolution at the expense of reduced accuracy.



(a) A two class gaze zone estimator gives weighted and unweighted accuracy of 98%.



(b) A five class gaze zone estimator gives weighted accuracy of 96% and unweighted accuracy of 97%.



(c) A nine class gaze zone estimator gives weighted accuracy of 90% and unweighted accuracy of 94%.

Figure 2.10: Performance evaluation on the number of gaze zone classes with the full scale face descriptor (i.e. head pose, horizontal gaze surrogate, vertical gaze surrogate and eye appearance descriptor) and with the multiple camera framework.

2.5 Concluding Remarks

Gaze zone estimation is still a new area of research in literature. While there are many ways to design a gaze zone estimator, the focus of this research is in carefully designing submod-

ules which will build up to achieve a continuous and robust gaze zone estimation. Key modules in this system include, face detection using deep convolutional neural networks, landmark estimation from cascaded regression models, head pose using relative configuration of 2D points in the image plane to 3D points in the head model, horizontal gaze surrogate based on geometrical formulation of the eye ball and iris position, vertical gaze surrogate based on openness of the upper eye lids and appearance descriptor, and finally, a 9-class gaze zone estimation from naturalistic driving data driven random forest algorithm. In addition to the contributions in each of these submodules, this chapter also gives an extensive analysis of the system as a whole on naturalistic driving data.

The gaze estimator described in this chapter is based on static features, which means no information in time window previous to the instance is leveraged. Near future work will extend this framework to dynamics features and therefore show significant improvement in current framework's sensitivity to occlusions and generic models.

2.6 Acknowledgments

Chapter 2 is a partial reprint of materials published in IEEE Intelligent Transportation Systems Conference (2013), by Sujitha Martin, Ashish Tawari and Mohan M. Trivedi, in the IEEE Transactions on Intelligent Transportation Systems (2014), by Ashish Tawari, Sujitha Martin and Mohan M.Trivedi, and in IEEE International Conference on Pattern Recognition (2016), by Kevan Yuen, Sujitha Martin and Mohan M. Trivedi. The dissertation author was one of the primary investigator and author of these papers.

Chapter 3

Gaze Dynamics, Modeling and Behavior Understanding

3.1 Introduction

In the age of self-driving and highly automated vehicles, the roles and responsibilities of human driver and “system” are still in the making. The Society of Automotive Engineers (SAE) International [88] has reduced some confusion and provided common terminology for automated driving by identifying six levels of driving automation from “no automation” to “full automation”. In addition, as show in Table 3.1, SAE also outlines who bares responsibility, human drivers or system, in three well defined categories: execution of steering and acceleration/deceleration, monitoring of driving environment, and fallback performance of dynamic driving task. Some examples of automated system capabilities which have already penetrated consumer vehicles and their corresponding level of automation include: blind spot alert and lane deviation alert which are considered to be a part of driver assistance, and adaptive cruise control and lane keep assist which are considered to be a part of partial automation. Higher levels of automation are still in their infancy.

Up to the level of partial automation, human drivers are expected to continuously monitor the driving environment and therefore the sole responsibility of safety lies with the humans. From the level of conditional automation and up, however, human drivers are not required to continuously monitor the driving environment until there arises a situation that the system cannot handle (e.g. system boundaries due to sensor limitations or ambiguous environment). In higher levels of automation, not requiring human drivers to monitor the driving environment is equivalent to giving human drivers permission to increasingly engage in non-driving secondary task. In which case, what if the driver is not ready to take-over control? Or what if the drive requires

Table 3.1: SAE identifies six levels of automation from “no automation” to “full automation” in order to provide common terminology for automated driving (for full table with definitions see [88]).

SAE level	Name	Execution of Steering and Acceleration/Deceleration	Monitoring of Driving Environment	Fallback Performance of Dynamic Driving Task	System Capability (Driving Modes)
0	No Automation	Human driver	Human driver	Human driver	n/a
1	Driver Assistance	Human driver and system	Human driver	Human driver	Some driving modes
2	Partial Automation	System	Human driver	Human driver	Some driving modes
3	Conditional Automation	System	System	Human driver	Some driving modes
4	High Automation	System	System	System	Some driving modes
5	Full Automation	System	System	System	All driving modes

t_{req} seconds to get ready, but system will reach its limit in t_{given} seconds, where $t_{given} < t_{req}$?

A recent study on how long it takes to get the driver back into the loop [44], attempted to address the question of “*at what point in time, prior to the occurrence of a system boundary, does the automation-system have to engage the attention of the driver in order to ensure a successful take-over process*”. In that study an experiment was conducted in a dynamic driving simulator, where two take-over times (i.e. 5-seconds and 7-seconds) were examined and compared with manual driving. The conditions of the experiment was such that during automation, drivers were engaged in secondary task and therefore inattentive to and disengaged from the task of driving. There are at least two interesting findings from this study. First is that with shorter TOR (take-over request) time, while the reaction was faster, the quality was worse. The quality here is measured based on reaction type (i.e. brake, steering or both), reaction time, gaze behavior (i.e. mirror checking), etc. This leads to the second interesting finding that in manual mode and automation with both TOR (take-over request) times, in the advent of unexpected incidents in the roadway, among the drivers who reacted by changing lanes, only a minor part of the subjects checked their blind spots. Such negligence can be overlooked in a driving simulator but during real, on-road driving it could lead to collisions or near-crashes.

Therefore, one of the major challenges in highly automated vehicles is ensuring the driver is able to intervene or take-over in a timely and safe manner. It is a challenge for at least three reasons. First, what does “able” to intervene or “readiness” to take-over mean? Second, how to measure “readiness” objectively? Third, how are “readiness” and time correlated? Unlike in manual driving, vehicle control based measures as a performance indicator cannot be used in highly automated vehicles; other metrics must be explored. As such, looking-in at the driver, their state and their behavior is key to developing readiness metrics for highly automated vehicles. For instance, the alertness of the driver to the changing road conditions, their engagement with other

passengers, their interactions with object are all useful indicators for deriving driver’s readiness.

Driver’s gaze is of particular interest because if and how the driver is monitoring the driving environment is the first step towards assuring safe and timely take-over. Literary works have addressed the problem of estimating driver’s awareness of the surround in a few different ways. Ahlstrom et al. [3], with an underlying assumption that the driver’s attention is directed to the same object as the gaze, developed a rule based 2-second ‘attention buffer’ which depleted when driver looked away from the field relevant to driving (FRD); and it starts filing up when the gaze direction is redirected toward FRD. One of the reasons for a 2-second buffer is because eyes off the road for more than 2 seconds significantly increases the risk of a collision by at least two times that of normal, baseline driving [54]. Tawari et al. [96], on the other hand, developed a framework for estimating the driver’s focus of attention by simultaneously observing the driver and the driver’s field of view. Specifically, the work proposed to associate coarse eye position with saliency of the scene to understand on what object the driver is focused at any given moment. Li et al.[59], under the assumption that mirror-checking behaviors are strongly correlated with driver’s situation awareness, showed that the frequency and duration of mirror-checking reduced during secondary task performance versus normal, baseline driving. Furthermore, mirror-checking actions were used as features (e.g. binary labels indicating presence of mirror checking, frequency and duration of mirror checking) in driving classification problems (i.e. normal versus tasks/maneuver recognition). However, the classification problem had as its input, features from CAN signal and external cameras, whereas the classification and recognition problem addressed in this chapter is purely based on looking at the driver’s gaze.

3.2 Related Research

Two of the main contributions and objectives in this work is: first, modeling the gaze behavior, and therefore the driver’s behavior, using unique and well representative features; second, modeling it from and evaluating it on naturalistic driving data. Therefore, in the following paragraphs, we present related works with respect to our objectives.

Gaze behavior modeling described in this chapter features the importance of glance duration and glance transition frequency as a stepping stone. To compute glance durations and glance transition frequency relies on the gaze zone estimation work presented in the previous chapter; glance duration is defined as the time driver spent looking in a specific gaze zone of interest in a given time window and glance transition frequency is the number of times driver went between two gaze zones of interest, respective or irrespective of the order, depending on preference, in a given time period. In literature, works that feature automatic gaze zone estimation very rarely go further to infer gaze behavior. Tawari et al. [100] use a similar approach for gaze zone estimation as presented in the previous chapter but only applied the estimator as a proof of concept to recreate the AttenD buffer [3]. AttenD buffer refers to a rule based 2-second ‘attention buffer’

framework referenced to earlier in this chapter. AttenD models the drivers attention to FRD but does not model behavior like the work presented in this chapter. Fridman et al. [40] presented learning based framework for gaze zone estimation and evaluations are commendably done on a large dataset. Interestingly, [40] defined an overall driver behavior metric where drivers were categorized as either "owl" or "lizard". The objective was to classify drivers based on how they looked at gaze zones, using their heads like owls or with their eyes like lizards, in order to build more customized gaze zone estimators. Our work here differs in that it is in modeling of gaze behavior for on-line performance of predicting driver's likelihood of exhibiting one of many gaze behaviors (e.g. right-lane-change-gaze-behavior, interacting-with-center-stack-gaze-behavior).

In terms of gaze modeling and behavior understanding, literary works have mainly conducted studies in a driving simulator but few recent works on on-road have emerged. In one on-road study, Birrell and Fowkes [11] explore the effects of using in-vehicle smart driving aid on glance behavior. The study is similar to ours in that it uses glance durations and glance transition frequencies to show difference in glance behavior between baseline, normal driving and when using in-vehicle devices. It differs in that no model is build and compared to new instance for prediction because the study was not intended for that purpose. In another on-road study, Li et al. [59] show that drivers exhibit different gaze behaviors when engaged in secondary tasks versus baseline, normal driving by using mirror-checking actions as indicators for differentiating between the two. Our work differs from [59] in two ways, first the latter study simplifies gaze classification into whether driver is mirror-checking or not and second when training to detect secondary tasks and maneuvers, features other than gaze are used (e.g. CAN bus, road cam). In our study, glance durations and frequencies are computed upon the nine classes and maneuver prediction is done using gaze relate features alone.

3.3 Naturalistic Driving Dataset

A large corpus of naturalistic driving dataset is collected using subjects' personal vehicles over the span of six months. Individual vehicles are instrumented with four Hero4 GoPro cameras: two for looking at the driver's face, one for looking at the driver's hands and one for looking at the forward driving direction. One of the driver facing camera is mounted on the windshield near the rear view mirror, while the other one is mounted on the left window near the A pillar without obscuring the view to left rear view mirror or obstructing the path of potential side airbag deployment; the hand viewing camera is positioned to have an over the shoulder view of the wheel and center stack region and the forward facing camera is mounted on the windshield near the rear view camera looking out. In this study, the focus is in analyzing the driver's face, while the forward view provides context for data mining; the hand looking camera is instrumented for future studies of holistic modeling of driver behavior.

All cameras are configured to capture data with 1080p resolution at 30 frames per second

Table 3.2: Dataset description of naturalistic driving for gaze modeling and behavior classification

Trip No.	Full drive		Left Lane Change		Right Lane Change		Lane Keeping	
	Duration [min]	Frames	No. of Events	Duration [sec]	No. of Events	Duration [sec]	No. of Events	Duration [sec]
1	81.2	14621	11	1650	16	2400	104	15600
2	88.5	159300	14	2100	15	2250	92	13800
3	89.6	161338	16	2400	17	2550	112	16800
4	93.9	169082	14	2100	15	2250	179	26850
All	353.3	635931	55	8250	49	9450	487	73050

(fps). While GoPro smart remote can start recoding of multiple cameras simultaneously, manual synchronization across all cameras, for example using a light source, is necessary at least once in the beginning; over a 90 minute drive, such a synchronization drifts only a couple of frames. With exact camera configuration and similar camera placements, the subjects captured data using their instrumented personal vehicles during their long commutes. The drives consisted of some driving in urban settings, but mostly in freeway settings with lanes ranging from a minimum of two up to six lanes. The data is especially collected in their personal vehicles in order to retain natural interaction with vehicle accessories and during the subjects’ usual commutes in order to study driver behavior under some constraints (e.g. familiar travel routes) with certain variables (e.g. traffic conditions, time of day/week/month).

This study analyzes data from a subset of the large corpus, to be exact four drives, whose details are as given in Table 3.2. From the collected data, with a special focus on freeway settings, events were selected when the driver executes a lane change maneuver, by either changing left or right, and when the driver keeps the lane. Table 3.2 shows the events considered, their respective counts and total frames analyzed with respect to one vision sensor. The duration of executing a maneuver varies depending on the traffic conditions, driving style, road conditions, etc. In Section 3.5.1, we define epochs that consistently exist in lane change maneuvers. These epochs are necessary to properly extract temporal features for analysis (see Section 3.5.2).

3.4 Modeling Scanpath for Driver Gaze Behavior Recognition

What is a gaze behavior? One definition of behavior is the way in which a person acts in response to a particular situation or stimuli. Then given a particular situation or stimuli, gaze behavior is more than where the driver is looking. Gaze behavior is derived from where the driver has been looking over a period of time. For example, when interacting with the center stack (e.g. radio, AC, navigation), the manner in which the driver looks at the center stack can be vastly different. One may perform the secondary task with one long glance away from the forward driving direction and at the center stack. Another time one may perform the secondary

task via multiple short glances towards the center stack, etc. However, while individual gaze patterns are different, together it is associated with an action of interest. Therefore, we define any individual gaze pattern as scanpath and the collection of scanpaths associated with a task performed by driver is what we define as gaze behavior.

3.4.1 Temporal Feature Descriptor

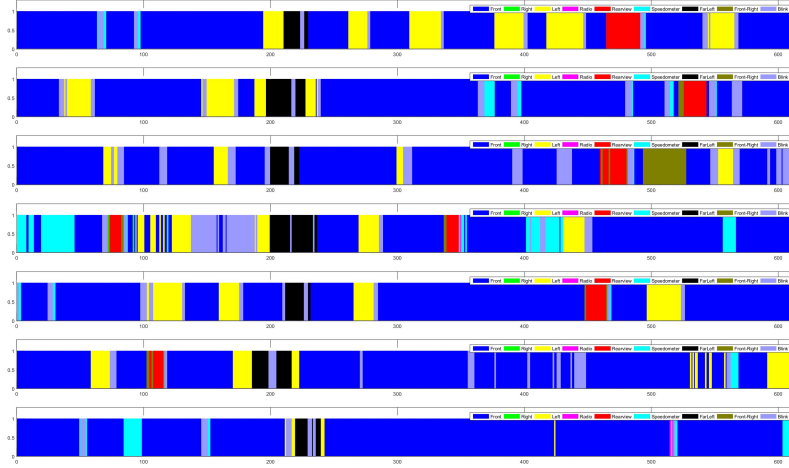
Figure 3.1 illustrates fourteen different scanpaths during a 20-second time window centered on lane change event, seven scanpaths during left lane change events and seven scanpaths during right lane change events. In the figure, the x -axis represents time and color displayed at a given time t represents the estimated gaze zone; note, frame number 300 represents when the car is half way in between the source and destination lane (let’s call this *SyncF*). Visually, 5-seconds before the *SyncF* there is some consistency observed across the different scanpaths within a given event (e.g. left lane change); consistencies such as the glance duration and glance transitions between gaze zones. For example, in the scanpaths associated with right lane change, the driver glances at the *rearview* and *right* gaze zones for a significant duration. However, the start point or the end point of the glances is not the same across the different scanpaths. Therefore, we represent the scanpaths using features called glance duration and glance transition frequency, which remove some temporal dependencies but still capture sufficient spatio-temporal information to distinguish between different gaze behaviors.

Both glance durations and glance transition frequencies are computed over a time window. First, we define signals necessary to compute them. Let Z , represent the set of all nine gaze zones as $Z = \{Front, Right, Left, Center Stack, Rearview, Speedometer, Far Left, Front Right, Eyes Closed\}$ and let $M = |Z|$ represent the total number of gaze zones. Then, let the vector $G = [g_1, g_2, \dots, g_N]$ represent the estimated gaze for an arbitrary time period of T , where $N = fps \times T$, $g_n \in Z$, and $n \in \{1, 2, \dots, N\}$.

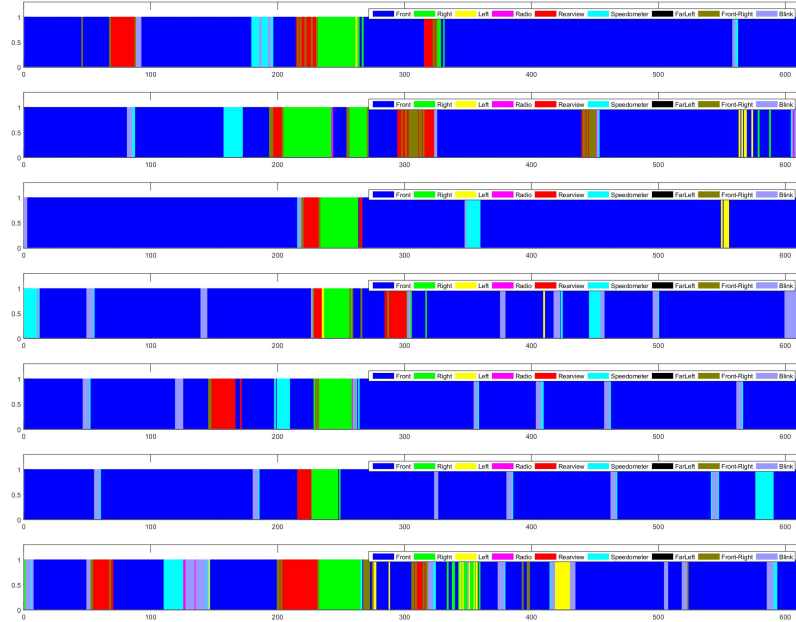
Glance duration is a function of the gaze zone. Given a gaze zone, glance duration for that gaze zone is the amount of time driver spends looking at the gaze zone within time period; which is then normalized by the time window for relative duration calculation. Glance duration then is calculated for each of the gaze zones, z_m , where $z_m \in Z$ and $m \in \{1, 2, \dots, M\}$, as follows:

$$Glance\ Duration(z_m) = \frac{1}{N} \times \sum_{n=1}^N \mathbb{1}(g_n == z_m)$$

Glance transition frequency is a function of two gaze zones. Given two gaze zones, glance transition frequency is the number of times the driver glanced from one gaze zone to the other sequentially, in a time period; this value is normalized by the time window for determining transition frequency with respect to time. The algorithm to compute a matrix, \mathbf{F}_{GT} , of glance transition frequency for all combinations of transitions is outlined in Algorithm 1. The matrix



(a) Left Lane Change Events



(b) Right Lane Change Events

Figure 3.1: Illustrates fourteen different scanpaths during a 20-second time window centered on lane change event, seven scanpaths during left lane change and seven scanpaths during right lane change event.

\mathbf{F}_{GT} is a transition frequency matrix, which means the diagonals are by definition 0:

$$F_{GT}(d, k) = \begin{cases} 0 & d == k \\ f_{dk} & d \neq k \end{cases}$$

```

input :  $G = [g_1, g_2, \dots, g_N]$ 
          $W$ , a positive time window threshold for consistency check,  $< N$ 
output: A matrix,  $F_{GT}$ , of glance transition frequency for all combination of
         transitions

LastGazeState =  $g_1$ 
for  $i \leftarrow W$  to  $N$  do
  if  $g_i \neq$  LastGazeState then
    if  $g_i == g_{i-1} == \dots == g_{i-W}$  then
       $F_{GT}(\text{LastGazeState}, g_i) ++;$ 
      LastGazeState =  $g_i$ 
    end
  end
   $i++;$ 
end

```

Algorithm 1: To compute a matrix of glance transition frequencies given estimated gaze zones over a time period.

where f_{dk} is the number of transitions from z_d , the gaze zone representing the d^{th} column, to z_k , the gaze zone representing the k^{th} column. Furthermore, in order to remove the order of transition, F_{GT} is first decomposed into its respective lower, L_{GT} and upper, U_{GT} , triangular matrix:

$$L_{GT}(d, k) = \begin{cases} 0 & d \geq k \\ f_{dk} & d < k \end{cases}$$

$$U_{GT}(d, k) = \begin{cases} 0 & d \leq k \\ f_{dk} & d > k \end{cases}$$

The decomposed triangular matrices are then combined by adding the upper triangular matrix to the transposed lower triangular matrix and normalized to produce the new glance frequency transition matrix:

$$F_{GT}^*(d, k) = \frac{1}{T} \times \begin{cases} 0 & d \leq k \\ f_{dk} + f_{kd} & d > k \end{cases}$$

The final feature vector representing a scanpath then is made up of two components. First is the glance durations over all gaze zones. Second is upper triangular matrix of the new glance transition frequency matrix, F_{GT}^* in vectorized form. Both these components are appended to produce the final feature vector \vec{h} .

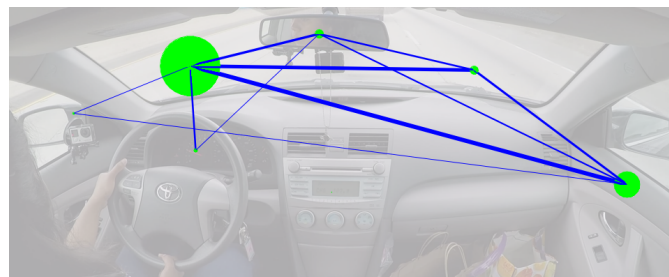
3.4.2 Gaze Behavior Modeling

Consider a set of feature vectors $\vec{H} = \{\vec{h}_1, \vec{h}_2, \dots, \vec{h}_N\}$, and their corresponding class labels $Y = \{y_1, y_2, \dots, y_N\}$. For instance, the class labels can be: *Left Lane Change*, *Right Lane Change*, *Merge*, *Secondary Task*. Given \vec{H} and Y , we compute the mean of the feature vectors within each class. Figure 3.2 illustrates mean of the vectors for the four above listed classes are illustrated. Note that the mean feature vector \vec{h} is composed of glance durations and glance transition frequency. Therefore, in Figure 3.2 the glance durations are represented by the solid green circles and glance transition frequencies are represented by the blue lines. The area of the circle is relatively proportional to the percentage of glances to the respective gaze zone within a time window and the thickness of the connecting lines are relatively proportional to the number of transitions between gaze locations also within the same time window. As expected in all the classes, the driver spends a significant portion of the time looking forward. An interesting observation occurs when comparing the glance duration and transition frequency decomposition between left lane change and merge because a merge is similar to a left lane change. However, the designed feature vector of glance duration and transition frequencies allows to differentiate between the two. In the rest of the study, only the following three classes are considered: *Left Lane Change*, *Right Lane Change* and *Lane Keep*.

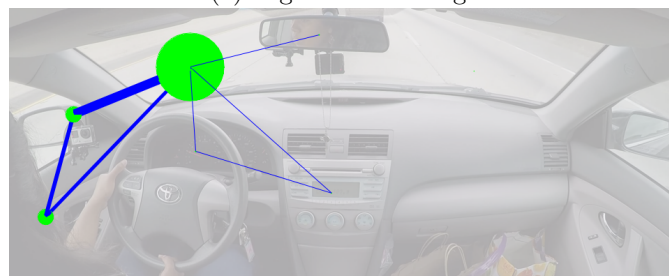
In this work we model the gaze behaviors of respective events, tasks or maneuvers, using a multivariate normal distribution (MVN). An unnormalized MVN is trained for each behaviors of interest:

$$M_b(\vec{h}) = \exp\left(-\frac{1}{2}(\vec{h} - \vec{\mu}_b)^T \Sigma_b^{-1}(\vec{h} - \vec{\mu}_b)\right)$$

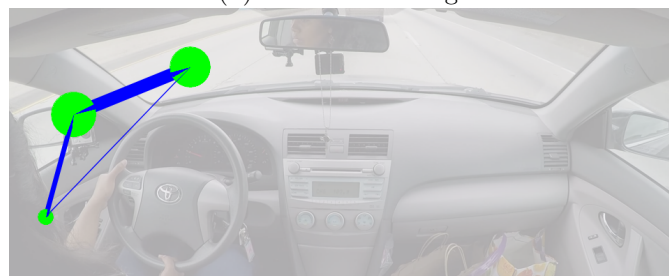
where $b \in B = \{\textit{Left Lane Change}, \textit{Right Lane Change}, \textit{LaneKeep}\}$, and μ_b and Σ_b represent the mean and covariance computed from the training feature vectors for the gaze behavior represented by b . One of the reasons for modeling gaze behavior using MVN is, given a new test scanpath \vec{h}_{test} , we want to know how does it compare to the average scanpath computed for each gaze behavior in the training corpus. One possibility is by taking the euclidean distance between the average scanpath, μ_b , and the test scanpath, \vec{h}_{test} , for all $b \in B$ and assign the label with the shortest distance. However, this assigns equal weight or penalty to each component of \vec{h} . We want the weights to be a function of component as well the behavior under consideration. Therefore, we use the Mahalanobis distance, which is the component in the exponent of the unnormalized MVN. By exponentiating the Mahalanobis distance, the range is mapped between 0 and 1. To a degree this can be used to asses the probability or confidence that a certain test scanpath, \vec{h}_{test} belongs to a particular gaze behavior model.



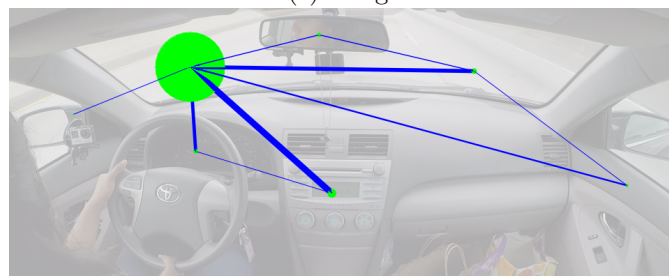
(a) Right Lane Change



(b) Left Lane Change



(c) Merge



(d) Secondary Task

Figure 3.2: Glance duration and transition decomposition analysis for (a) right lane change, (b) left lane change, (c) merge and (d) instrument cluster (averaged over multiple runs in the naturalistic driving scenarios). The area of the circle is proportional to percentage of glances and the thickness of the connecting lines proportional to the number of transitions between gaze locations.

3.5 Experimental Design and Analysis

3.5.1 Event Description

Every instance of driving on a freeway can be broken or categorized exclusively into one of these three categories, left lane change, right lane change and lane keep. As a point of synchronization, for lane change events, when the vehicle is half in the source and half in the destination it is marked at annotation. A time window of 5-seconds before this instance defines the left and right lane change events. For the lane keeping events, a lengthy stretch of lane keeping is broken into non-overlapping 5-sec time windows to create lane keeping events. Table 3.2 contains the number of such events annotated and considered for the following analysis. The analysis in the following sections are conducted from modeling the gaze for the events described here with sufficient separation of the training and testing corpus.

3.5.2 Evaluation on Gaze Modeling

All evaluations conducted in this study is done with a four-fold cross validation; four because there are four different drives as outlined in Table 3.2. Instead of the usual leave one out cross-validation, training is done with events from one drive and tested with events from the other drives. With this setup of separating the training and testing samples, two sets of evaluations are conducted in this study. First, we explore the precision and recall of the gaze behavior model in predicting lane changes as a function of time. Second, we explore the balance of learning from the training data without overfitting and generalizing to unseen testing data.

In the first case, training occurs on the 5-second time window before *SyncF* (*SyncF* is defined in Section 3.4.1) as represented by the events in Table 3.2. While testing, however, we want to test how early the gaze behavior models are able to predict lane change. Therefore, starting from 5-seconds before *SyncF*, sequential samples with $\frac{1}{30}$ of a second overlap are extracted up to 5-seconds after *SyncF*; note that the time window at 5-seconds before the *SyncF* encompasses data from 10 seconds before the *SyncF* up to 5-seconds before the *SyncF*. Each of the samples are tested for fitness across the three gaze behavior models, namely models for *left lane change*, *right lane change* and *lane keep*. The sample is assigned the label based on the model which procures the highest fitness score and if the label matches the true label the sample is considered a true positive. Note that each test sample is associated with a time index of where it is sampled from with respect to *SyncF*. By gathering samples at the same time index with respect to *SyncF* from drives not included in the training set, precision and recall values are calculated as a function of the time index. When calculating precision and recall values, true labels of samples were remapped from three classes to two classes; for instance, when computing precision and recall values for left lane change prediction, all right lane change events and lane keep events were considered negatives samples and only the left lane change events are consid-

Table 3.3: The recall and precision of lane change prediction (averaged over multiple runs in naturalistic driving scenarios, 88 min each) via gaze behavior modeling using multivariate Gaussian.

Time before maneuver		Left Lane Change Prediction		Right Lane Change Prediction	
Frame	Milliseconds	Precision	Recall	Precision	Recall
60	2000	0.4357	0.6931	0.6098	0.7978
54	1800	0.4605	0.7954	0.6066	0.7872
48	1600	0.4740	0.8295	0.6250	0.8510
42	1400	0.4780	0.8636	0.6279	0.8617
36	1200	0.4810	0.8636	0.6391	0.9042
30	1000	0.4873	0.8750	0.6391	0.9042
24	800	0.4906	0.8863	0.6391	0.9042
18	600	0.4969	0.9090	0.6418	0.9148
12	400	0.4815	0.8863	0.6364	0.8936
6	200	0.4780	0.8636	0.6391	0.9042
0	0	0.4753	0.8750	0.6279	0.8617

ered positive samples. Similar procedure is observed for computing precision and recall values for right lane change prediction. Table 3.3 shows the development of the precision-recall values for both left and right lane change prediction in an interval of 200 milliseconds starting from 2000 milliseconds prior to *SyncF* up to 0 milliseconds prior to *SyncF*. Interestingly, the peak of precision and recall values occur around 600 milliseconds before the driver is half way in each lanes. There are several possibilities for such an occurrence. One, it may be likely that there is a strong indication at that time index of intended lane change using gaze analysis only. Other possibilities include the misalignment of training and testing samples, the time window considered, the point of marking what is the start of lane change, etc. With respect to the precision rate, the values are expected to be low because even during lane keeping, drivers may exhibit lane change like behavior without lane changing. This is especially observed with the precision rate for left lane change prediction, where checking left rear view mirror is a strong indicator of lane change but also part of driver’s normal mirror checking behavior during lane keep. One of the main cause is the gaze model for *lane keep*, which encompasses a broad spectrum of driving behavior and future work will consider finer labeling of classes.

Lastly, we explore the gaze models ability to adapt to unseen data. Unlike the previous experiment where we explored the early predictability factor of the gaze model, here we explore whether modeling gaze behavior as an average of observed glance duration and transition frequency may be overfitting the training data. In this experiment, training and testing are accumulated at the same time index, which is the 5-second time window before *SyncF* uptill *SyncF*. Figure 3.3a shows the confusion matrix when the same samples are used for training and testing; model is trained from samples on one drive and tested on the same training samples, with values averaged over all four drives. Figure 3.3b shows the confusion matrix when training and testing data are separate; model is trained from samples of one drive and tested on samples from other drives, with values averaged using four-fold cross-validation. With a weighted accuracy or recall rate of 97% with the former case and 86% with the latter case, the results are promising with



Figure 3.3: Confusion matrix, where rows are true classes and columns are predicted classes, from two experiments. (a) The same data is used for training and testing; model trained on one drive is tested on the same drive, with values average over all four drives. (b) Training and testing data are separate; model trained on one drive and tested on the other drives, with values averaged using four-fold cross validation.

indications of room for improvement.

3.5.3 Discussion on Gaze Behavior Understanding

One of the motivations behind this work is to estimate the driver’s readiness to take over in highly automated vehicles. Situations where system cannot handle thus requiring take-over include system boundaries due to sensor limitation and ambiguous environment. In such situations, looking inside at the state of the driver and how much time is required to reach readiness to take-over is important. Therefore, in this study we developed a framework to estimate his or her readiness to handle the situation by modeling gaze behavior from non-automated naturalistic driving data. In particular, gaze behavior during lane changes and lane keep is modeled. Figure 3.4 illustrates the fitness or confidence of the model around left and right lane change maneuvers. The figure shows mean (solid line) and standard deviation (semitransparent shades) of three models (i.e. *left lane change*, *right lane change*, *lane keep*) for two different maneuvers (i.e. left lane change and right lane change), using the events from naturalistic driving dataset described in Table 3.2. The model confidence statistics are plotted 5 seconds before and after the lane change maneuver, where time of 0 seconds represents when the vehicle is half way between the lanes. Interestingly, during left (right) lane change maneuver, fitness of the right (left) lane change gaze model is very low within the 10 second bracket and left (right) lane change model peaks in fitness in a tighter time window around the maneuver. Furthermore, for both maneuvers, the lane keep model as desired is high in the periphery of the lane change maneuvers hinting at how this model performs during actual lane keep situations.

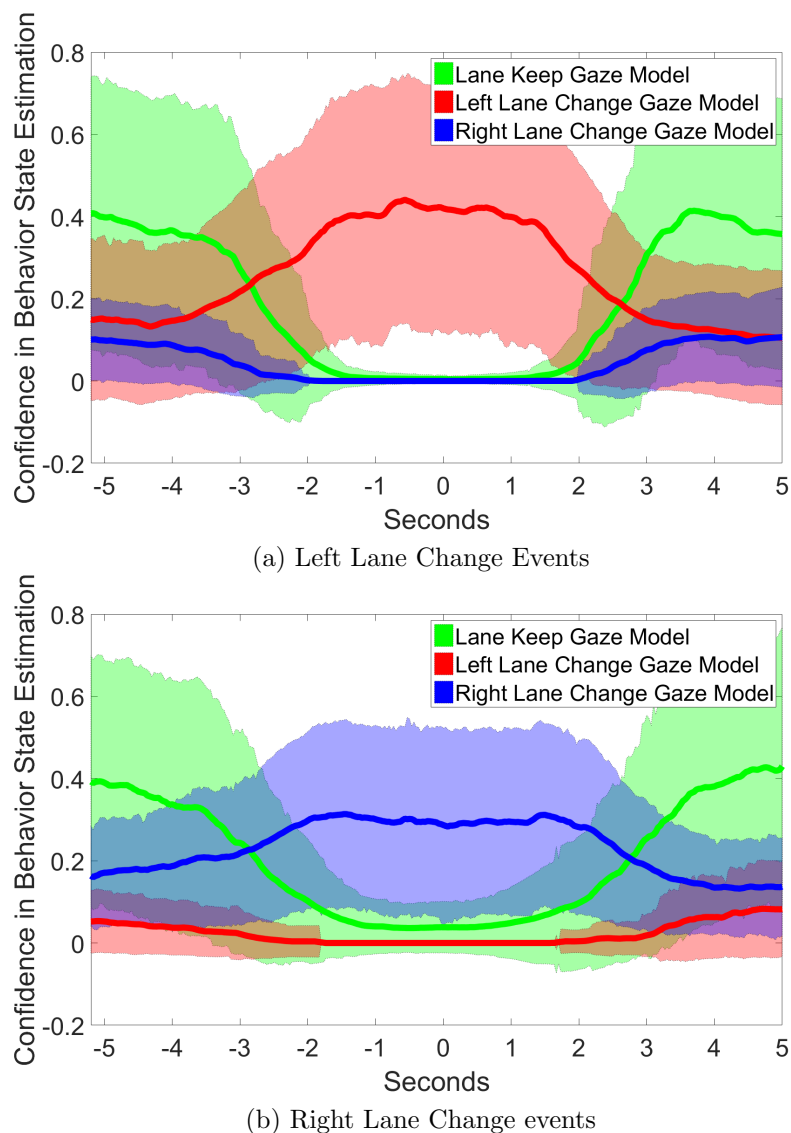


Figure 3.4: Illustrates the fitness of the three models (i.e. *Left lane change*, *Right lane*, *Lane keep*) during left and right lane change maneuvers. Mean (solid line) and standard deviation (semitransparent shades) of the three models as applied to the lane change events described in Table 3.2 are shown.

3.6 Concluding Remarks

In this study, we explored modeling driver's gaze behavior in order to predict maneuvers performed by drivers, namely left lane change, right lane change and lane keep. The particular model developed in this study features three major aspects: one is the spatio-temporal features to represent the gaze dynamics, second is in defining the model as the average of the observed instances and interpreting why such a model fits the data of interest, third is in the design of the metric for estimating fitness of model. Applying this framework in a sequential series of time

windows around lane change maneuvers, the gaze models were able to predict left and right lane change maneuver with an accuracy above 80% around 1600 milliseconds before the maneuver and reaches a peak of 90% recall rate around 600 milliseconds.

The overall framework, however, is designed to model driver's gaze behavior for any tasks or maneuvers performed by driver. In particular, the spatio-temporal feature descriptor composed of glance duration and glance transition frequencies are powerful tools to capture the essence of recurring driver gaze dynamics. To this end, there are multiple future directions in site. One is to quantitatively define the relationship between the time window from which to extract those meaningful spatio-temporal features and the task or maneuvers performed by driver. Another is in exploring and comparing different modeling approaches, including HMM, SVM and bag of words approaches. Other future directions include exploring unsupervised clustering of gaze behaviors and exploring the effects of quantity and quality of events (e.g. same vs. different drives, different drives from same or different time of day) on gaze behavior modeling.

3.7 Acknowledgments

Chapter 3 is a partial reprint of material currently being prepared for submission for publication to the IEEE Transactions on Intelligent Vehicles, by Sujitha Martin, Sourabh Vora, Kevan Yuen and Mohan M. Trivedi. The dissertation author was the primary investigator and author of this paper.

Chapter 4

Driver Modeling by Multi-cue Fusion of Head, Eyes and Hands

4.1 Introduction and Related Works

In a recent report, NHTSA revealed that urban automotive collisions accounted for some 46% of 33,782 fatal crashes in the United States in 2012; intersection-related collisions accounted for 26% of fatal crashes [1]. Furthermore, intersections, especially stop-controlled, have been found to be the most demanding driving scenario among all typical scenarios in driving [60]. The visual demands include pedestrians, bicyclists, cross-traffic vehicles, etc. and these demands typically vary depending on the maneuver executed at the stop-intersection [55]. For instance, going straight at stop-controlled intersections does not demand as much of driver's attention to a pedestrian crossing in parallel to the path traveled when compared to turning right or turning left. These preconceived notions of what and where demands more attention depending on the maneuver can lead to varying preparatory actions observable of the driver prior to executing the maneuver [80]; preparatory actions such as head movements, eye glances and hand movements.

Understanding, extracting and temporally modeling these preparatory actions before the onset of the respective maneuver has tremendous benefits in the development of intelligent vehicles. Benefits include catering driver assistance for the maneuver and using computational resources more efficiently. While the natural progression of this work is towards prediction and intent detection [24], the scope of this paper is to use real driving data and data driven approaches to guide in the search for signals which best represent these preparatory motions.

Therefore, in order to better understand and properly exploit the preparatory motions a reality, we equipped an experimental vehicle with cameras and sensors to capture the vehicle dynamics, view of the road ahead, driver's face and driver's hand. This paper investigates how

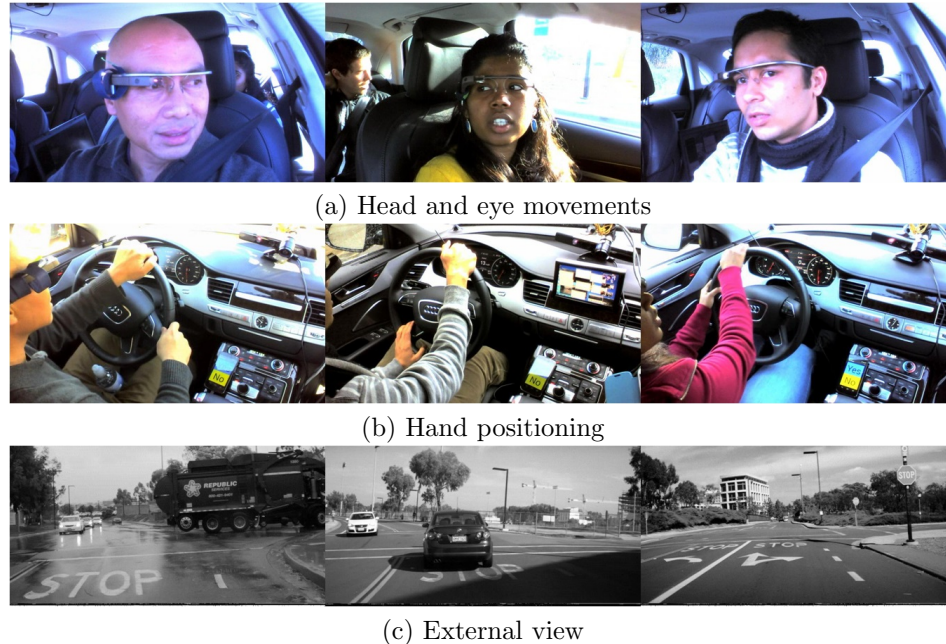


Figure 4.1: Sample instances of (a) Head and eye movements, (b) Hand positioning and (c) external context prior to the start of respective maneuvers (i.e. stop and right/left turns, stop and go straight).

and to what extent information from driver’s face and hands can be used to differentiate between maneuvers at a stop-controlled intersection before the maneuver is executed. Figure 4.1 shows sample instances of driver’s preparatory motions prior to the start of respective maneuvers. Note that while external views are shown and discussed throughout the paper, external views are shown and used only for the sake of context. To this end, the contributions are in three folds:

- **Naturalistic driving dataset collection:** synchronized capture of sensors looking-in and looking-out, multiple drivers driving in urban environment, and segmenting events at stop-controlled intersections
- **Extracting reliable features:** eye movements [100], head pose [95] and hand location [86] respective to the wheel from purely vision sensors looking in at the driver
- **Data driven processing:** construction of features using temporal pyramids and using the random forest algorithm to extract optimal feature subset, where optimal features reveal when and what cue is most relevant for representing the preparatory motions.

Given the above contents summary, this work is most relatable to the study of head, eyes and hand coordination for driver hand activity classification [62]. The common ground is with respect to the type of extracted features (i.e. head, eyes and hands cues) and showing with data the relationship between head, eye and hands. In addition to the fundamental difference of

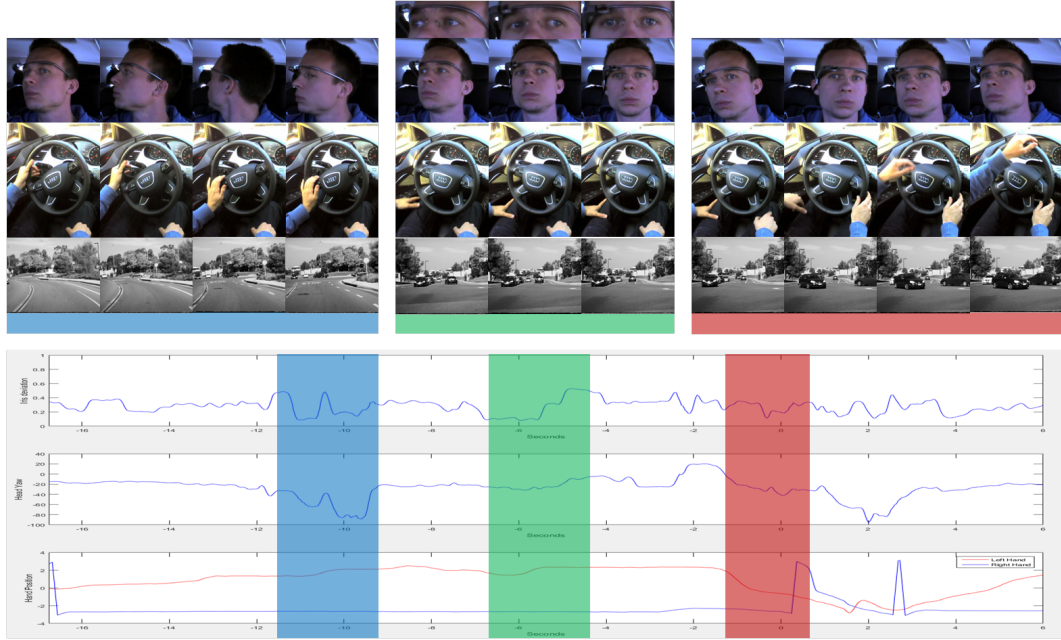


Figure 4.2: A two part figure illustrating an interesting coordination of head, eye and hand movements prior to the driver starting a right turn at the stop-controlled intersection. The bottom part is a time series plot of the cues: eye deviation, head pose and relative hand location. The top part presents instances from looking-in and looking-out. The colored coded boxes in the two parts shows where the images are taken from in the time series.

[62] focusing on driver hand activity recognition while this paper is concerned with maneuvers at stop-controlled intersections, the process of extracting and congregating features temporally are also fundamentally different, as described in later sections. With regards to stop-controlled intersections and extracted features, work presented in [60] is similar, but the focus of the study is on detecting driver cognitive distraction and the findings are from data collected in driving simulators. Another work which deals with intersection data and deals with cues from looking inside at the driver, is the work on turn-intent analysis using body pose [14], however the cues and signal selection and interpretation is vastly different.

There are many other literary works related to the work presented in this paper. The difference among them and this work is usually in two regards: analysis of signals/cues are conducted similarly but applied in highway driving conditions or application is urban setting but no thorough analysis on feature selection and relevance study. As an example of the latter case, Jain et al. [52] proposes a framework which anticipates maneuvers performed few seconds in the future. This study is indeed very interesting and useful, as it uses a data driven approach using real-driving data. The lane change intent [24] and overtaking/braking intent [75] work similarly uses a data driven approach, and explores the importance of integrating inside looking cues. Both however, deal with highway driving.

The emphasis of this paper is in understanding, extracting and representing the coor-

dination of head, eyes and hands cues in the preparatory stages leading up to maneuvers at stop-controlled intersections. Figure 4.2 illustrates one right turn event from our dataset, where time 0 represents the start of the turn. The time series plots at the bottom show the three important cues of interest: iris deviation, head pose in yaw, and hand location with respect to center of steering wheel. In this particular event, there is evident head motions 10 seconds before, there is a unique presence of iris deviations 6 seconds before and finally strong hand cues are observable up to 1 second before the right turn maneuver is executed. This is but an exemplary show of head, eye and hand coordination for a particular turn event. This paper will analyze multiple events across varying drivers and different intersections, and show promising results as observed in this example.

4.2 Extraction, Representation and Fusion of Spatio-Temporal Features

4.2.1 Head and Eye Features

In this work, the state of the head and eye are represented using head pose and normalized iris deviation. Chapter 2 provides details on obtaining facial landmarks and head for a given image. This section contains detail on the second representative feature, iris deviation. Iris deviation is defined as the distance from the iris to one of the eye corners, normalized by the distance between the eye corners. For example, iris deviation of the left eye is calculated as follows:

$$\Delta d_{LEI} = \frac{d(q_{LELC}, q_{LEI})}{d(q_{LELC}, q_{LERC})} \quad (4.1)$$

with d as the Euclidean distance. The vectors q_{LELC} , q_{LERC} , and q_{LEI} are the left eye’s left corner coordinates, left eye’s right corner coordinates, and left eye’s iris coordinates, respectively. Iris deviation of the right eye is calculated similarly.

Given a video sequence $V^T = \{I_1, I_2, \dots, I_T\}$ of length T , the final output of head and eye features are represented as:

$$\begin{aligned} \mathcal{X}_{head}^f &= \{\phi_1, \phi_2, \dots, \phi_T\}, \\ \mathcal{X}_{eye}^f &= \{\Delta d_1, \Delta d_2, \dots, \Delta d_T\} \end{aligned}$$

where ϕ is the yaw rotation of the head and Δd is iris deviation of the eye with the most visibility; the visibility is determined by head pose with respect to camera perspective.

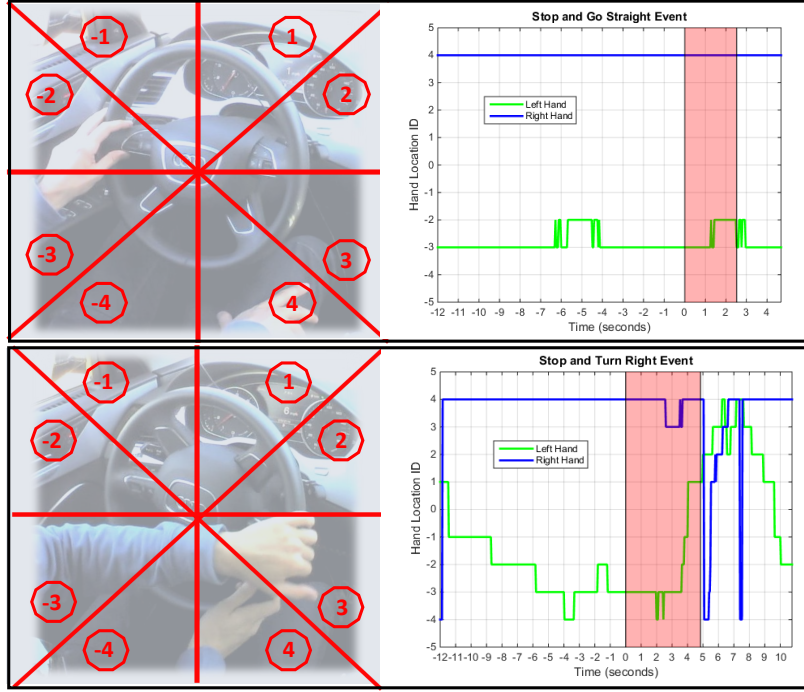


Figure 4.3: The left panel depicts the labels assigned to regions around the wheel. These are then used to map hand locations to the corresponding label values. The right panel shows an exemplar hand track sequences encoded using the method described.

4.2.2 Hand Analysis

Hand motion patterns can be leveraged as an important cue to understand driver activity. To this end, driver hands are detected and tracked in order to produce continuous trajectories using the tracker proposed in [86, 85].

Given a video sequence $V^T = \{I_1, I_2, \dots, I_T\}$ of length T , the hand tracker outputs the trajectory corresponding to each hand of the driver:

$$\begin{aligned} \mathcal{X}_{left} &= \{\vec{x}_1^l, \vec{x}_2^l, \dots, \vec{x}_T^l\}, \\ \mathcal{X}_{right} &= \{\vec{x}_1^r, \vec{x}_2^r, \dots, \vec{x}_T^r\} \end{aligned} \quad (4.2)$$

where each

$$\vec{x}_i = (x_i, y_i), \forall i \in \{1, 2, \dots, T\} \quad (4.3)$$

represents the location corresponding to the center of the hand in the image plane.

The trajectories obtained as such are not informative in their original form. In addition to this, the tracks may be fragmented and subject to spatial noise or jitters. To solve the above

issues, we propose to encode the spatial location in an intuitive yet discriminative manner. First, we change the origin of the image plane to the center of the wheel. Given the location of this center in the image plane (\vec{x}_c), the transformed location of the hands are calculated as follows:

$$\vec{x}_i' = \vec{x}_i - \vec{x}_c. \quad (4.4)$$

To make the tracks robust against noise and fragmentation, we propose a region based mapping,

$$\mathcal{M} : \mathbb{R}^2 \mapsto \{-4, -3, -2, -1, 1, 2, 3, 4\}$$

to encode tracks meaningfully. This mapping is visualized in Figure 4.3 along with sample plots for left and right hand tracks. The regions are assigned labels with a motive of associating distinct motions with larger skips in region labels. For instance, keeping one hand on each side of wheel with little or no movement would constrain your hand tracks to regions 2 and 3 for the left, and -2 and -3 for the right hand respectively. However, substantial hand motions would result in large changes in the region labels through a given temporal window.

With the above mapping in place, the final trajectory may then be represented as :

$$\begin{aligned} \mathcal{X}_{left}^f &= \{c_1^l, c_2^l, \dots, c_T^l\}, \\ \mathcal{X}_{right}^f &= \{c_1^r, c_2^r, \dots, c_T^r\} \end{aligned} \quad (4.5)$$

where,

$$c_i = \mathcal{M}(\vec{x}_i'), \forall i \in \{1, 2, \dots, T\}. \quad (4.6)$$

4.2.3 Temporal Modeling and Feature Ranking

In this subsection, we describe the feature generation and ranking process given temporal sequences from all three modalities - head, hand and eyes. From here on, we denote any generic temporal sequence as \mathcal{X} . For any given \mathcal{X} , we extract a set of statistical features (detailed in Table 4.1) and concatenate them to create a feature vector.

The above feature representation ignores any temporal structure in the data. This makes it hard to decipher the evolution of an event in time (starting 10-12 seconds before the event). To encode temporal information into our framework, we represent features in a temporal pyramid [83], where at the top level features are extracted over the full temporal extent of a video sequence, the next level is the concatenation of features extracted by temporally segmenting the video into two halves, and so on. We obtain a coarse-to-fine representation by concatenating all such

Table 4.1: Candidate features for temporal sequences

Feature	Description
<i>hist</i> ¹	Histogram
<i>mean</i> ²	Sample mean
<i>std</i> ²	Sample standard deviation
<i>min</i>	Minimum value
<i>max</i>	Maximum value
<i>mode</i> ³	Mode
<i>range</i>	Range (<i>max</i> - <i>min</i>)
<i>q</i> ₁	25 th percentile
<i>q</i> ₂	50 th percentile
<i>q</i> ₃	75 th percentile

¹ only for head and hand sequences

² only for head and eye sequences

³ only for hand sequences

features together to generate a feature vector \vec{f} .

Consider a set of feature vectors $\vec{F} = \{\vec{f}_1, \vec{f}_2, \dots, \vec{f}_N\}$, and their corresponding class labels $Y = \{y_1, y_2, \dots, y_N\}$. In the context of our study, the class labels are one of the following: *stop and turn right*, *stop and turn left* or *stop and go straight*. Given \vec{F} and Y , we train a random forest (RF) with an ensemble of 1000 decision trees on the entire corpus. The maximum depth of each tree is restricted to 10 to prevent overfitting. This RF is then used to perform feature selection across modalities. The advantage in choosing RFs over other feature selection techniques is three fold: First, we sidestep the entire process of tuning hyper-parameters through cross-validation. Second, RFs make no assumption about the linear separability of the data. Third, RFs are capable of handling different data formats like integers, floats or labels. Thus, no standardization or regularization is necessary.

The trained RF measures the importance of each feature as the averaged impurity decrease computed from all decision trees in the forest. The impurity measure used is the *gini impurity* [4], which acts as a criterion to minimize the probability of mis-classification. Using these importance scores, we rank the set of all features in decreasing order of discriminative power.

4.3 Use Case: Stop-controlled Intersection

4.3.1 Naturalistic Driving Dataset

A large corpus of naturalistic driving data was collected using a highly instrumented test vehicle dubbed LISA-A [97]. The testbed was built to provide a near-panoramic sensing field of view with synchronized internal vision, external vision, radar, lidar and GPS. Of vision sensors, of relevance to this work are four camera views; the fourth camera gives another perspective of the driver’s face.

Using this testbed, multiple drivers were asked to naturally drive on local streets and freeways in Southern California. The subjects were given no instructions on how to drive, resulting in naturalistic driving data. From the collected data, we selected events when the driver passes through stop-controlled intersections, by either turning right/left or going straight. In particular, we narrowed down the events to the stop-controlled intersections geographically illustrated in Figure 4.4. The accumulated events were gathered from 7 unique drives lasting at least 60 minutes each of 5 different drivers (4 male and 1 female). The table in Figure 4.4 shows the events considered, their respective counts and total frames analyzed with respect to one vision sensor. The duration of how long a driver stops varied depending on the intersection, the driver and the maneuver executed at the intersection. In Section IV.A, we define epochs that consistently exist in all stop-controlled intersection related maneuvers. These epochs are necessary to properly extract temporal features for analysis (see Section IV.B for more details).

4.3.2 Event Description

With an intent of analyzing driver behavior at stop-controlled intersections, we define the following epochs during each event:

- *stop* : the time at which the vehicle approaching an intersection comes to complete halt.
- *start* : the time at which the vehicle starts moving after the *stop* epoch.
- *mid* : the time halfway between the *stop* and *start* epochs.
- *end* : the time at which the driver completes the entire maneuver.

With the above definitions in place, we analyze each event for a period starting 12 seconds prior to the *stop* epoch, and ending with *end* epoch.

4.3.3 Data Driven Event Analysis

To understand the interplay between head, hand and eye modalities, we extract a set of features for each of the 24 stop-controlled intersection events, as described in section II. These

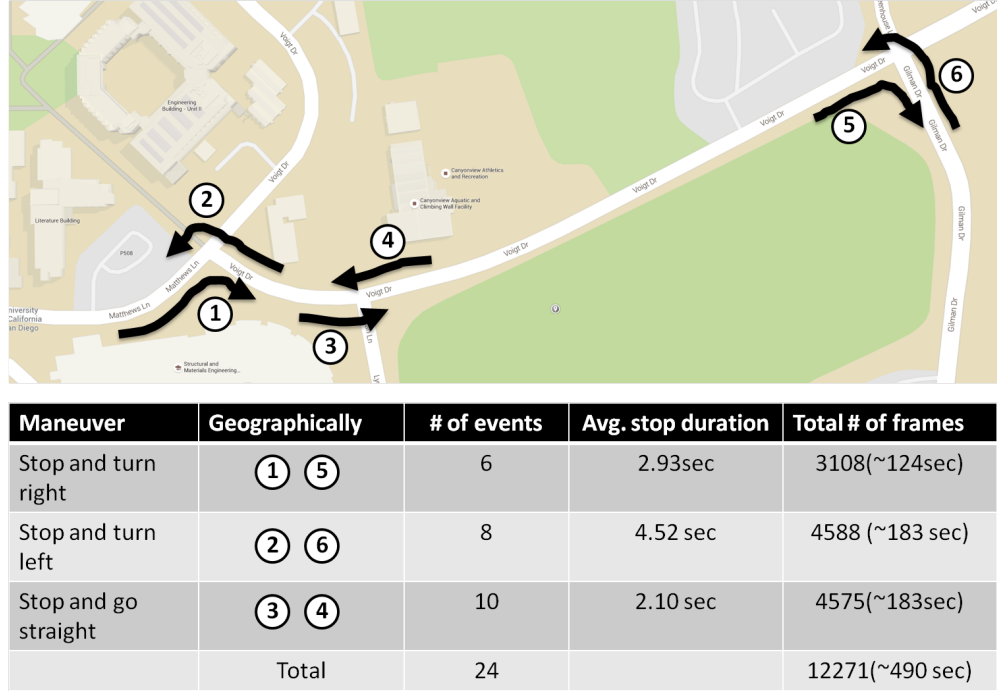


Figure 4.4: Geographical location of the stop-controlled intersections where data is extracted from and statistics on the events analyzed.

features are labeled based on the end maneuver performed by the driver (i.e. right turn, left turn, go straight).

To see how the feature importances change as a function of time, we train one RF each for every 0.25 seconds starting from 10 seconds prior to the *stop* epoch, up until and including the *stop* epoch. Each forest is trained only on features extracted up to that given point in time. Similarly, we train one RF each for every 0.25 seconds after and including the *start* epoch. Since the time interval between the *stop* and *start* epochs is variable, we simply train a RF at the *mid* epoch for each event. To encode temporal structure into the features, we use a temporal pyramid of three levels to train each RF.

For each time instant considered above, all features are ranked based on their importance scores. We consider the top 25 features and plot a histogram based on their modality. Figure 4.5 depicts the evolution of the histogram through time. In addition to the plot, we provide exemplar image sequences extracted from different time periods for each maneuver type.

The eye modality is seen to hold significance very early in the event. It provides discriminative feature up to a few seconds before the *stop* epoch. This indicates long-term planning on the drivers' part. The head modality is clearly the most discriminative of all three modalities in this context. This is indicated by a large number of highly ranked head features throughout the entire event. This modality seems to be most effective just as the vehicle is coming to a

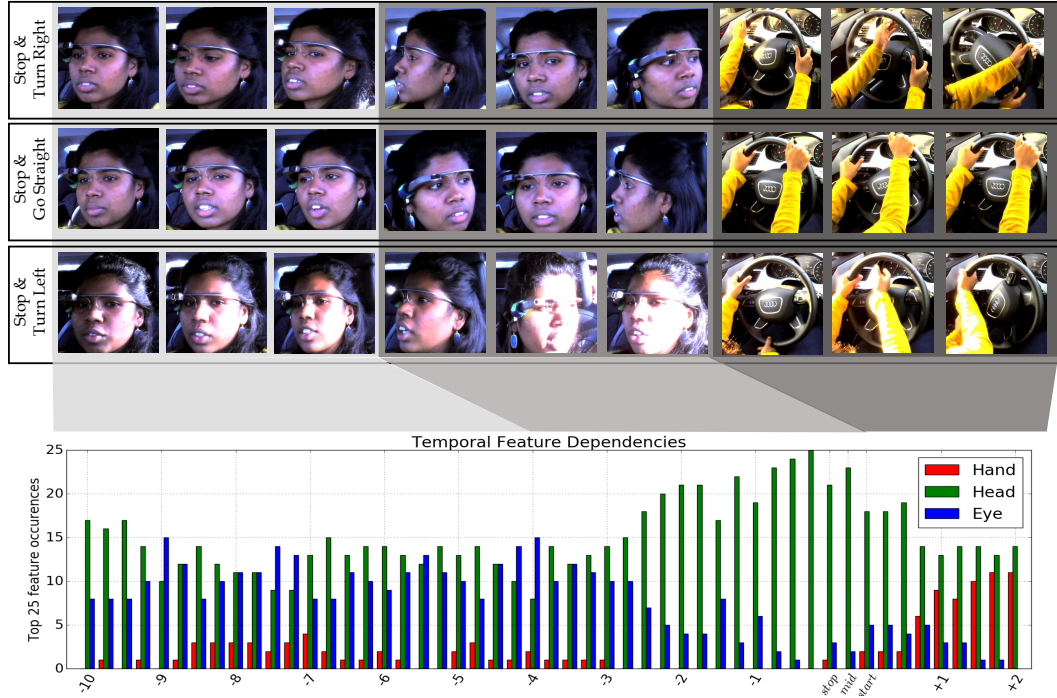


Figure 4.5: A plot of the histogram of top 25 features with respect to the modality it originates from as a function of time (bottom panel). Exemplar image sequences extracted from different time intervals for each of the 3 maneuvers. Best viewed in color.

stop, up until the *start* epoch. This corresponds to the side-to-side head movements of a driver to scan the intersection for pedestrians or vehicles. This is reflected in the corresponding image sequences from this temporal region. Finally, the hand modality comes into play later in the event. This modality is very short-sighted in the sense that its prediction window is very small in relation to other modalities. However, it provides the most reliable features for classification late in the event. This is verified by the rapid increase in discriminative features obtained from hands right after the *start* epoch. Additionally, the hand modality is a very strong indicator of preparatory movements before the actual maneuver is executed (seen from corresponding image sequences).

Figure 4.6 shows the frequency of occurrence of each individual feature in the top 25, for each modality, across all time intervals. A higher count generally corresponds to a higher relevance value for coordination analysis in the dataset. It can be observed that the *range*, *std*, *min*, and *max* features are useful across the different modalities studied. These motion cues are useful for capturing the extent of the motion in the temporal window, such as the side-to-side movements of the head (captured in the yaw). Because the hand movement is most active towards the end of the event, hand cues are selected less frequently. Among those selected, the *hist* features seem to work best for the left hand, while *range* is seen to be of most use for the

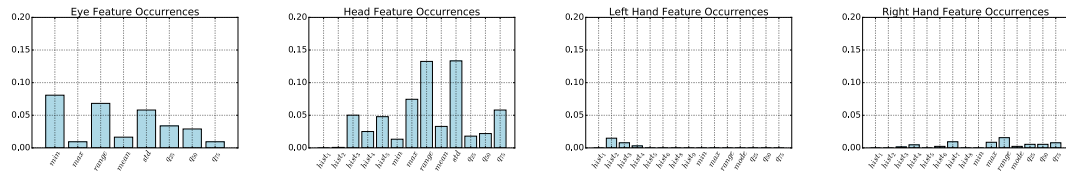


Figure 4.6: A plot of relative frequency with which each feature occurs in the top 25 across all modalities and for all time intervals.

right hand. This may be indicative of the fact that the right hand is more prone to crossover from one side of the wheel to the other in comparison to the left hand.

As another experimental analysis, the three class problem is broken into three groups of two-class problem: go straight vs. turn left, turn left vs. turn right, go straight vs. turn right. Figure 4.5 depicts the evolution of the histogram through time for each of the three binary classes. In these plots of feature importance as a function of time, there are at least three interesting observations. One is the order in which the importance of modalities occur; here modalities refer to eye, head and hand. Another is the time when the relative importance of a modality is overtaken by another modality. Finally, the duration of a particular modality’s importance over other modalities.

Elaborating on the first observation, in all three groups, the significance of modalities is seen to be temporally coordinated. The eye modality is seen to hold significance very early in the event, followed by head modality and finally by hand modality. The significance of eye modality early in the event indicates long-term planning on the driver’s part. The significance of head modality closer to the *stop* epoch corresponds to the side-to-side head movements of a driver to scan the intersection for pedestrians or vehicles. The significance of the hand modality, on the other hand, is very short-sighted in the sense that it is very close to or during the execution of the maneuver. This is verified by the rapid increase in discriminative features obtained from hands right after the *start* epoch.

The other observations of when the relative importance of a modality first surfaces and how long the modality’s importance stays above the others, unlike the first observation of the order of modality importance, varies significantly among the groups. For instance, the relative significance of head modality is earlier when classifying between turn right and go straight than when classifying between turn right and turn left; also comparatively the latter is more concentrated in significance duration as the driver approaches the intersection. This knowledge is important in many ways, one of which is when intersections are three ways and drivers have only two options for executing maneuver. For instance, as shown in Figure 4.7, if the only options are going straight or turning right, with head pose alone, the indication of executing one maneuver versus another can be as early as 6 seconds before coming to a stop at the intersection. We especially make a pitch with head pose feature because in many situations eyes maybe be

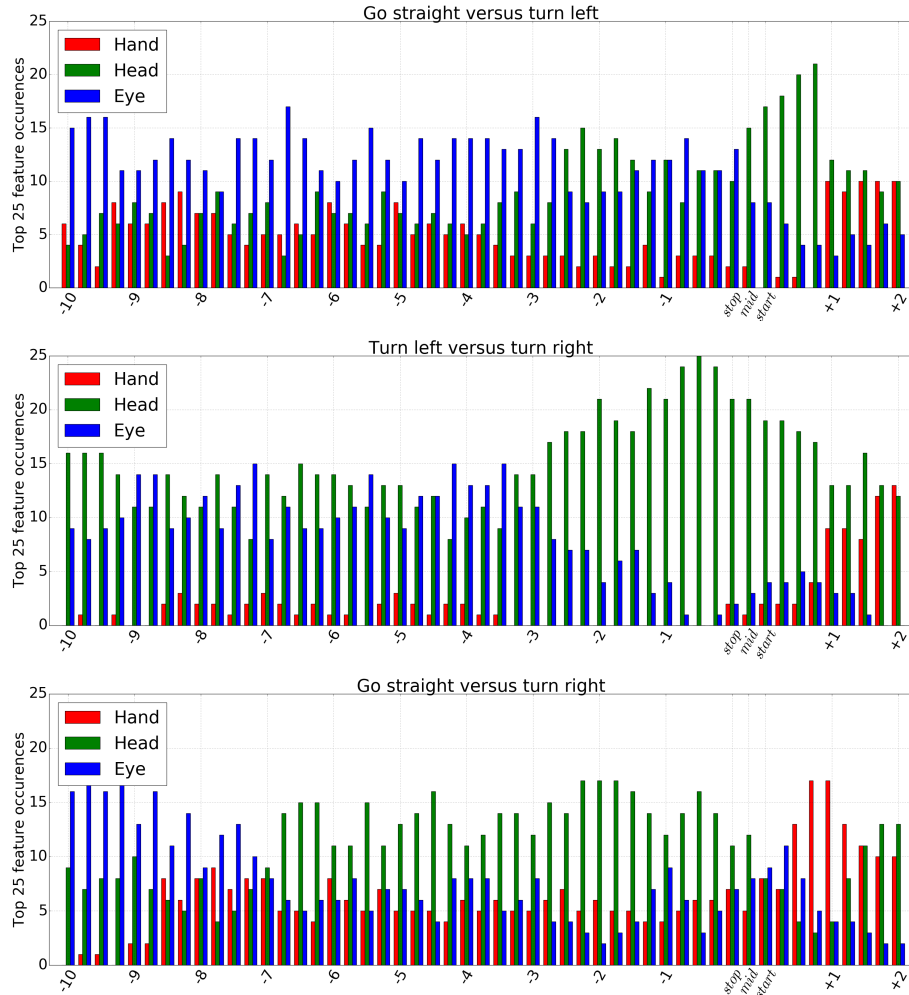


Figure 4.7: A plot of the histogram of top 25 features with respect to the modality it originates from as a function of time for three groups of two-class problem: go straight vs. turn left (top row), turn left vs. turn right (middle row), go straight vs. turn right (bottom row) .

occluded or noisy. Similarly for turning left versus turning right, strong indications may occur as early as 3 seconds.

4.4 Concluding Remarks

In this study, we provide a data-driven approach to understand and analyze the temporal interplay between three modalities: head, hands and eyes of the driver, in the context of stop-controlled intersections. A naturalistic driving dataset was employed to show that preparatory motions range in the order of a few seconds to a few milliseconds, depending on the modality, before maneuver events at stop-controlled intersections. Features generated from the head are seen to be most useful in terms of predictive power. Eye based features play an important role in

prediction at a very early stage of the maneuver, while hand features dominate towards the end. These findings are in line with the general flow of most human activities- first see, then perceive and finally actuate.

Future work encompasses extending this understanding of the temporal influence of each modality to different maneuvers performed under different contexts. Additionally, this study will serve as a basis to come up with a strong predictive algorithm for intersections. Other information sources like vehicle dynamics and external camera sensors can be integrated.

4.5 Acknowledgments

Chapter 4 is in full a reprint of material that is published in the IEEE Intelligent Vehicles Symposium (2016), by Sujitha Martin, Akshay Rangesh, Eshed Ohn-Bar and Mohan M. Trivedi, and a partial reprint of material published in the IAPR International Conference on Pattern Recognition (2016), by Sujitha Martin, Akshay Rangesh, Eshed Ohn-Bar and Mohan M. Trivedi. The dissertation author was the primary investigator and author of these papers.

Chapter 5

Naturalistic Driving Studies (NDS) database for algorithm benchmark and development

5.1 Introduction

A holistic perception and understanding of inside and outside the vehicle is absolutely necessary, and vision based techniques are expected to play an increasing role in this holistic view. The question is, how well do these vision techniques work in order to be used in time and safety critical driving situations?

In the past decade, Viola-Jones's adaboost cascade with haar features [112] for face detection has been a favorite in the community of intelligent vehicles not only because of its real-time processing speed but also because of its implementation in OpenCV which allows for easy training and testing. Works as early as [10] and as late as [63] have used the boosted cascade with haar features in order to monitor driver vigilance and generate drive reports, respectively. Viola-Jones's face detector has also been a key stepping stone in estimating driver's head pose [72], predicting driver's intent to change lane [24], monitoring driver's alertness [78] and estimating driver's gaze [100, 111]. A list of literary works in the development of driver assistance systems selected because of presenting their work on real driving data is given in Table I. The table emphasizes how the research community for driver assistance has relied heavily on Viola-Jones's face detector. Evaluations of face detection for faces from naturalistic driving in these selected publications, however, is negligible.

The driving environment presents a unique set of challenges and vision algorithms, like face detection, should be properly evaluated under these challenges. The challenges include

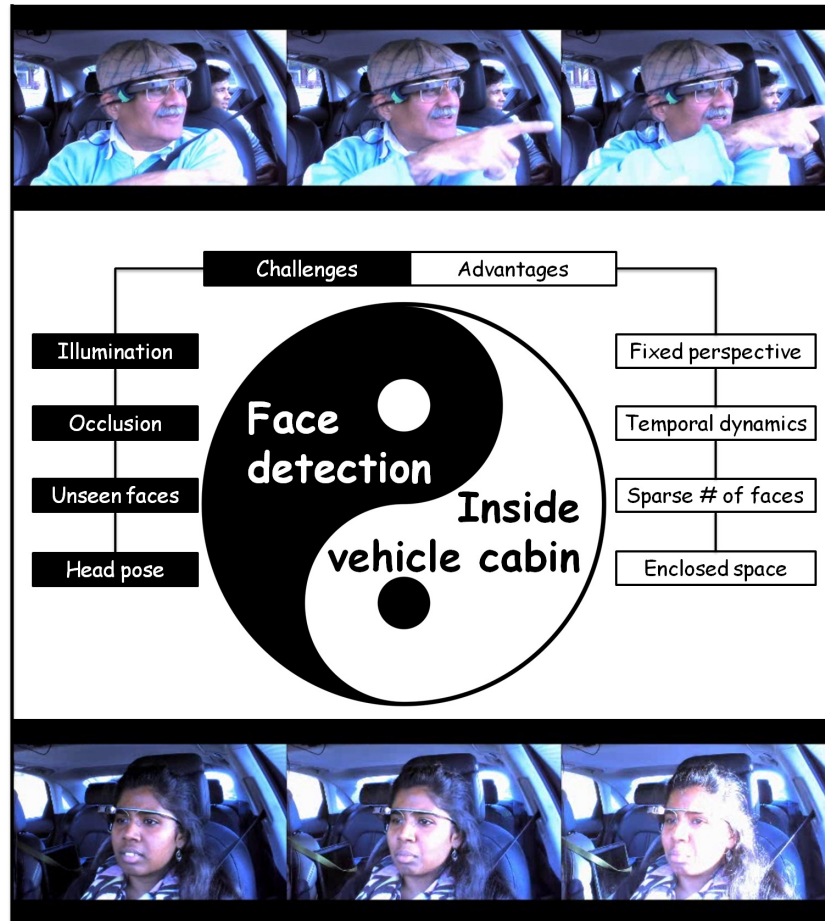


Figure 5.1: Face detection in general is challenging due to illumination, occlusion and unseen faces. However, detecting faces inside the automobile cabin has advantages such as sparse number of faces in any given image and correlated sets of images from a fixed perspective. A system that best leverages these challenges and advantages is key towards a robust face detection system when looking inside the vehicle cabin.

occlusion from object (e.g. eyewear and hats) or actions (e.g. hand and head movements), lighting conditions (e.g. sunny and cloudy), camera perspective and resolution, and unknown faces. There are, however, advantages unique to driving conditions which can aid in building a robust system (see Figure 5.1); for instance, the numbers of faces in any frame is sparse (i.e. number of faces is less than or equal to the maximum occupancy of the vehicle). In addition to evaluating vision algorithms on naturalistic driving data, it important that the data be public for proper comparison.

Vision for intelligent vehicles & applications (VIVA) [64] is a challenge set up to serve two major purposes. First is to provide the research community with a common pool of naturalistic driving data of videos from looking -inside and looking-outside the vehicle to present the issues and challenges from real-world driving scenarios. Second is to challenge the research community to highlight problems and deficiencies in current approaches and simultaneously, progress the

Table 5.1: Comparison of Selected Studies in Literature on Face Analysis Methodologies (with emphasis on face detection) as Applied to Naturalistic Driving Data

Research Study	Objective	Face Detection Method	Evaluation on Face Detection or Head Pose
Fletcher et al., 2003 [36]	Monitoring driver’s visual behavior	Commercial	Neither on NDS
Bergasa & Nuevo, 2006 [10]	Monitoring vigilance	Viola-Jones	87.5% face detection accuracy in night time driving
Fletcher & Zelinsky, 2009 [37]	Driver inattention detection	Commercial	Neither on NDS
Tran & Trivedi, 2009 [104]	Driver activity analysis	Skin color segmentation & Viola-Jones	Neither
Murphy-Chutorian & Trivedi, 2010 [72]	Head pose estimation	Viola-Jones	90% face detection accuracy and MAE > 10° & STD ≈ 17° in head pose
Morris, Doshi & Trivedi, 2011 [24]	Lane change intent prediction	Commercial	Neither on NDS
Oyini Mbouna et al., 2013 [78]	Alertness monitoring	Viola-Jones + adaptive template matching	Neither
Ahlstrom et al., 2013 [3]	Distraction warning system	Commercial	Neither
Tawari, Martin & Trivedi, 2014 [95]	Continuous head pose	Viola Jones & MTS	96% success rate with head pose
Tawari et al., 2014 [100]	Attention estimation and merge recommendations	Commercial	Neither
Liu & Graeser, 2015 [61]	Face detection	Viola-Jones	Face detection accuracy of 97.9% on only faces with shades.
Paone et al., 2015 [79]	Face detection, head pose & coarse direction	Commercial	89% face detection accuracy & MAE of 6° in yaw
Rodemerk et al., 2015 [87]	Predicting driver’s turn intentions	Commercial	Neither
Jain et al., 2015 [52]	Anticipating maneuvers	Viola-Jones	Neither

development of future algorithms. There are benchmarking competitions and databases available for general vision problems, such as the KITTI Vision Benchmark Suite [43], which in comparison is the closest to VIVA due to data collected from driving. However, one of the major difference is, VIVA contains datasets and challenges for looking-inside, while KITTI does not.

In this chapter we introduce one part of the VIVA challenge, namely the VIVA Face challenge. The current face challenge includes face detection and head pose. This section describes the dataset for and current submissions to the respective challenges.

5.2 Dataset Description

The VIVA Face dataset is composed of selected images from day time naturalistic driving data collected at LISA-UCSD, as well as selected images from, to the best of our knowledge, naturalistic driving videos from YouTube. The images are selected from a total of 39 video

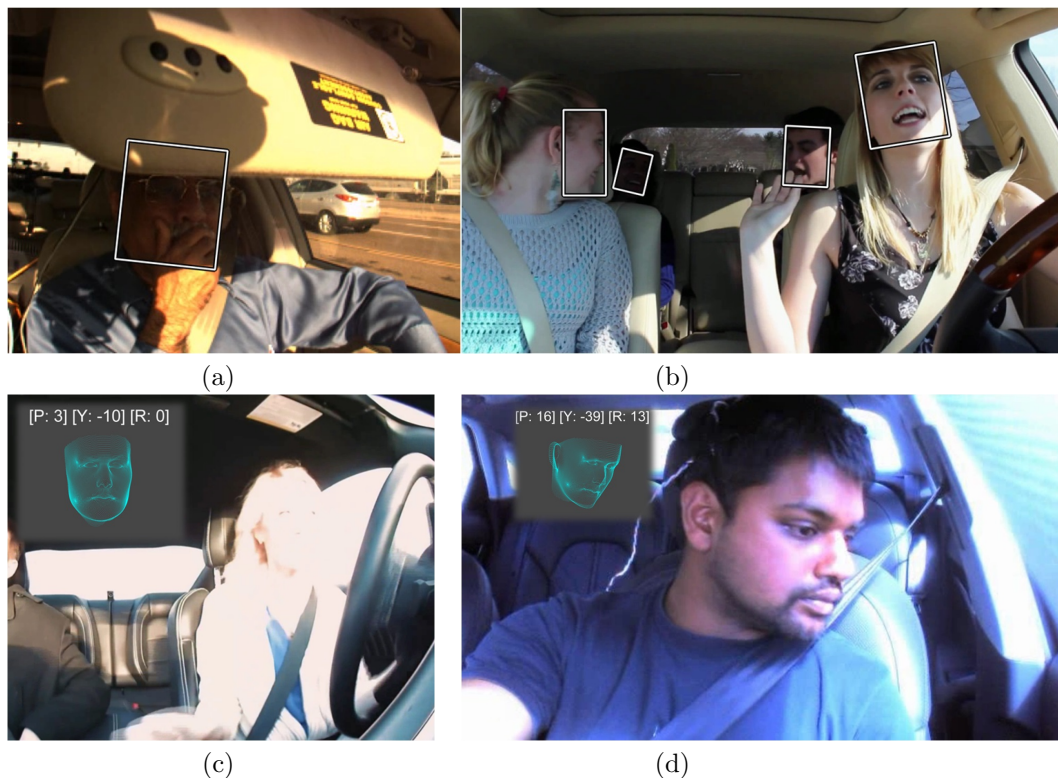


Figure 5.2: Challenges in the dataset: varying illumination (a,b,c), head rotation away from frontal (b,d), occlusion (a,b,c), multiple faces (b). Realistic driving scenarios are prone to such volatile conditions. Thus the inclusion of these settings in various degrees in the faces dataset is vital for proper performance evaluation and benchmarking.

sequences and are selectively chosen based on harsh lighting conditions and facial occlusions (e.g. hands or the vehicle’s sun visor), as shown in Figure 5.2, in order to highlight some of the challenges in face detection and head pose estimation in a naturalistic vehicle environment. The viewpoints range from observing only the driver to the entire vehicle cabin, which resulted in at least one face per image to a maximum of four faces. To keep computational time roughly the same per image, all images were resized to have the same height as the LISA-UCSD dataset at 544 pixels in height, while keeping the aspect ratio of the original images. All images provided in the dataset are only for testing and is available for download here [64]. No training images are provided, but participants are referenced and encouraged to use publicly available face databases for training (more details are available on the website).

5.2.1 Ground Truth Generation

For our two challenges, face detection and head pose estimation, our ground truth is composed of 4 corner points and 3 angles corresponding to a face box and head pose, respectively. For the face detection challenge, results will be evaluated against our ground truth face

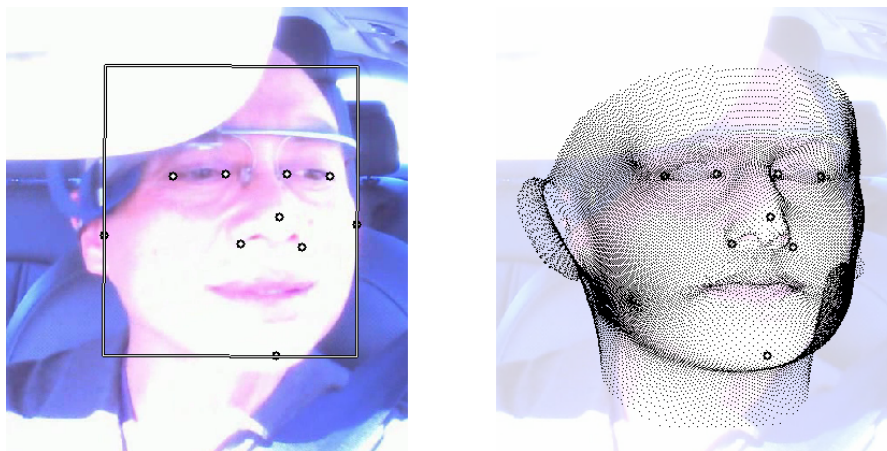


Figure 5.3: Example annotations from the VIVA-Face dataset. (Left) shows the ground truth face box derived from the 10 annotated points. (Right) shows an overlay of 3D general face model used to verify the quality of head pose estimation.

box represented by 4 corners, e.g. $(x_1, y_1), \dots, (x_4, y_4)$. Results for head pose challenge will be evaluated against our ground truth angles for pitch, yaw, and roll of the head. Both ground truths are derived from 10 landmark annotations composed of the four eye corners, nose tip, left/right nose corners, chin, and left/right most part of the face (e.g. the left and right side burn areas by the ears for a frontal face, and left side burn area and nose tip area for an extreme profile face looking towards the right), as shown in Figure 5.3. During this process, the annotator also subjectively labeled each part (i.e. left eye, right eye, nose, and mouth) as occluded or not occluded. Each part was considered occluded if four pre-defined points within each part region was not visible.

To generate the face box from the 10 landmark annotations, the bottom edge of the provided face box is placed at the chin and approximately parallel to the line fitted through the four eye corners. The top edge of the face box, set to be parallel to the bottom edge, is placed above the four eye corners with distance being a fraction of the distance between the eyes and the chin. The left/right edges are placed at the annotated left/right most part of the faces and set to be perpendicular to the top/bottom edges. A visualization of the annotated landmarks and resulting face box is shown in Figure 5.3. The distribution of the height and width of the face boxes is visualized in Figure 5.4a, showing a large range of face sizes. One thing to note is that the face box height is generally slightly larger than the width.

Head pose is annotated by estimation from the annotated facial landmarks (i.e. four eye corners, nose tip, left/right nose corners) and their corresponding points on a 3D generic face model [65], as shown in Figure 5.3. The quality of this estimation is verified by overlaying the 3D generic face model and visually inspecting to ensure it is accurate. In cases where it is not accurate, the annotated points were re-adjusted slightly such that both the annotated landmarks

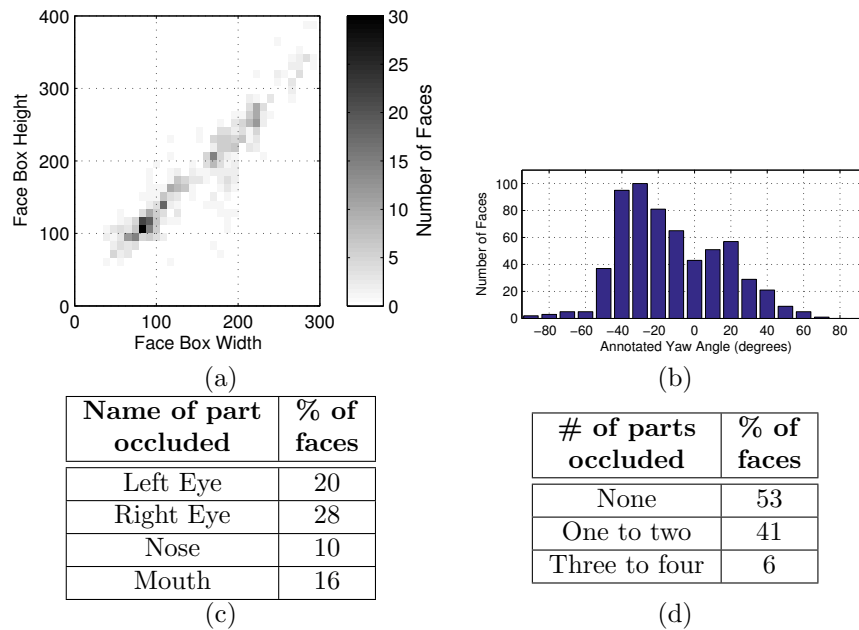


Figure 5.4: VIVA-Faces Dataset Statistics: (a) Annotation bounding box size shows a good spread of different face sizes. (b) Number of faces by head pose in yaw rotation angle. (c) Percentage of images where face is occluded with respect to (c) part type (i.e. left eye, right eye, nose and mouth) and (d) number of parts.

and the head pose remain precise. Yaw rotation angle distribution of all the annotated head pose (Figure 5.4b) shows the variations in the images with respect to head pose. The odd two-peaked distribution is due to the camera positioning being off to the side on the left or right, so as to not block the driver’s view, with the faces generally looking straight ahead.

Variations in faces occur in at least one other form, that is occlusion. Occlusions range from no part of the face to most of the face. In order to quantify the percentage of occlusion subjectively, we consider four major parts of the face: left eye, right eye, nose, and mouth. Occlusion statistics of the VIVA-Face dataset is provided in the tables in Figure 5.4c,d.

5.3 Metrics and Performance Evaluation

The dataset as described earlier has many variations in faces including head pose, illumination, and object occlusion. These particular examples can all cause some form of occlusion to the face as shown in Fig 5.2. In order to evaluate a face analysis system on its strengths and weakness it is necessary to evaluate with respect to these variations by partitioning the dataset. And within each variation, standard evaluation metrics are used to evaluate face detection and head pose estimation. For instance, how well does the face detection and head pose estimation perform under varying occlusion (i.e. no occlusion to partial/full occlusion of face)? To evaluate

the performance under these variations, we split the evaluation into three categories: all faces, faces without any occlusion, and faces with at least one part occluded. However, in a given image of more than one face, each face may be occluded to different degrees. If an image contains both non-occluded and occluded faces, we set an ignore flag for the ground truth face boxes which do not belong in a given category as done in [22] so that detections in the ignore regions are ignored and do not count as a false positives. If the area of the intersection of the detection and the ignored ground truth divided by the area of the detection is greater than 0.5, then the detection is ignored, e.g. a detection entirely inside an ignore region will result in a value of 1.0 and is thus ignored.

Within each category, standard performance metrics are used to evaluate face detection and head pose estimation. For face detection, precision recall (PR) curve with the well accepted overlap of greater than 50% between annotation and detection is used to classify the detections along with their scores into true versus false positive. For head pose estimation, we first compute the detection rate (DR) as simply the maximum recall value in the precision recall curve, i.e. we ignore the score value for the detections and compute the percentage of ground truth faces that the head pose estimator was able to detect with more than 50% overlap which we will call *correctly detected faces*. For a given head pose estimator, we compare only their correctly detected faces' head pose with the corresponding ground truth to compute the success rate (SR) along with mean (μ_{AE}) and standard deviation (σ_{AE}) of the absolute error. The success rate is defined as the percentage of head pose that were estimated within 15 degrees of the ground truth. For this paper, we only present results for yaw angle due to limited submissions with the other two angles.

Generating PR-curves for face detection and statistics for the head pose estimation challenges with the dataset as a whole versus partitioned based on occlusion, as shown in Figure 5.5 and Table 5.2 respectively, gives perspective on a given system's performance with respect to occlusion. In Figure 5.5, a few different face detectors, which are open source and publicly available online (i.e. Boosted cascade with Haar [112], Boosted cascade with ACF [21], Mixture of Tree Structure [122]), are shown as benchmarks for the VIVA Face dataset. In Table 5.2, we present evaluation results on yaw angle of the head pose with one benchmark system. We also present results for verified submissions from those who wish to remain anonymous until their work has been published. While the results shared in this are preliminary, the authors are in correspondence with researchers working on state-of-the-art methods [49] to keep the benchmarking results relevant and up-to-date for the vision and intelligent vehicles community.

5.4 Concluding Remarks

The Laboratory for Intelligent and Safe Automobiles (LISA) team at UCSD has made a major commitment to offer its data sets to the worldwide research community so that anyone

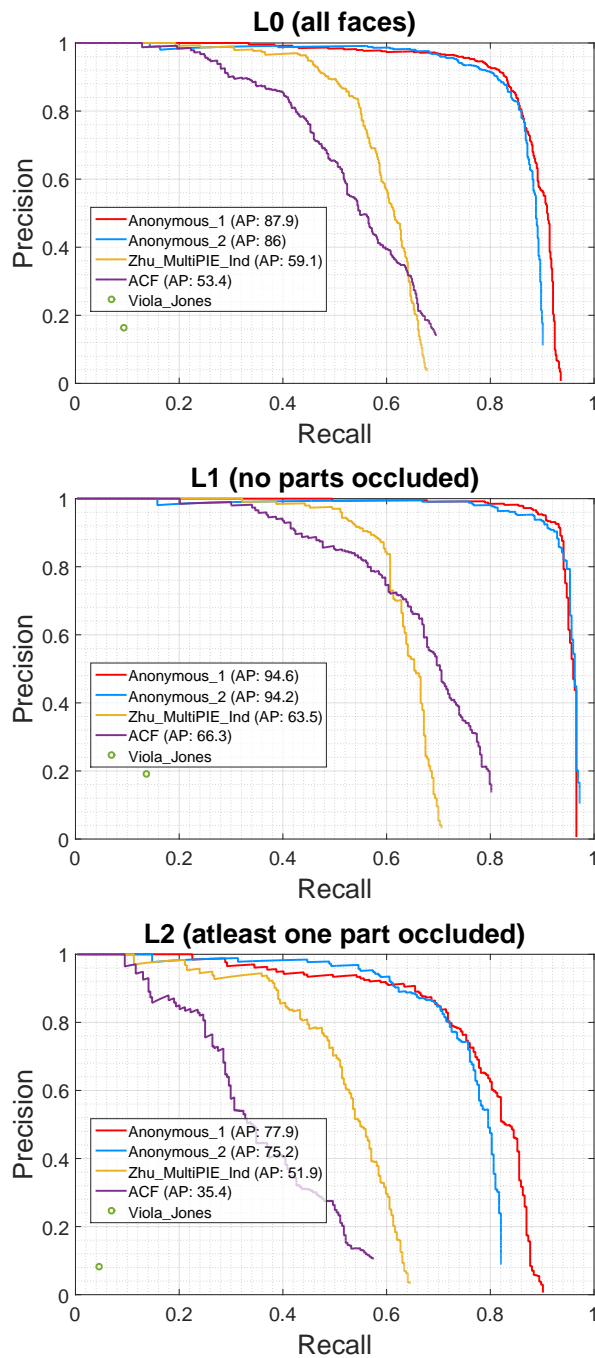


Figure 5.5: Benchmark evaluations for face detection on the VIVA-Face dataset for (left) all 607 faces in 458 images, (middle) 323 non-occluded faces in 289 images, (right) 284 faces with at least one occluded part in 240 images.

interested in the research can get open access and evaluate their own algorithms and benchmark them against others. The main motivators for such effort is to engage wider community of

Table 5.2: Evaluation results from benchmark and submissions on yaw angle estimation of the head pose. The evaluation is split by occlusion levels: L0 (all 607 faces in 458 images), L1 (323 non-occluded faces in 289 images), L2 (284 faces with at least one occluded part in 240 images). DR, detection rate, is the percentage of images for which the face detector was able to detect a face with at least 50% overlap with the ground truth face. SR, success rate, is the percentage of correctly detected faces for which the estimated yaw angle was within 15 degrees of the annotation. μ_{AE} and σ_{AE} are the mean and standard deviation of the absolute yaw error (in degrees), respectively, calculated only from the correctly detected faces.

Benchmark/ Submission	Occlusion Level											
	L0 (all faces)				L1 (no parts occluded)				L2 (at least one part occluded)			
	DR	SR	μ_{AE}	σ_{AE}	DR	SR	μ_{AE}	σ_{AE}	DR	SR	μ_{AE}	σ_{AE}
Anonymous.3	0.76	0.75	10.7	9.2	0.85	0.80	9.3	6.6	0.66	0.66	12.7	11.8
Anonymous.4	0.60	0.67	13.9	12.5	0.71	0.69	13.7	12.7	0.48	0.64	14.2	12.0
Zhu.MultiPIE	0.68	0.65	15.6	16.3	0.71	0.70	13.8	13.8	0.64	0.59	17.9	18.7

students, scholars, developers in this exciting and challenging field and also to provide credible, quantifiable benchmarks for evaluation and rapid progress in the field.

In this regard, the paper presented important issues, challenges, and metrics associated with development of robust vision based systems for intelligent vehicles. Specifically, this paper presented issues related with face detection and head pose in static images. A major step in the future direction for face related challenges alone is in the temporal domain; this includes face tracking, head pose tracking, gaze estimation, activity analysis and expression recognition. A parallel effort is also in the integrative vision framework of looking-in (e.g. faces, hands) and looking-out (e.g. vehicles, traffic signs) for activity recognition and intent prediction, to name a few.

5.5 Acknowledgments

Chapter 5 is in full a reprint of material that is published in the IEEE Intelligent Vehicles Symposium (2016), by Sujitha Martin, Kevan Yuen and Mohan M. Trivedi. The dissertation author was the primary investigator and author of these papers.

Chapter 6

Balancing privacy and safety: Protecting driver identity in naturalistic driving video data

6.1 Introduction

Although there is a widespread consensus for intelligent vehicles to improve safety, the study of driver behavior to design and evaluate intelligent vehicles requires large amounts of naturalistic driving data. In current literature, however, there is a lack of publicly available naturalistic driving data largely due to concerns over individual privacy.

Camera sensors looking at a driver, which are an integral part of intelligent vehicles [108, 106, 23], are of particular concern for invasion of privacy as they can be used to recognize the drivers identity. Publicly available naturalistic driving data should be used for the study of driver behavior to improve driving safety and not to implicate drivers on their driving behavior. Typical protection of the privacy of individuals in a video sequence include blacking out and blurring of faces or people, which is commonly referred to as deidentification. Although this will help protect the identities of individual drivers, it impedes the purpose of sensorizing vehicles to look at a driver and his or her behavior. In an ideal situation, a deidentification algorithm would protect the identity of drivers while preserving sufficient details to infer driver behavior (e.g., eye gaze, head pose, and hand activity).

Many existing state-of-the-art algorithms on driver behavior are trained to work on undistorted raw images from camera sensors with possible privacy implications, as shown in Figure 6.1. Trivedi et al. [107], for example, proposed a smart airbag system by sensing the passenger occupancy and the body posture to ensure safe airbag deployment. Similarly, in

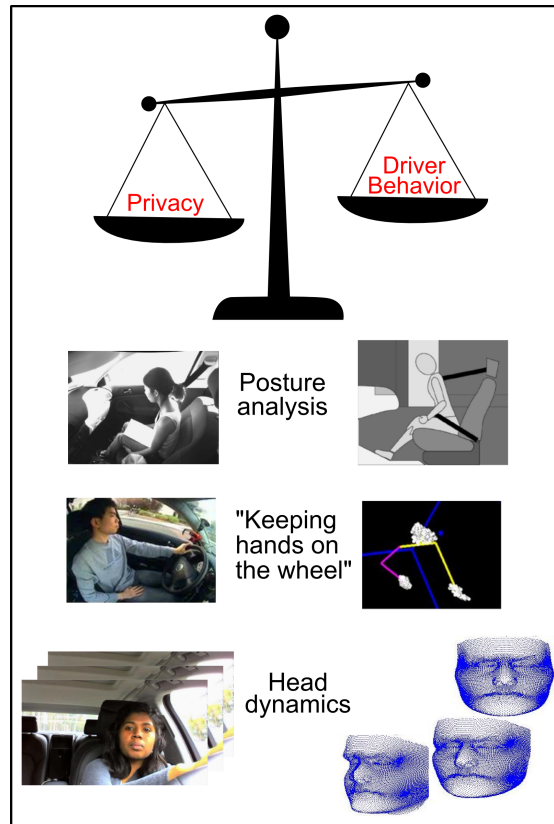


Figure 6.1: Illustration of privacy implications in using existing systems on raw camera sensory output to infer driver behavior: posture analysis [107], “keeping hands on the wheel” [104], and head dynamics [95].

[104, 105], cameras pointed at the drivers face and hands are used to discern whether the driver is keeping hands on the wheel and eyes on the road, which is a general tip for safe driving. In a larger scheme of looking inside and looking outside, Doshi et al. [24] used vision-based head dynamics estimation as a proxy for gaze direction to predict a drivers intent to change lanes. To eventually realize these active safety systems in commercial vehicles, parallel efforts need to be made on both privacy protecting and driver-behavior monitoring systems.

A drivers absolute gaze direction is of particular interest because it can be used as a proxy to determine what information the driver is processing [101, 26, 27]. We say absolute gaze, which should not be confused with eye gaze relative to head pose, and henceforth any reference to gaze estimation implies absolute gaze. Interestingly, the same facial features that are explicitly or implicitly used for gaze estimation play a key role in recognizing a persons identity. Any form of deidentification on video sequences of looking at a driver can only degrade the performance of gaze estimation. However, the degradation in performance could be minimized by using appropriate methods for deidentifying drivers. Therefore, we propose and implement a deidentification filter, which, semantically, protects the identity of the driver and preserves the behavior of the driver;

to our interest, it should preserve gaze direction. With such filters, researchers may be more inclined to publicly share deidentified naturalistic driving data. The research community can then tremendously benefit from large amounts of naturalistic driving data and focus on the analysis of human factors in the design and evaluation of intelligent vehicles [102, 110, 77].

The remainder of this paper consists of the following. Section 6.2 extensively describes the properties and challenges of deidentification in the driving scenario and how it compares to deidentification in other applications. Section 6.3 describes one possible design for a deidentification filter and explores its varying parameters. Section 6.4 presents how the driving data are collected, which filter parameters are used, what is gathered from the face recognition and gaze-estimation user study, and the discussion on the level of deidentification and of gaze estimation. Section 6.5 provides the concluding remarks with future directions.

6.2 DeIdentification Filter: Definition, Challenges and Related Research

A de-identification filter is that which takes an image or a sequence of images where a person’s identity could be recognized and makes it unrecognizable, while at the same preserving the necessary information for which the image was captured. Semantically, in an ideal situation, a de-identification filter applied to a raw image of looking at the driver will output a similar image where the identity of the driver is protected and the behavior of the driver is preserved. In the first part of this section, we briefly discuss the properties associated with a de-identification filter and the challenges in addressing those properties when looking at the driver. In the second part, we discuss work related to de-identification of people in other applications.

6.2.1 Looking inside the vehicle

There are two main properties associated with de-identification filters: region of interest and what should be preserved. First, given an image containing the driver’s face, whether de-identification should be applied locally inside a **region of interest** (e.g. the driver’s face) or over the entire image. The former is challenging because if the face is detected or tracked incorrectly, then de-identification on the wrong portion of the image leaves the driver’s face vulnerable for identification. Robustness of face tracking algorithms is made difficult by the changing illumination conditions and large spatial head movements present in typical naturalistic driving data. While applying de-identification uniformly to the entire image is appealing for its limited or lack of dependence on face detection or tracking modules, it would, however, mean sacrificing resolution or key details in the background, which do not reveal the driver’s identity. For example, distorting the whole image can deteriorate the performance not only of recognizing driver’s identity, but also of inferring whether the driver is wearing a seat belt or in inferring

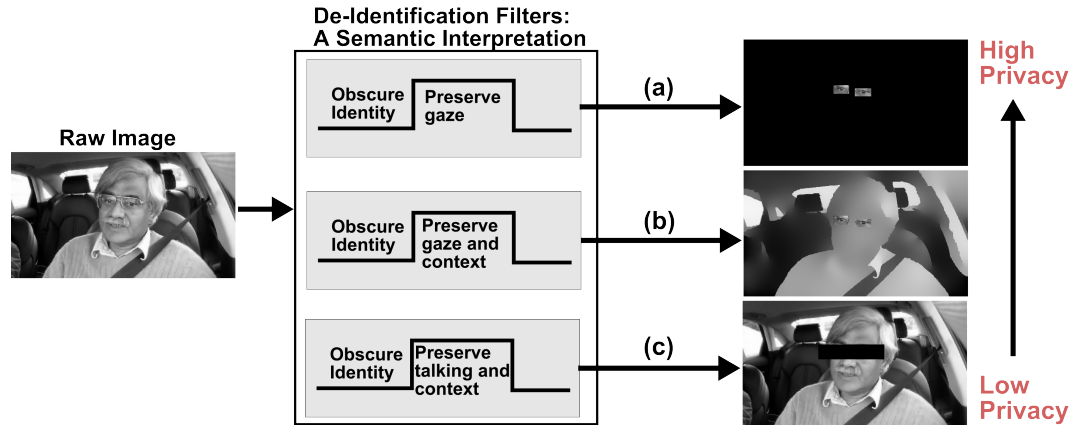


Figure 6.2: Illustrating three different de-identification filters, which semantically share the same goal of obscuring driver’s identity and preserving driver’s behavior, but in different degrees. The last filter demonstrates de-identification by region of interest (e.g. eyes), which is very weak in protecting privacy but reveals behaviors such as whether the driver is talking, frowning, smiling or yawning, and roughly where the driver is looking. The first and second filters are geared towards gaze direction estimation without and with spatial context, respectively. As a decreasing number of driver’s face parts are preserved the less likely it is to identify the driver.

driver’s hand movements.

This leads to the second issue of **what should be preserved** in the process of de-identification. Driver fatigue monitoring systems would benefit largely from preserving mouth behavior such as the number of times the driver yawned [114], head dynamics behavior such as the nodding frequency [91], and eye behavior such as the proportion of time the eyes are closed (PERCLOS) [81] and fixed gaze [10]. In addition to fatigue monitoring, preserving eye gaze behavior can be used as a proxy to determine what information the driver is processing [101]. For example, coarse gaze direction estimation is a good indicator of driver’s intent to change lanes [25]. Figure 6.2 illustrates three different de-identification filters which share the same goal of obscuring driver identity and preserving driver behavior, but in varying degrees. Therefore, depending on the study of driver behavior, specific de-identification algorithms can be designed.

These key ideas on de-identification of driver’s face on naturalistic driving data are to some extent addressed in literature for de-identification of faces and people in other applications, as described in the following section.

6.2.2 Related Studies

The term “De-Identification” has been popularly used in literature to address the concern of privacy invasion from sensorized environments. A sensorized environment can be anything from when a photographer captures a moment in time, where the camera is the sensor and the scene is the environment, to constant surveillance, such as airports and convenience stores. The need for de-identifying people in sensorized environments are typically for two reasons, either the person is

not intended to be in the image or the presence and action of the person is intended but not their identity. The former is a prevalent reason for de-identification in applications like Google-street view [41, 38] while the later has applications in surveillance [2] and vehicle cockpit. A summary of selected studies are described and illustrated in Table 6.1, with the following important elements to be considered about respective means of de-identification:

- Approach: What is the process and object of de-identification? Is it a region based de-identification or de-identification applied to the entire image?
- Preserving: In the process of de-identification, what was preserved? Was it the background, the scene, the action of people, etc.?
- Application domain: What is application domain that requires de-identification? Is it surveillance (e.g. airports, convenience stores, banks)? Is it intelligent vehicle space (e.g. Google street view)?
- Sample: An illustration of respective de-identification applied to an image portraying respective application domain.
- Evaluation: How was the recognition of people or faces evaluated? Is it by user studies or by machine algorithms?

To put our contributions in perspective, we present related research on de-identifying people with application to real-world, unconstrained environments such as intelligent vehicle space and surveillance. Note that we compare our work mainly with de-identification in other applications because there is a lack of studies in driver specific de-identification in literature.

Intelligent Vehicle space

Intelligent Vehicle space where vehicles instrumented with camera sensors observe the environment outside and/or inside the vehicle for various purposes. Google street view is a popular context for observing the environment outside an instrumented vehicle to capture geographical information. In the 360° panoramic view, the street view car captures not only the appearance of location specific objects such as buildings, billboards and street signs but also privacy sensitive material such as people and license plates. Google protects individual privacy by introducing a system that automatically blurs faces and license plates [41]. However, recent studies have shown that face is one of many identifiable features associated with people, such as silhouette, gait and articles of clothing [15, 119]. To this end, many researchers have proposed to remove persons from the Google street view and replace them with background pixels using multiple views of the same scene [38] or with similar looking pedestrians from a controlled dataset [73].

Looking inside the vehicle has been equally popular among researchers of intelligent vehicle space. While the matter of privacy invasion has been of much concern, to the best of our

knowledge, no work has explicitly proposed or evaluated de-identification algorithms for inside the vehicle. We say explicitly because some researchers have used a derivative of camera sensor data to do further analysis. In [107], Cheng and Trivedi represent data from multiple camera sensors as voxel data and perform occupant posture analysis. Similarly, [104] uses EXtremity Movement OBservation (XMOB) for 3D upper body pose tracking to determine whether the driver’s hand is on the wheel. However, these derivatives of raw data require further analysis to address privacy concerns. In this study, we intentionally apply de-identification filter on data from camera looking at the driver, and quantify the level of de-identification and the effects of de-identification on estimating driver’s gaze direction.

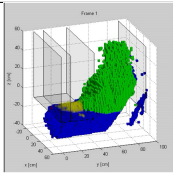

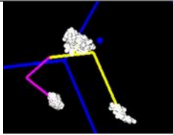



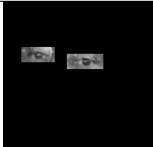
Surveillance

Surveillance of public areas such as airports, convenience stores and shopping malls is becoming increasingly prevalent. In many of these sensorized environments, the scene and the actions of people to determine irregular activities is highly desired over the identity of individuals. This can be accomplished in one of two ways: first, by overlaying distinguishable features, such as face, with elementary shapes (e.g. ellipsoids, rectangular boxes) [34, 90] and second, by segmenting the person from the background and representing the raw pixels differently [84, 2]. While overlaying faces with ellipsoids preserves raw pixels on the rest of the body parts [90] and may allow for better action recognition, it leaves the rest of the distinguishable features vulnerable to person identification. On the other hand, replacing the entire person with a rectangular box guarantees de-identification and can provide rudimentary information such as the velocity and the direction in which the person is moving [34], but lacks the detail to do finer action recognition such as use of cellphones and carrying of luggage.

Segmentation of person from background, however, provides finer details on actions of individuals. Person segmentation is more feasible in surveillance with stationary cameras to allow the system to learn the background and distinguish it from the foreground. Once segmented, raw pixels of individual persons can be replaced with the background or color coded depending on actions of interest [84]. Interestingly, [2] addresses the concern over recognizing individuals using gait information by smooth temporal blurring on segmented persons and test for recognition by humans in a user study. While algorithmic recognition systems using gait information are yet to materialize, gait information is a key distinguishing factor in human recognition system. Future de-identification algorithms on videos should collectively de-identify images in the sequence. This is especially true for videos looking at the driver inside the vehicle, where a number of de-identified frames containing one driver can accumulate to reveal the driver’s identity.

In this chapter, we first address the concern of de-identifying driver’s faces in individual frames independent of other frames in the video and leave collective de-identification over the entire video sequence to future studies.

Table 6.1: Comparison of selected studies in literature of de-identification of faces or people.

Research Study	Approach	Preserving	Context	Sample	Evaluation
Cheng & Trivedi, 2004 [107]	Voxel reconstruction of scene using constrained multi-camera setup	Action	Looking inside the vehicle		Not applicable
Schiff, Meingast & Mulligan, 2009 [90]	Obscures faces with solid ellipsoidal overlays	Scene and action of people	Surveillance		Hand labeled: false positives and false negatives.
Tran & Trivedi, 2009 [104]	Extremity movement observation for 3D upper body pose representation	Action	Looking inside the vehicle		Not applicable
Flores & Belongie et al., 2010 [38]	Removing pedestrians using multiple views	Scene	Google street view		Visual samples
Agrawal & Narayan, 2011 [2]	Applying transformations on individuals in a conservative voxel space	Scene and action of people	Surveillance		Algorithmic face & people detection results, and user study for recognition
Nodari et al., 2012 [73]	Replacing pedestrians with similar pedestrians from a controlled database	Scene and action of pedestrians	Google street view		Algorithmic detection, segmentation, matching and replacing results
Proposed framework	Isolated segmentation of eyes on a black background	Driver's eye gaze	Looking inside the vehicle		User study of face recognition and gaze estimation

6.3 How to Protect Identity and Preserve Gaze?

As presented in Section 6.2.2, there are many ways of de-identifying faces and people. The matter, however, lies in what should be preserved after applying de-identification, such as the action. In this study, we are especially interested in preserving driver's gaze. There is a trade-off, however, between preserving driver's gaze and protecting the driver's identity, because

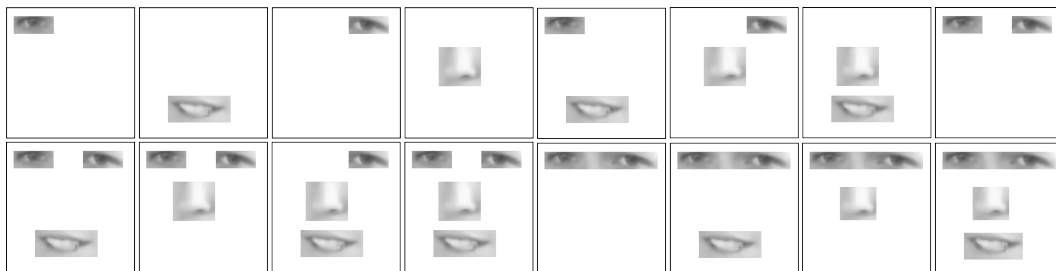


Figure 6.3: Illustration of different combinations of patches around facial landmarks to estimate or predict driver behavior while protecting driver’s identity. Combinations which include eyes can provide information such as the driver’s gaze and blink rate. Combinations which include mouth can be used to discern whether the driver is talking or to interpret the driver’s emotion (i.e. happy, sad). However, some combinations are more susceptible than others to face recognition.

the same facial features that are explicitly or implicitly used for gaze estimation play a key role in recognizing a person’s identity. In the following sections, we explore methods to de-identify the driver yet allow for gaze estimation by preserving key facial regions in the foreground and obscuring other regions in the background.

6.3.1 Foreground Preservation for Gaze Estimation

Facial recognition occurs due to a combination of facial features: eye shape, eye color, hair texture, nose size, mouth shape etc. When these facial regions are isolated, it could represent many number of people. Yet isolated facial regions provide powerful information about the state of the driver. For example, eyes alone could be used to estimate the driver’s gaze, calculate blink rate, etc. and mouth alone could discern whether the driver is talking, frowning, smiling, etc. While it would be more powerful to preserve a combination of facial regions to derive the complete state of the driver, certain combinations are more susceptible than others to face recognition. Figure 6.3 shows a subset of these combinations of facial regions.

Gaze estimation can be accomplished using any one of the combinations of facial regions in Figure 6.3. Some, however, are more robust to large head movements, facial deformations, lighting conditions etc. In this study, we explore the benefits of preserving only the region around the eyes. To extract the location of the eyes, facial landmarks can be reliably detected using state of the art algorithms such as Constrained Local Model [18], Pictorial Structure Matching [33] and Supervised Descent Method [117]. These detected facial landmarks are used not only to extract the eye regions, but also to estimate head pose. Head pose is computed using seven facial landmarks (eye corners, nose corners and nose tip) and their corresponding points on a 3D-generic face model [65]. Using head pose, a de-identification scheme which preserves only one-eye can determine which eye is more visible from the camera perspective.

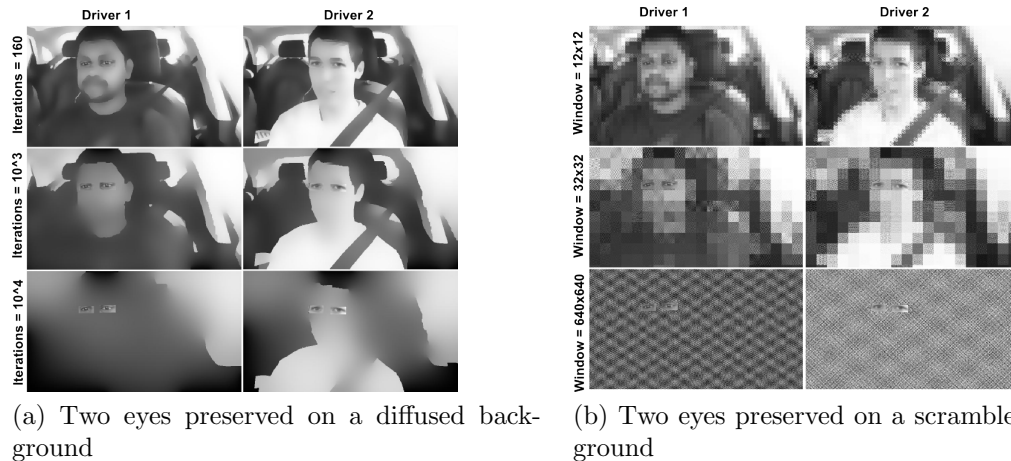


Figure 6.4: Visual demonstration of de-identification by preserving the eyes in the foreground while scrambling or diffusing the context in the background. (a) Scrambling in the transform domain occurs inside non-overlapping blocks. Small block size (window = 4×4) leaves the driver in the image vulnerable to identification, whereas a large block size (window = 640×640) is equivalent to replacing the background with random noise. (b) Anisotropic diffusion preserves edges and lines while smoothing out finer details in the image. Illustrated are images from two different drivers under varying iterations of diffusion. As shown with two different drivers, the outcome is dependent on contrasting color between skin, clothing, background etc.

6.3.2 Background Distortion for Privacy Protection

When facial landmark tracking becomes unreliable, precautions must be taken to protect the driver’s identity. One way is to replace everything around the region of interest with black pixels. Other possibilities include replacing the background using one of the methods mentioned below to give some spatial context, as illustrated in Figure 6.4.

Anisotropic Diffusion

Anisotropic Diffusion reduces noise or high frequency detail in an image, and preserves the overall model or low frequency details such as edges and lines. In fact, one of the criteria for anisotropic diffusion is “intraregion smoothing should occur preferentially over interregion smoothing” [82]. Figure 6.4 a shows three different numbers of iterations for two images of different drivers. First row of images, with an iteration of 160, are clearly identifiable. The second row of images, with an iteration 1000, has sufficient contextual information but de-identification is still not sufficient because enough distinguishing features exist in order to correctly identify the driver from a list of possible candidates. The last row of images, with an iteration of 10 000, takes significant computational time yet provides no contextual information.

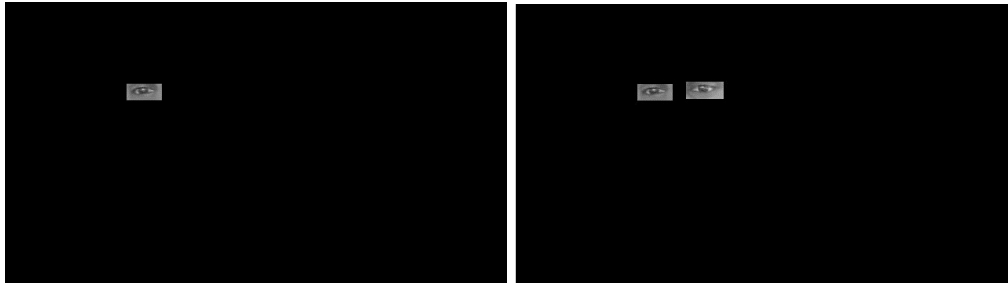


Figure 6.5: Illustration of a de-identification method where region around the eyes are preserved and the background is replaced with black pixels. While pixel replacement for the background ensures more privacy, it removes context information often helpful in determining driver’s gaze.

Transform Domain Scrambling

Transform Domain Scrambling is random permutation of AC-coefficients in the transform domain [29]. Given an image, transform domain coefficients, $q[i]$ with $i = 0, \dots, (n^2 - 1)$, are computed for each non-overlapping blocks of size $n \times n$ in the image. Within each block, randomly permuting only the AC-coefficients and applying inverse transform results in the scrambled image. Block size plays a key role in the level of de-identification. If the block size is too small, local scrambling will preserve most of the identifiable facial features. On the other hand, if the block size is the size of the image, then the background is like random noise. Figure 6.4b illustrates the increased loss of key facial features as well as spatial context with increasing block size.

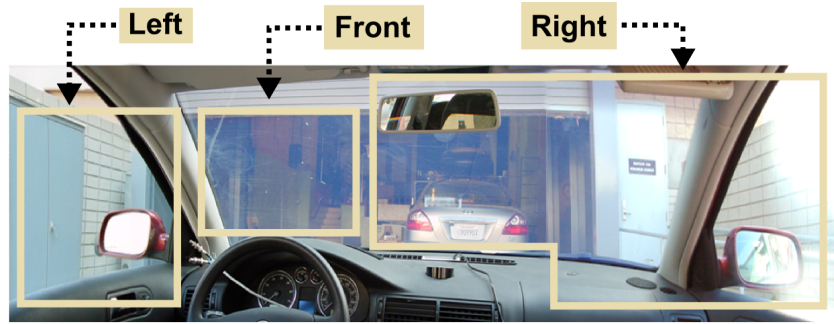
Visual querying raises concern that any form of de-identification on the background, except for pixel replacement, provides some information towards driver’s identity. For this reason, we use the pixel replacement approach, as illustrated in Figure 6.5, for further analysis. However, removing background information can remove contextual information often helpful in e.g. determining driver gaze (look) zone. Hence, a thorough experimental analysis is conducted to address the two aspects.

6.4 Experiment Design, Evaluation and Discussion

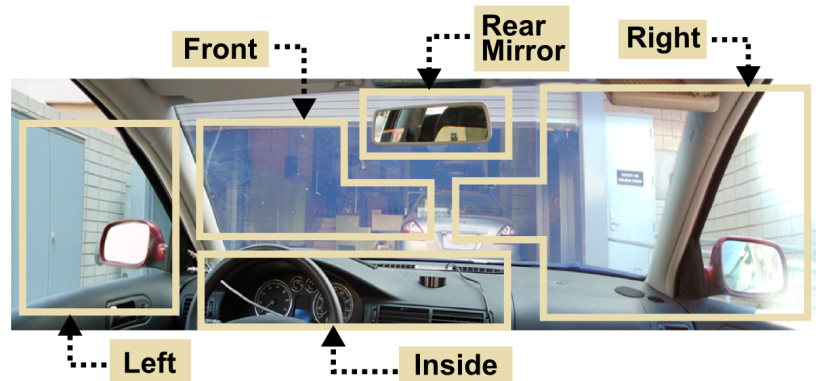
Looking inside the vehicle cockpit at the driver, we are more interested in driver behavior than driver’s identity. In the following sections, we describe the collected dataset and the user study with human participants on face recognition and on gaze zone estimation of de-identified images of drivers.

6.4.1 Experiment Design

The dataset is collected on the LISA-A testbed [98] with four drivers. The dataset contains video feed from a camera mounted to the left of the driver on the front wind shield near



(a) Three Gaze Zones



(a) Five Gaze Zones

Figure 6.6: Illustrates (a) the three gaze zone regions of interest: Left, Front, Right and (b) the five gaze zone regions of interest: Left, Front, Right, Rear Mirror, and Inside.

the A-pillar. All drives are from urban and freeway settings around the University of California, San Diego (UCSD) campus. While each drive lasted approximately 20 minutes, we choose sample images from discrete gaze zones to conduct our recognition and gaze zone estimation study. From a human factors perspective, we are interested in knowing whether the driver is looking inside or outside, and if outside, which regions. Hence, we designed the experiments with the following sets of discrete gaze zones.

One set of gaze zones, as shown in Figure 6.6a, consists of three regions: Left, Front, Right. The second set of gaze zones, as shown in Figure 6.6b, consists of five regions: Left, Front, Right, Rear-view mirror, Inside. The former set of gaze zones is self explanatory and except for ambiguity in the borders between regions, the regions are mutually exclusive. In the latter set, however, some of regions are not mutually exclusive. For example, a driver glancing at the rear-view mirror can also be considered as looking right. The inside gaze zone, representing the gauge and center console regions, can be confused with Front and Right gaze zones, respectively, for similar reasons.

Two user studies are conducted to test the level of recognition and the ability to estimate gaze zone in de-identified images. A total of 20 human subjects were asked to participate in this two part study. In the user study for recognition of faces in de-identified images, we use five

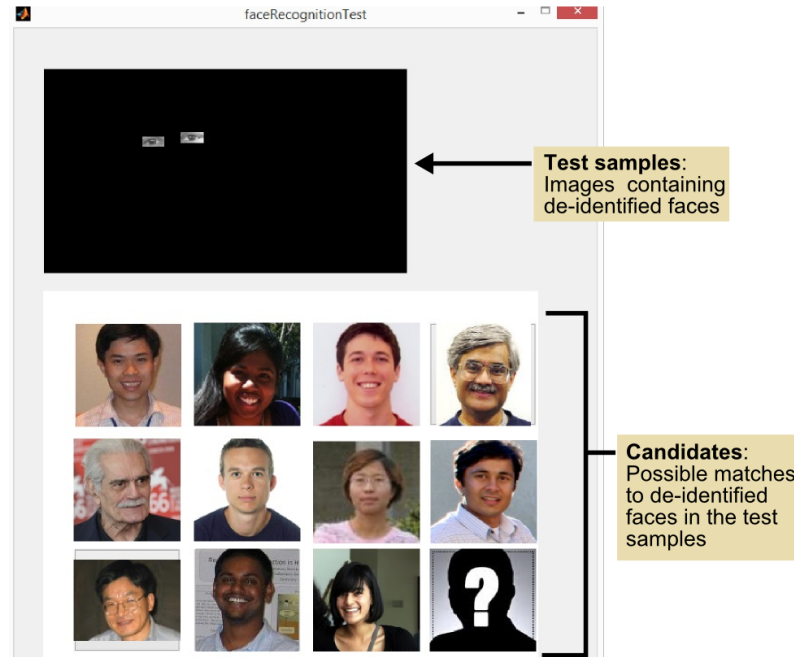


Figure 6.7: Layout of the face recognition testing toolbox for user study. Given a de-identified image, participants choose one of the 12 candidates that best represents the driver in the de-identified image.

images of each driver for the five gaze zones. In addition, approximately 5 random images from four other drivers are used to increase variability in the dataset. So there are a total of $(5 \times 4 + 19) \times 2 = 78$ de-identified images, representing de-identification with one-eye and two-eyes, presented to each of the ten participants. Figure 6.7 shows the layout of the user study as seen by participants.

In this layout for face recognition, the location indicated by the arrow for test samples is updated with de-identified images during the study and the 12 images of candidates are possible matches to the de-identified faces in the test samples. Given a de-identified image, participants choose one of the persons in the option who best identifies with the person in the de-identified image. Among the 12 options, the image with a question mark is an option for if the participant cannot conjecture who is in the de-identified image. The participants are instructed that people in the de-identified images may not necessarily be available as one of the options and not all people in the options are necessarily represented in the de-identified images. This represents a realistic situation of person identification where the subject in real life encounters many unknown faces to match with known (available) faces.

The second user study is comprised of nine expert participants estimating the gaze of the driver in the de-identified image by choosing one of the six categories: Left, Front, Right, Rear Mirror, Inside, and Unknown. These experts represent researchers who are interested in

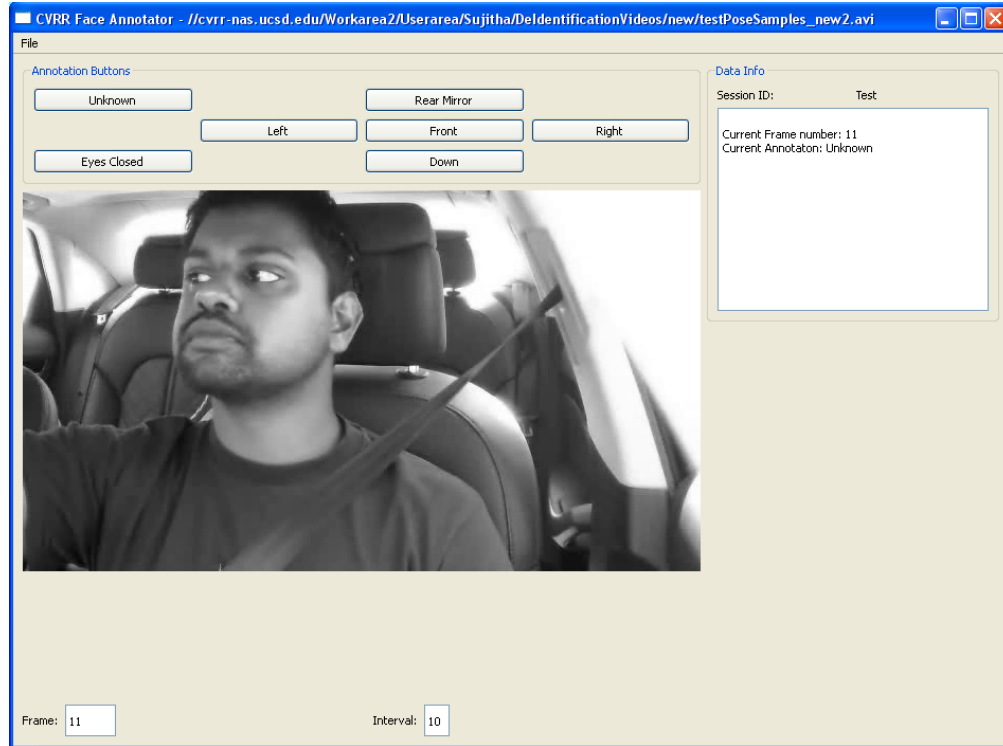


Figure 6.8: Layout of gaze zone estimation tool box for user study. Given a de-identified image, participants choose one of following categories that best represents the driver’s gaze: Left, Front, Right, Rear Mirror, Down, Unknown.

driver behavior study and are familiar with the camera perspective. By watching unperturbed videos of drivers not in this study but from the same camera perspective, the expert participants are familiarized with estimating gaze zones. This is especially challenging because the camera position is biased to the left. In the study, we use five images of two drivers for each of the five gaze zones. There are a total of $5 \times 2 \times 5 \times 2 = 100$ de-identified images, representing de-identification with one-eye and two-eyes, presented to each participant. Figure 6.8 shows the layout of the user study as seen by participants.

6.4.2 Face Recognition

Evaluating the degree of de-identification is a key part of designing the filter. Because failure to provide adequate protection of privacy is unacceptable and in one case, a failed de-identification attempt resulted in a lawsuit [113]. Two types of de-identification methods are implemented: the first preserves the region around one eye with black pixel replacement for background and the second preserves the region around both eyes with black pixel replacement for the background. These de-identification methods, henceforth, will be referred to as *One-Eye* and *Two Eyes*, respectively. Evaluations of face recognition on de-identified images can occur in

Table 6.2: Face Recognition User Study: Evaluation of Participants' Response

De-Identification Method	Drivers	Samples	Recognition Rate	Unknown Rate
One-Eye	1	50	0%	34%
	2	50	8%	38%
	3	50	2%	38%
	4	50	10%	46%
	All	200	5%	39%
Two-Eyes	1	50	0%	46%
	2	50	8%	40%
	3	50	8%	30%
	4	50	16%	54%
	All	200	8%	43%

one of two ways: human user study and machine vision.

First, an automatic machine vision based approach for face recognition is implemented to get an objective measure on the recognition rate. Training samples are taken from the list of candidate images, as presented in the user study, and testing samples are the de-identified images of the four drivers described earlier. As features, eye regions are extracted, scaled, aligned and intensity normalized. The nearest neighbor algorithm is then used to find the candidate image closest in the feature space to the de-identified image. Using 800 testing samples (200 samples per driver) randomly chosen from video sequences of the four drivers, the results show an 8% recognition rate, which is less than random chance. Given there are 11 possible candidates (excluding the *Unknown* available to participants in the user study), the random chance of recognition is $1/11 = 9.1\%$. While machine face recognition is more objective, it does not compare to a human's ability to recognize faces.

The second type of face recognition evaluation is with a user study using human subjects. In the user study, the first step of evaluation lies in justifying the pictures used in options to recognize drivers in the de-identified images. To do this, one participant performed the user study for recognition on images before they are de-identified. With a 100% recognition rate, results show the pictures used in options satisfactorily represent the raw images of looking at the driver. Second part of the face recognition user study is evaluating the level of recognition after de-identification. Ten participants took part in this second part of the evaluation.

In the user study, we present de-identified images starting with the de-identification method that reveals the least amount of information to the most about the driver's identity. Therefore, *One-Eye* de-identification is presented first followed by the *Two-Eyes*. Sequence of the images are randomly presented in each of the two cases. Table 6.2 details the number of drivers, the number of samples accumulated over all participants per driver, the recognition rate and the percentage of times participants responded with *Unknown* for de-identification with one-eye and two-eyes. Given there are 12 possible candidates to choose from, the random chance

of recognition is $1/12 = 8.3\%$. Table 6.2 shows the mean recognition rate is less than or equal to chance for de-identified images with *One-Eye* and with *Two-Eyes* for most of the drivers considered. On average, the recognition rate is higher with *Two-Eyes* than with *One-Eye*, as expected, however, both are below chance level. Notice that, participants responded with a high percentage of *Unknown* with both *One-Eye* and *Two-Eyes*, indicating the difficulty in recognizing a person with eyes only.

It's important to mention that the nature of the experiment, where the choices are given to pick one from, is very conservative and subjects could use elimination tactics without actually identifying the driver. For example, one participant noted that eye color (e.g. dark or light) was one of his criteria for choosing a candidate. Despite this, recognition rate of any particular driver as well as overall is very close to chance level.

6.4.3 Gaze Zone Estimation

We consider the gaze zones as illustrated in Figure 6.6b: Left, Front, Right, Rear Mirror and Inside. In a user study, a total of nine expert participants classify the driver's gaze using de-identified images of looking at the driver. The de-identified image samples in the user study were chosen after two experts used respective raw images, as well as image sequences around the interested frame for temporal context, to independently agree upon the correct label of gaze. Similar to the user study for recognition, de-identified images with *One-Eye* is presented first, followed by de-identified images with *Two-Eyes*. The participant pool between the user study for face recognition and gaze estimation, however, is non overlapping. In Table 6.3, the number of samples over all participants, the classification accuracy, and the percentage of times participants responded with *Unknown* for each method of de-identification and for each gaze zone is described. Notice that, in comparison to the percentage of *Unknown* responses for face recognition, the percentage of *Unknown* responses for gaze zone estimation is very small. This strengthens our goal of de-identifying drivers yet preserving driver's gaze.

Furthermore, gaze zone classification accuracy on de-identified images with *One-Eye* and *Two-Eyes* is 65% and 71%, respectively, which is well above chance of 20% for the five gaze zones considered. Confusion matrix, as given in Figure 6.9, gives insight into the misclassification of gaze zones. As expected, there is misclassifications among neighboring gaze zones. Left gaze, for instance was confused only with the Front gaze while Front gaze was confused with all but Rear Mirror gaze. On the other hand, Rear Mirror gaze was confused with Front gaze sometimes because some rear-view mirror glances were very subtle. Furthermore, Rear Mirror gaze was significantly confused with Right gaze. Even though participants were asked to chose the zone that better matches the driver's gaze than all others, it is not incorrect to assume a rightward gaze when a driver is gazing at the rear-view mirror. Similarly, Inside gaze is significantly confused with Front gaze and Right gaze. It is expected since gazing at the gauge and instrument panel invoke

Table 6.3: Gaze Zone Estimation User Study: Evaluation of Participants' response

De-Identification Method	Gaze-zone	Samples	Accuracy	Unknown Rate
One-Eye	Left	90	67%	1%
	Front	90	82%	0%
	Right	90	76%	9%
	Rear Mirror	90	67%	0%
	Inside	90	32%	7%
	All	450	65%	3%
Two-Eyes	Left	90	92%	2%
	Front	90	87%	2%
	Right	90	77%	1%
	Rear Mirror	90	61%	1%
	Inside	90	39%	3%
	All	450	71%	2%



Figure 6.9: Confusion matrix for the five gaze zone classification by participants of de-identified images with (a) one eye and (b) two eyes. Gaze zones are depicted in Figure 6.6. Each row represents true gaze and each column represents the participant's estimate of the gaze zone. Most of the elements in the diagonal have higher percentages than off-diagonal entries. On the average, gaze zone estimation is accurate 65% and 71% for de-identification with one-eye and with two-eyes, respectively.

gazes similar to Front and Right, respectively. In Figure 6.10, an extension of the confusion matrix is presented where each row represents true gaze and each image is one of the most recurring de-identified image in respective positions in the confusion matrix for de-identifications with *One-Eye* and *Two-Eyes*.

Interestingly, there are instances where de-identified image with *One-Eye* is sufficient to estimate the driver's gaze. One such instance is portrayed in Figure 6.11a, where the driver's gaze is leftward. In Figure 6.11, each of the four collages is made up of three images: bottom is the raw image, top left and top right are cropped images of de-identification with *One-Eye* and *Two-Eyes*, respectively. On the other hand, sometimes even the raw image is insufficient for gaze estimation because of the lack of temporal context. For example, in Figure 6.11b, it is hard to strongly support Front or Inside gaze without more reference into what happened before and possibly after this instant. A more expected result is when more visual cues (e.g.

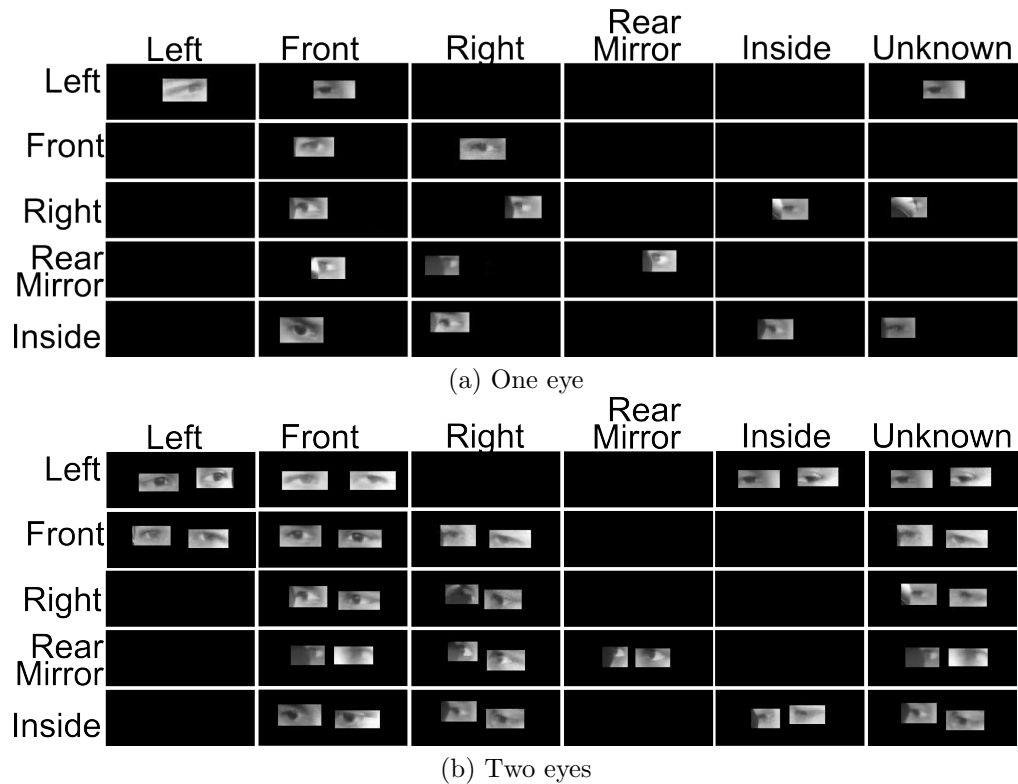


Figure 6.10: Five gaze zone performance of de-identified images with (a) one eye and (b) two eyes. Gaze zones are depicted in Figure 6.6. Each row represents true gaze and each column represents the participant’s estimate of the gaze zone. Each element in this matrix of images is a cropped image of one of the highest participant response in respective categories.



Figure 6.11: Illustrates multiple instances where typical confusion between gaze zones could occur. Each collage of images is comprised of three images: bottom is the raw image, top left and top right are cropped images of de-identification with one-eye and two-eyes, respectively. Each collage illustrates when (a) de-identification with one-eye is sufficient, (b) raw image doesn’t contain sufficient information, (c) de-identification with two-eyes provides sufficient context and (d) de-identification with two-eyes introduces more uncertainty than with one-eye.

two eyes over one eye) lead to correct gaze estimation. Figure 6.11c is such an instance where participants found difficulty in inferring gaze with one-eye but had no problem estimating gaze with two-eyes. Surprisingly, however, there are instances where participants misclassified more often with two-eyes than with one-eye. One of these instances is portrayed in Figure 6.11d,

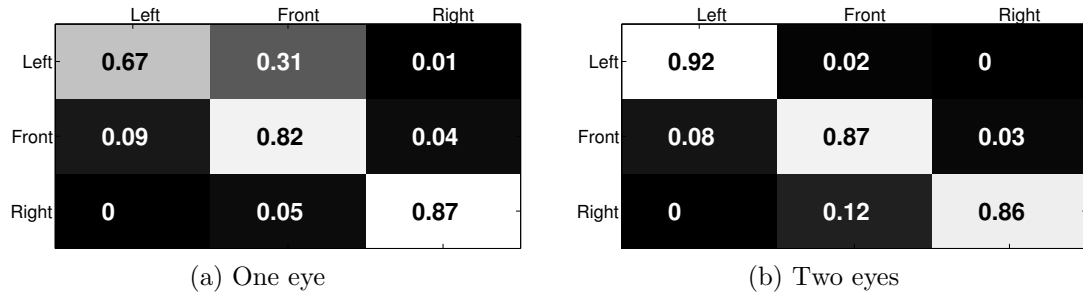


Figure 6.12: Three gaze zone performance of de-identified images with (a) one eye and (b) two eyes. Accuracy for Right gaze zone goes up significantly when Rear-Mirror gaze zone is considered to be part of Right gaze zone. Average accuracy for (a) one eye is 0.79 and (b) two eyes is 0.88.

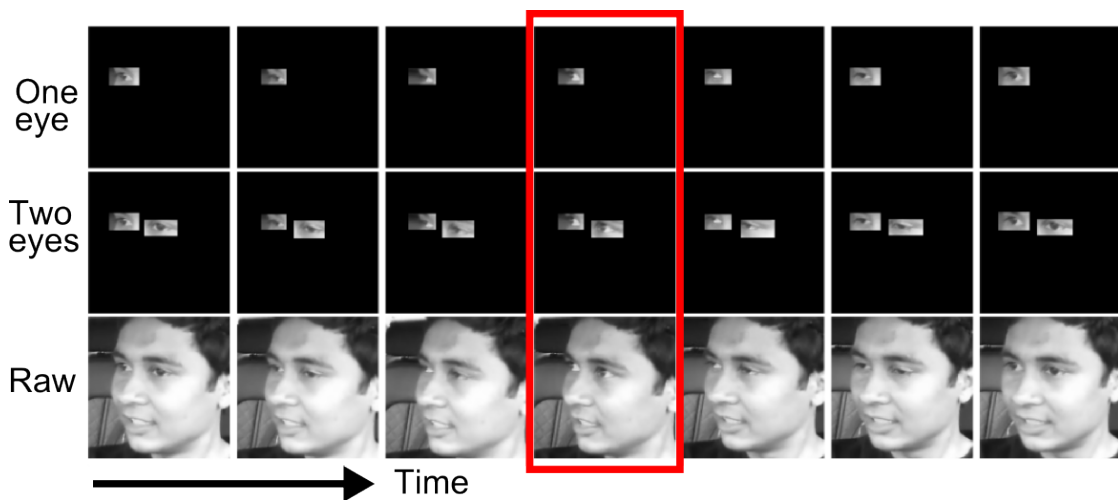


Figure 6.13: A sequence of de-identified images with respective raw images from a video sequence of the driver glancing at the rear view mirror. The red box indicates an instance when the driver's gaze is on the rear-view mirror. Gaze estimation of the de-identified images in this instance can be more accurate when provided the sequence leading up to it and possibly following it.

where the ground truth gaze is Rear Mirror. One possible reason is, more information on the gaze introduces alternative possibilities.

Finally, due to the inherent confusions in some zones we consider a simpler allocation of gaze zones as illustrated in Figure 6.6a: Left, Front and Right. One of the main differences is the absorption of Rear-Mirror into Right gaze zone. The second is the removal of Inside gaze zone from the list, because looking inside at the gauge is similar to looking front whereas looking at the center console is similar to looking right. Figure 6.12 shows the confusion matrix for a three gaze zone classification. The overall accuracy is higher for both forms of de-identification since there is less ambiguity between regions. As expected, the average accuracy of 88% for de-identified images with two-eyes is higher than 79% for de-identification with one-eye when considering a

simple allocation of gaze zones.

Estimating the gaze of the driver, in this user study, proved especially difficult because the participants viewed an image out of context. Figure 6.13 shows a sequences of de-identified images with respective raw images from a video sequence of the driver glancing at the rear-view mirror. The red box indicates the actual act of looking at the rear-view mirror, which is typically the image presented to a participant in our study. By providing the sequence leading up to the event and possibly the sequence following, the participants are provided with more context and can thus make a more informed decision.

6.5 Concluding Remarks

In the design of driver assistance system, when looking at the driver, driver’s identity is irrelevant to understanding and predicting driver behavior. We explored a de-identification scheme which preserves the facial region around the eyes in the foreground and obscures everything else in the background. Eyes especially because it can provide finer detail on gaze zone estimation. A user study using human participants showed face recognition to be well below chance and gaze estimation accuracy for the five gaze zones to be 65% and 71% with *One-Eye* and *Two-Eyes*, respectively. Gaze zones were misclassified mostly due to lack of spatial and temporal context.

Our ongoing research is focused on preserving different combinations of facial regions, on testing different background obscuring algorithms for spatial context and on providing a sequence of images representing the act of looking at a gaze zone for temporal context [94]. The latter is of special interest because knowing the history of driver’s gaze is useful in robustly predicting the current gaze of the driver [100].

6.6 Acknowledgments

Chapter 6 is in full a reprint of material that is published in the IEEE Transactions on Intelligent Transportation Systems (2014), by Sujitha Martin, Ashish Tawari and Mohan M. Trivedi. The dissertation author was the primary investigator and author of this paper.

Chapter 7

Conclusions

In this dissertation, we have tackled research challenges related to using cameras, fusing modalities and preserving privacy to equip intelligent vehicles with an understanding of the driver’s state. In relation to this central research goal, we have made several critical research contributions

Gaze zone estimation is still a new area of research in literature. While there are many ways to design a gaze zone estimator, the focus of this research is in carefully designing submodules which will build up to achieve a continuous and robust gaze zone estimation. Key modules in this system include, face detection using deep convolutional neural networks, landmark estimation from cascaded regression models, head pose from geometrical correspondence mapping from 2-D points in the image plane to 3-D points in the head model, horizontal gaze surrogate based on geometrical formulation of the eye ball and iris position, vertical gaze surrogate based on openness of the upper eye lids and appearance descriptor, and finally, a 9-class gaze zone estimation from naturalistic driving data driven random forest algorithm. In addition to the contributions in each of these submodules, this work also gives an extensive analysis of the system as a whole on naturalistic driving data. The gaze estimator described in this work is based on static features, meaning no information in time window previous to the instance is leveraged. Near future work will extend this framework to dynamics features and therefore show significant improvement in current framework’s sensitivity to occlusions and global models.

In this study, we explored modeling driver’s gaze behavior in order to predict maneuvers performed by drivers, namely left lane change, right lane change and lane keep. The particular model developed in this study features three major aspects: one is the spatio-temporal features to represent the gaze dynamics, second is in defining the model as the average of the observed instances and interpreting why such a model fits the data of interest, third is in the design of the metric for estimating fitness of model. Applying this framework in a sequential series of time windows around lane change maneuvers, the gaze models were able to predict left and right

lane change maneuver with an accuracy above 80% around 1.6 seconds before the maneuver and reaches a peak of 90% recall rate around 600 milliseconds. The overall framework is designed to model driver’s gaze behavior for any tasks or maneuvers performed by driver. In particular, the spatio-temporal feature descriptor composed of glance durations and glance transition frequencies are powerful tools to capture the essence of recurring driver gaze dynamics. To this end, there are multiple future directions in sight. One is to quantitatively define the relationship between the time window from which to extract those meaningful spatio-temporal features and the task or maneuvers performed by driver. Another is in exploring and comparing different modeling approaches, including HMM, SVM and bag of words approaches. Other future directions include exploring unsupervised clustering of gaze behaviors and exploring the effects of quantity and quality of events (e.g. same vs. different drives, different drives from same or different time of day) on gaze behavior modeling.

In this study, we also provided a data-driven approach to understand and analyze the temporal interplay between three modalities: head, hands and eyes of the driver, in the context of stop-controlled intersections. A naturalistic driving dataset was employed to show that preparatory motions range in the order of a few seconds to a few milliseconds, depending on the modality, before maneuvers occurred at intersections. Features generated from the head are seen to be most useful in terms of predictive power. Eye based features play an important role in prediction at a very early stage of the maneuver, while hand features dominate towards the end. These findings are in line with the general flow of most human activities- first see, then perceive and finally actuate. Future work in this scope encompasses extending this understanding of the temporal influence of each modality to different maneuvers performed under different contexts. Additionally, this study will serve as a basis to come up with a strong predictive algorithm for intersections. Other information sources like vehicle dynamics and external camera sensors are to be integrated and explored.

The Laboratory for Intelligent and Safe Automobiles (LISA) team at UCSD has made a major commitment to offer its data sets to the worldwide research community so that anyone interested in the research can get open access and evaluate their own algorithms and benchmark them against others. The main motivators for such effort is to engage wider community of students, scholars, developers in this exciting and challenging field and also to provide credible, quantifiable benchmarks for evaluation and rapid progress in the field. In this regard, this work presented important issues, challenges, and metrics associated with development of robust vision based systems for intelligent vehicles. Specifically, this dissertation presented issues related with face detection and head pose in static images. A major step in the future direction for face related challenges alone is in the temporal domain; this includes face tracking, head pose tracking, gaze estimation, activity analysis and expression recognition. A parallel effort is also in the integrative vision framework of looking-in (e.g. faces, hands) and looking-out (e.g. vehicles, traffic signs) for activity recognition and intent prediction, to name a few.

In the design of driver assistance system, when looking at the driver, driver's identity is irrelevant to understanding and predicting driver behavior. We explored a de-identification scheme which preserves the facial region around the eyes in the foreground and obscures everything else in the background. Eyes especially because it can provide finer detail on gaze zone estimation. A user study using human participants showed face recognition to be well below chance and gaze estimation accuracy for the five gaze zones to be 65% and 71% with *One-Eye* and *Two-Eyes*, respectively. Gaze zones were misclassified mostly due to lack of spatial and temporal context. Our ongoing research is focused on preserving different combinations of facial regions, on testing different background obscuring algorithms for spatial context and on providing a sequence of images representing the act of looking at a gaze zone for temporal context [94]. The latter is of special interest because knowing the history of driver's gaze is useful in robustly predicting the current gaze of the driver [100].

The work in this dissertation has been oriented towards enabling intelligent vehicles to understand and estimate the state of the driver by mainly observing the face and some works on fusion with hand analysis. Study of driver state in naturalistic driving studies is still in its infancy relative to surround analysis. One of the future research direction is in understanding, modeling and predicting driver behavior in highly automated vehicles. Reason being the role of drivers in highly automated vehicles is constantly changing and being redefined and studies on driver state is mainly in driving simulators. Another future research direction is in understanding and modeling the interaction of humans in the vehicles with those surrounding the vehicle, because the aim is global safety. Research challenges will lie ahead, in correlating the observable actions to thoughts and intentions of humans in and around the intelligent vehicle.

Appendix A

Continuous Head Movement Estimator: Framework and On-Road Evaluations

A.1 Introduction

Driver head and eye dynamic behaviors are of particular interest, as they have the potential to derive where or at what the driver is looking. Traditionally, eye gaze and movement are considered good measures to identify an individual's focus of attention. Vision based systems are commonly used for gaze tracking as they provide non-contact and non-invasive solution. However, such systems are highly susceptible to illumination changes, particularly in real-world driving scenario. Eye-gaze tracking methods using corneal reflection with infrared illumination have been primarily used indoor [47] but are vulnerable to sunlight. Robustness requirement of IDAS has suggested the use of head dynamics. While precise gaze direction provides useful information, head pose and dynamics provide coarse gaze direction, which are often sufficient in a number of applications [57, 28]. Recent studies have used head motion along with lane position and vehicle dynamics to predict a driver's intent to turn [13] and intent to change lanes [67]. In fact, head motion cues when compared to eye gaze cues were shown to better predict lane change intent, 3s ahead of the intended event [25]. A significant amount of research has gone towards fatigue and attention monitoring using driver head dynamics [5, 8]. In a more recent study, head dynamics was used to estimate driver's awareness of traffic objects by learning which objects attract the driver's gaze depending on the situation [6].

Many state-of-the-art vision based head pose algorithms have taken the necessary steps to be automatic, fast and person invariant [68]. These systems have shown good performance

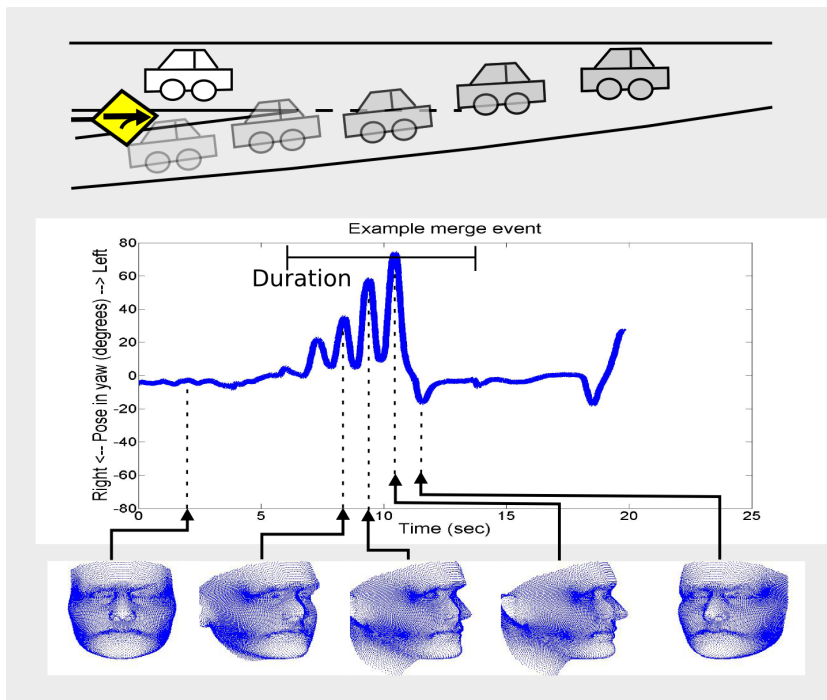


Figure A.1: Head movements during a merge event. The 3D model of a head illustrates observed facial feature from a fixed camera perspective and self-occlusion induced by large head movements.

when the head pose is near frontal. Martin et al. show that during a typical ride a driver spends 95% of the time facing forward. Then, a system may be able to perform reliably 95% of the time but it is during those 5% non-frontal glances which are of special interest since interesting events, critical to driver safety, occur during those times. Figure A.1 illustrates a typical temporal dynamics of head pose seen from a fixed single camera perspective during a merge maneuver. It can be seen that head pose quickly goes far from forward facing (about 0^0 in yaw angle). It is during those times when performance of monocular based systems degrades significantly due to decreased visibility of facial features and texture caused by self occlusion.

Hence, we require a system with new sensing approaches to continuously estimate driver’s head movement. A natural choice for the design of such a system is the use of multiple cameras [66, 98]. Multi-camera systems exist in many other applications such as gesture recognition [51, 103], human body pose and activity recognition [50], face detection, tracking and pose estimation in intelligent space etc. A thorough study of such systems in a vehicular setting utilizing naturalistic driving data, however, is lacking in the literature. Towards this end, we propose a continuous head movement estimator (CoHMEt), a key component for the uninterrupted driver monitoring system.

Our contributions are three folds. First, we propose a distributed camera solution and conduct a thorough study comparing different configurations of multiple cameras. Second, we

propose two solutions for head pose estimation based on a geometric method utilizing state-of-the-art techniques for facial feature tracking. We introduce spatio-temporal constraints available in driving context to improve head pose tracking accuracy as well as computation time. Furthermore, we compare the two solutions for different configurations and show that the choice of the algorithm determines ‘best’ camera configuration. Finally, we quantitatively demonstrate the success of this system on the road. For this, we gather a dataset which targets spatially large head turns (away from the frontal pose) during different vehicle maneuvers. Although this makes the dataset challenging, it sets realistic requirements for the vision based system to be a viable commercial solution. We evaluate our proposed systems using both metrics, error in angular calculation in 3 degree of freedom (pitch, yaw and roll) and failure rate, which is a percentage of the time the system’s output is unreliable. The part hardware and part software solution of multiple camera perspectives will be shown to improve continuous head dynamics estimation during critical events such as merges, lane changes and turns.

A.2 Related Research

Naturalistic driving presents unique challenges for vision based head dynamic estimation and tracking methods. Amongst research thrust and commercial offerings that can provide automatic head pose estimation, most of them lack rigorous and quantitative evaluation in an automobile. In a car, ever-shifting lighting conditions cause heavy shadows and illumination changes, and as a result, techniques that demonstrate high proficiency in stable lighting often will not work in on-road driving situations. In this work, our effort is to advance state-of-the-art technology for head pose and dynamics estimation targeted for drive assistance systems. In this context, we review past work with a focus on systems that have been evaluated in naturalistic driving or studies conducted in-lab/driving-simulator setup that have potential for applying but have yet to be tested under naturalistic driving conditions. For a good overview of head pose estimation in computer vision, readers are encouraged to refer to a survey by Murphy-Chutorian and Trivedi [68].

Head pose estimation algorithms can generally be classified into the following main categories: geometric/shape feature based, appearance/texture feature based, and hybrid (shape+texture) feature based methods. Methods based on shape features analyze geometric configuration of facial features along with face model (e.g. cylindrical [109], ellipsoidal [57] or mean 3D face [65]) to recover head pose. Smith et al. analyzed color and intensity statistics to find both eyes, lip corners, and the bounding box of the face [19]. By using these facial features, they estimated continuous head orientation and gaze direction. However, this method cannot always find facial features when the driver wears eyeglasses or makes conversation. Kaminski et al. analyzed the intensity, shape, and size properties to detect the pupils, nose bottom, and pupil glints to estimate continuous head orientations and gaze direction [20]. By using the foregoing

Table A.1: Selected studies on vision based head pose and dynamics estimation systems which are already tested or have potential to work in automobile environment.

Research study	Objective	Methodology				Evaluations		
		Feature	Perspective	Resolution	DOF	Operation	Dataset	Metrics
Zhu and Ji '04 [123]	Head pose estimation from uncalibrated monocular camera	Texture, IR	Single	Continuous	yaw, pitch, roll, and x-, y- translations	10fps	In lab	Unspecified
Guo et al. '06 [48]	head pose estimation for driver surveillance	Texture	Single	225 discrete poses	yaw, pitch, roll	Real-time	In-lab	CCR
Wu and Trivedi '08 [116]	Two-stage, coarse and fine, head pose estimation	hybrid	Single	86 discrete poses	yaw, pitch	Not real-time	In lab	Each discrete pose: classification accuracy
Murphy-Chutorian et Trivedi '10 [71]	Head pose estimation	Texture	Single	Continuous	yaw, pitch, roll, and x-, y-, z- translation	Real-time	Naturalistic driving	MAE and STD. Function of true angle: ME and STD
Martin et al. '12 [65]	Head pose estimation	Shape	Single	Continuous	yaw, pitch, roll	Real-time	Naturalistic driving	STD and FR.
Fu et al. '13 [42]	Automatic estimation of driver's gaze zone	Texture	Single	Continuous	yaw, pitch	Real-time	ambiguous	MAE
Bär et al. '13 [6]	Driver's head pose and gaze estimation	Texture & depth	Single	Discrete	yaw	Real-time	On-Road	CCR
Proposed CoHMET	Continuous head dynamics estimation for spatially large head movements	Shape	Multi	Continuous	yaw, pitch, roll	Real-time	Naturalistic driving	Overall: MSE, STD. Function of yaw: 1st three error quartiles.

methods considering both eye and head orientations, detailed and local gaze direction can be estimated. However, the accuracy of the eye location significantly drops in the presence of large head movements. Thus, the algorithm delivers poorer performance for deviation from the frontal pose.

To circumvent precise localization of detailed facial feature, Ohue et al. proposed simple facial features - the left and right borders, and the center of the face [76]. Along with these features, the authors used a cylindrical face model to find the driver's yaw direction. Lee et al. [57] used similar shape feature with ellipsoidal face model to improve the yaw estimate when the head rotates significantly away from frontal pose. The authors trained gaze classifiers in a supervised framework to determine 18 gaze zones. Fu et al. designed a system that categorizes head pose into 12 different gaze zones based on facial features [42]. The system automatically learns the zones based on different calibration points such as side mirrors, rear-view mirrors etc. It takes, however, several hours of driving before automatic calibration reaches similar accuracy as of supervised training based method. It is unclear whether the evaluations are performed in stationary or moving vehicle and whether drivers were asked to look towards defined zones during data collection. A study conducted on naturalistic driving data by Martin et al. [65], tracked prominent facial features (e.g. eye corners, nose corners and nose tip) and analyzed their geometric configurations to estimate head pose. This is very similar to proposed approach but it is limited to single perspective. Appearance-based approaches attempt to use holistic facial appearance, where face is treated as a two-dimensional pattern of intensity variations. They assume that there exists a mapping relationship between 3D face pose and certain properties of the facial image, which is constructed based on a large number of training images. In [48], Guo et al. utilizes template face images distributed in the pose space to determine the head pose. The system operates by first finding the face using a cascade of face detectors and then a best face exemplar in the training dataset is found. The head pose of the exemplar is the estimated head pose. The study provides little information on testing methodology and how the ground truth is obtained. Such methods, however, require precise localization of faces as matching is often sensitive to localization errors. Bär et al. [7], estimate driver's head pose using RGB-D images. Multiple templates are used to align 3D point cloud data using Iterative Closest Point (ICP) algorithm to obtain head pose and subsequently drivers line of gaze by analyzing the angles of the eyes in RGB image. Template matching algorithm (ICP) also suffers from initialization error.

Zhu and Ji [123] proposed a system to track 2D face location and 3D face pose simultaneously. 3D face pose is tracked using Kalman Filtering which in turn guides 2D face localization. The system uses a planar face appearance template to match with current frame to obtain the best pose parameters. It, however, requires initialization with frontal face and tracking is performed from this initial position. Like other holistic approaches, the use of full face appearance template can be very limiting, specially in driving scenario due to constant varying illumination condition. The authors propose to dynamically update the face model or when the track is lost,

use eye detection and fiducial facial feature to estimate the rough pose parameters. The evaluation is performed in a lab setting. With the limited information about the characteristics of the database, it is not clear how the system performs as a function of out-of-plane rotation.

Wu et al. detected discrete yaw and pitch by using a coarse-to-fine strategy using quantized pose classifier [116]. First, a coarse pose estimate is obtained by nearest prototype matching with Euclidean distance in the subspace of Gabor wavelets. In the second stage, the pose estimate is refined by analyzing finer geometrical structure of facial features. This is a hybrid approach combining shape and texture features. Evaluations are performed in laboratory settings. More recent studies have taken their research to naturalistic driving, where driver is asked to drive on highways or urban roads as they would do in their normal commute. One notable work by Murphy-Chutorian et al. estimated the initial head orientation using a local-gradient-orientation (LGO) feature and support vector regression [69], called a static pose estimator. Finer head orientations were computed by fitting and tracking a 3D face model. While the tracking module showed good performance, the combined system suffered from inaccurate initialization.

A summary of select studies with emphasis on applicability to driver assistance system is provided in Table A.2. Apart from their original objective, the following important elements related to employed methodology and evaluation strategies are mentioned for comparison against the proposed CoHMEt framework:

- Objective: What is the purpose of the study (e.g. gaze estimation)?
- Methodology:
 1. Feature: Type of features used (shape, texture or hybrid).
 2. Perspective: Whether system utilizes single or multiple cameras.
 3. Resolution: Whether the system provides discrete or continuous head pose estimate.
 4. Degrees of freedom: Number of degrees of freedom in the system output e.g. rotation - pitch, yaw and roll, and position - x,y and z values from a reference frame.
- Evaluations:
 1. Operation: Real-time vs non real-time
 2. Dataset: In what environment, the evaluation is performed (naturalistic driving, stationary vehicle or lab)
 3. Metrics: Type of metrics used for the performance evaluation.

Our proposed approach falls in the category of shape feature based methods. Unlike appearance based methods, they are intuitive and simple to implement (since the cause of failure can be reasoned out well). The challenge, however, lies in robust and accurate localization of the facial features. With the recent advancements in facial feature tracking methods, we revisit them

and perform a thorough evaluation in naturalistic driving scenario. Furthermore, with multiple cameras, we improve the operational range while maintaining good accuracies. Unlike stereo camera (an instance of multi-camera) setup, we do not have any assumption of the visibility of the faces in both cameras nor do we require lengthy calibration process. In fact, our cameras have wide baseline and are uncalibrated. The proposed framework utilizes them independently in a parallel fashion and the results are further analyzed by later stages to provide the final output.

A.3 Issues and Challenges in Continuous and Robust Head Movement Analysis

Researchers working on driver monitoring systems, in particular, for driver head dynamic analysis, face unique challenges. As argued earlier, methods designed and tested in controlled lab settings provide no guarantee of robust performance in automobile environment. Hence, a proper evaluation on naturalistic driving database is very much required. The challenge lies in the design of a reliable, configurable and yet affordable database collection module. It's not just a matter of mounting cameras, but we also require ground truth for proper evaluation. Automobile setting during driving, however, precludes conventional methods used in lab environment e.g. asking individuals to look to certain fixed direction. Care needs to be taken to not distort imagery input e.g by placing marker on face. Manually labeling either direct head pose information or more objectively, annotation of facial features can provide head pose measurement. However, with video data at 30fps, it quickly becomes a daunting task and renders itself practically infeasible. One good candidate could be magnetic sensors. They are used extensively in lab settings without cluttering or obscuring visual data. However, they can be unreliable in an automobile due to their high susceptibility to noise and the presence of metal in the environment. Optical motion capture systems do provide a very reliable solution. But they are often very expensive with bulky equipments and require lengthy calibration. Inertial sensors utilizing accelerometer, gyroscopes or other motion sensing devices can provide compact, inexpensive and clean solution. But they often suffer from drift associated with gyroscope. However, this can be solved as proposed in this work by small amounts of manual annotation.

Another important aspect is number of camera(s) and their placement. Camera should neither block the driver's view for safe driving nor should its presence alter driver's behavior. At the same time, the placement should not be prone to frequent occlusions. A choice of placement can very much be application dependent, though a desirable choice would be one that covers as large a pose space, generally exhibited by a driver during a typical ride, as possible. From a computer vision perspective, however, intrinsic properties of head dynamics present a challenge to the robustness of many existing algorithms. As described earlier, many existing state of the art head pose estimation algorithms, explicitly or implicitly, rely on a portion of the face to

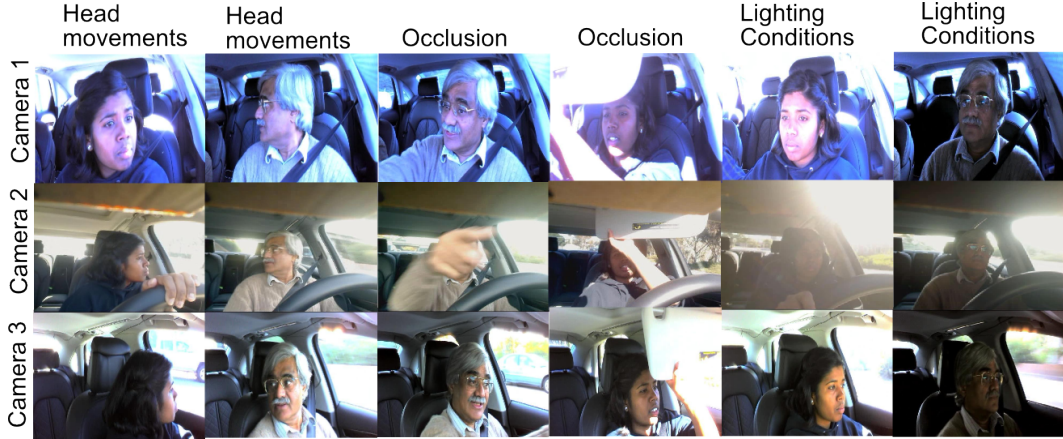


Figure A.2: Multi-perspective data collected during naturalistic on-road driving. Each row of images shows images are from a particular camera location and each column of images are time-synchronized. Locations of the camera: Camera 1 is near the left A-pillar, Camera 2 is close to the dashboard, and Camera 3 is near the rear-view mirror. Notice, challenges (e.g. external-/self-occlusion, shadows, illumination change) present in real world data.

be visible in the image plane to estimate head pose. This means that even during large head movements, algorithms require the visibility of facial features to continuously track the state of the head. With a single perspective of the driver’s head, however, large spatial head movements induce self-occlusions of facial features as illustrated in the first two columns of Figure A.2. In Figure A.2, each row of images are taken from a different camera perspective and each column of images are time-synchronized. Clearly, the availability of multiple perspectives decreases the severity of self-occlusions at any instant in time which translates to an increase in the robustness of continuous head tracking.

Occlusions of facial features can also occur due to external objects (e.g. hand movements near the face region, sunglasses). Depending on camera perspective, hand movements on steering wheel during vehicle maneuvers, to adjust sunshade, to point, etc. can cause occlusion. The middle two columns in Figure A.2 show examples of the latter two scenarios with hand movements. Effects of lighting conditions are also highly dependent on the camera location. In Figure A.2, the last two columns illustrate the effects of lighting conditions. Therefore, a multiple perspective approach with suitable camera placements, can mitigate the adverse effect of any one camera perspective being unreliable to track the head.

A.4 CoHMEt: Framework and Algorithms

Continuously and accurately monitoring driver’s head movement even during large deviation from the frontal pose requires improved operating range of the head pose tracking system. For this, we propose a distributed camera framework inside the vehicle cockpit. The framework

treats each camera perspective independently and a perspective selection procedure provides the final head pose estimation by analyzing temporal dynamics and the current quality of the estimated head pose in each perspective. For head pose estimation, we present a geometric method where local features, such as eye corners, nose corners and nose tip, and their relative 3D configurations determine the pose. In the following sections, we present automatic facial feature detection and tracking methods, pose estimation approach and perspective selection procedure in detail.

A.4.1 Facial Feature Detection and Tracking

In this work, facial features refer to salient landmarks on the face such as eye corners, nose corners, nose tip, mouth contour and outer face contour as shown in Figure A.5. We present two formulations for automatic facial feature detection and tracking based on two separate feature detection methods: Constrained Local Model (CLM) introduced by Cristinacce and Cootes [17, 19] and Pictorial Structure Matching (PSM) proposed by Felzenszwalb and Huttenlocher [33]. Unlike images, video data provides temporal constraints; moreover, a driving setting imposes spatial constraints on the detected facial features in the image plane. In our formulations, we introduce these spatial and temporal constraints to improve the tracking accuracy by reducing false detections, and the computation cost by reducing the search space.

Constrained Local Model

CLM represents objects, in our case faces, using local appearance descriptions centered around landmarks of interest, and a parameterized shape model of those landmarks. Local representation of appearance circumvents many drawbacks of holistic approach (e.g. Active Appearance Model (AAM)), such as modeling complexity and sensitivity to illumination changes, and shows superior generalization performance to novel unseen faces. The local descriptors are generally learned from labeled training images for each landmark. These local representations, however, are often ambiguous mainly due to small support region with large appearance variation in the training data. The effect of the ambiguity is typically reduced by the shape model that constraints the joint positioning of the landmarks.

A parametrized shape model to capture plausible deformations of landmark locations is given in Equation A.1. This is also known as a point distribution model (PDM), a term coined by Cootes and Taylor [16]:

$$\mathbf{p}_i = sR_{2D}(\bar{\mathbf{p}}_i + \Phi_i \mathbf{q}) + \mathbf{t} \quad (\text{A.1})$$

where $\mathbf{p}_i = (x_i, y_i)$ is the 2D location of the i^{th} landmark in the image \mathcal{I} . $\bar{\mathbf{p}}_i$ is the 2D location of the i^{th} landmark of the mean shape and Φ_i encodes the shape variations. Rigid parameters

$\boldsymbol{\theta}_{rg} = \{s, R_{2D}, t\}$ with global scaling s , in-plane rotation R_{2D} and translation t , along with non-rigid parameters \mathbf{q} , represent parameters of the PDM.

Let's define $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{rg}, \mathbf{q}\}$. The objective of CLM, then, can be defined in a probabilistic framework as maximizing the likelihood of the model parameters such that all of the facial landmarks are aligned to their corresponding locations. With the assumption of conditional independence amongst detections of each landmark, the objective function becomes:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}|\{l_i = 1\}_{i=1}^n, \mathcal{I}) &= p(\{l_i = 1\}_{i=1}^n|\boldsymbol{\theta}, \mathcal{I}) \\ &= \prod_{i=1}^n p(l_i = 1|\boldsymbol{\theta}, \mathcal{I}) \end{aligned} \quad (\text{A.2})$$

where $l_i \in \{+1, -1\}$ is a discrete random variable denoting the i^{th} landmark is aligned or not.

To facilitate the optimization process so that it is efficient and numerically stable, the true response map $p(l_i = 1|\boldsymbol{\theta}, \mathcal{I})$ of the local detectors are approximated by various models, such as parametric representation - Gaussian density with diagonal covariance [16], full covariance [115], Gaussian mixture model [46], or nonparametric representation - kernel density estimate (KDE) [89]. In our current implementation, we chose KDE for it's fast convergence property with good tracking ability [89]. It has shown its efficacy in other applications too, such as face expression recognition [99]. In the method, landmark locations are optimized via subspace constrained meanshifts while enforcing their joint motion via shape model. The maximum likelihood estimate (MLE) of the parameters, however, does not exploit the constraint setting present in the driving context. Since driver's seat location is fixed while driving, the body and head locations are restricted along with head orientation observed from the fixed camera perspective. To incorporate these constraints, we learn the parameter space, particularly for rigid parameters $\boldsymbol{\theta}_{rg}$ online. We need to learn this online since each driver has a different seat setting suitable for their driving.

To learn the probable face location and face size, face detection is used to find bounding boxes, B^i , for the first N_B face detected frames:

$$\mathbf{B}^i = \begin{bmatrix} x_{min}^i & y_{min}^i & x_{max}^i & y_{max}^i \end{bmatrix}$$

where $i \in 1, \dots, N_B$. A restricted face region \mathbf{B}^* is obtained as:

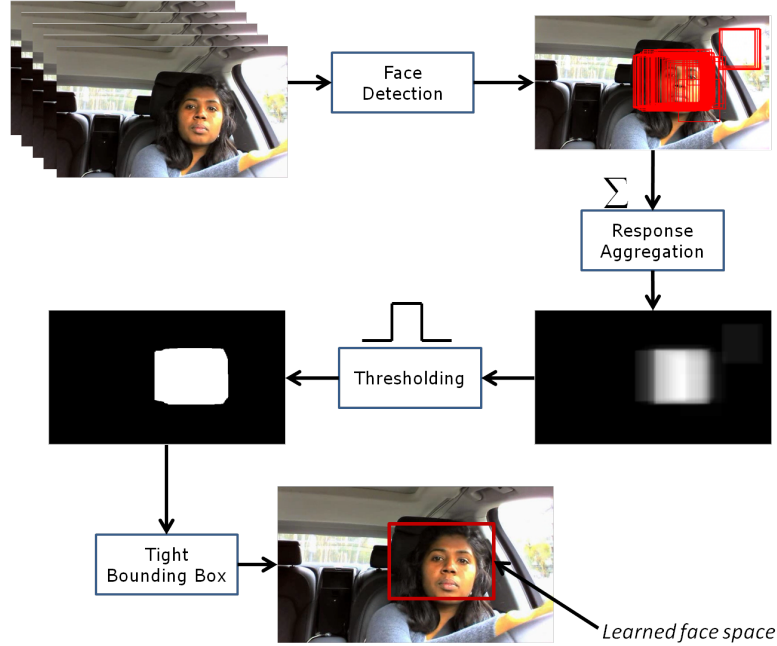


Figure A.3: Illustration of the online learning process of estimating restricted face region in the image plane.

$$H(x, y) = \sum_{\mathbf{B}^i} U[x - x_{min}^i]U[y - y_{min}^i] \quad (\text{A.3})$$

$$- U[x - x_{max}^i]U[y - y_{max}^i]$$

$$M(x, y) = U \left[\frac{H(x, y)}{\gamma} - \alpha \right] \quad (\text{A.4})$$

$$\mathbf{BR}(\mathbf{P}) = \left\{ \left(\min_{x \in \mathbf{P}} x, \min_{y \in \mathbf{P}} y, \max_{x \in \mathbf{P}} x, \max_{y \in \mathbf{P}} y \right) \right\} \quad (\text{A.5})$$

where, $U[\cdot]$ is the unit step function and the normalization factor $\gamma = \max_{x,y} H(x, y)$. $\alpha \in (0, 1)$ is a tuning parameter to control the size of the expected face region $M(x, y)$. A tight bounding rectangle $\mathbf{BR}(\cdot)$ as defined in Eq. A.5 is calculated using set of points $\mathbf{P} = \{\mathbf{p} = (x, y) \mid M(x, y) > 0\}$. We call this minimum bounding rectangle B^* , the restricted face region. Figure A.3 depicts the overall process. Estimated facial landmark location \mathbf{p}_i within \mathbf{B}^* is considered admissible. Similarly, the probable size of face is proportional to the size of \mathbf{B}^* . Finally, the rotation parameter is inferred from the estimated roll angle of the driver's head. The estimate value within $\pm 20^\circ$ is considered admissible. When the estimated parameters do not satisfy above conditions, they are discarded and the system is re-initialized. This helps reduce false detection and also improve tracking quality (accuracy and failure rate). This is the case since during the optimization process an initial guess of the parameters is based on the previous output. When

there is no output from the previous frame, the face detection output in the current frame is used to initialize the parameters. Discarding the estimation, however, amounts to no system output which we account for in one of the performance metrics as explained in the Section A.5.2.

Mixture of Pictorial Structures (MPS)

Using pictorial structures, a face is modeled by a collection of parts arranged in a deformable configuration, where each part captures the local visual descriptions of the face and spring-like connections between certain pair of parts capture the deformable configuration [35]. This is naturally represented by an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the vertices $\mathcal{V} = \{v_1, \dots, v_n\}$ correspond to the n parts, and for each pair of connected parts, there exists an edge $(v_i, v_j) \in \mathcal{E}$. A mixture of pictorial structures further captures the topological changes of the face due to varying head orientations. The best configuration of parts is found by maximizing a score function that measures both the appearance similarity, $S_A(I_t, \mathbf{p}_i, m)$, of placing the i^{th} part (i.e. node v_i) at location $\mathbf{p}_i = (x_i, y_i)$, and the likely deformation, $S_D(I_t, \mathbf{p}_i, \mathbf{p}_j, m)$, for each pair of connected parts. Optimization proceeds by maximizing over all mixtures:

$$S(\mathcal{I}_t, \mathbf{P}, m) = \sum_{i=0}^{n_m-1} S_A(\mathcal{I}_t, p_i, m) + \sum_{(v_i, v_j) \in \mathcal{E}} S_D(\mathcal{I}_t, \mathbf{p}_i, \mathbf{p}_j, m)$$

$$S^*(\mathcal{I}_t, \mathbf{P}^*) = \max_{m \in \{\hat{m}_t\}} \left[\max_{\mathbf{p} \in \{\hat{\mathbf{P}}_t\}} S(\mathcal{I}_t, \mathbf{p}, m) \right]$$

where $\hat{m}_t \in M$ is a subset of all mixtures M and $\hat{\mathbf{P}}_t$ is a rectangular region of interest defined by Eq. A.6 and Eq. A.8 respectively.

In literature, the appearance similarity $S_A(I_t, \mathbf{p}_i, \cdot)$ is modeled in various ways e.g. Gaussian derivative filter response around a point [33], feature based description such as Histogram of Gradient (HoG) [122], Haar-like feature [31] etc. $S_D(I_t, \mathbf{p}_i, \mathbf{p}_j, \cdot)$ is a distance function e.g. a Mahalanobis distance in some transformed space of \mathbf{p}_i and \mathbf{p}_j . In our implementation, we use a discriminative, max-margin framework [122] to model the two scoring functions. Here, \mathcal{G} for each mixture is a tree and the optimization is performed efficiently with dynamic programming [33]. To improve the computation time, we further incorporated spatio-temporal constraints to reduce the search space. First, the possible solutions for a configuration of parts are constrained to lie within a region where the head was found in the previous frame. Second, the enumerations over all mixture components for the current frame can be reduced to neighboring mixture components

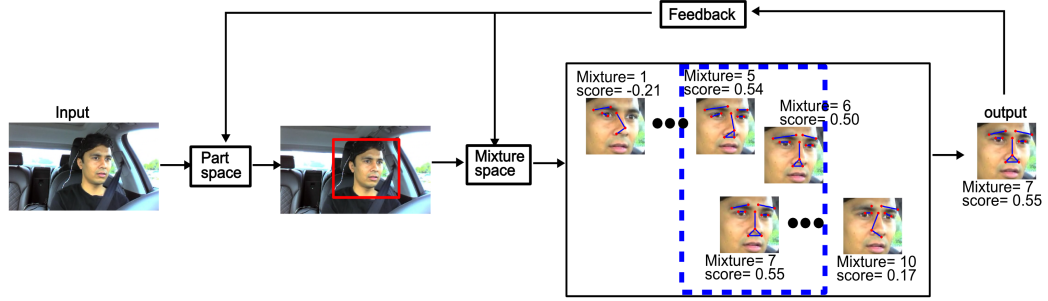


Figure A.4: Process of reducing search space for video analysis using mixture of pictorial structures. Part space is reduced by constraining to region around the face location in the previous frame as illustrated by the red box. Similarly, mixture space is reduced by searching over neighboring mixture components around the estimated component from the previous frame.

around the estimate from the previous frame:

$$\hat{m}_t = m_{t-l}^* + \{-1, 0, 1\} \quad (\text{A.6})$$

$$m_{t-1}^* = \arg \max_{m \in \{\hat{m}_{t-1}\}} \left[\max_{\mathbf{p} \in \{\hat{\mathbf{P}}_{t-1}\}} S(I_t, \mathbf{p}, m) \right] \quad (\text{A.7})$$

$$\hat{\mathbf{P}}_t = \{\mathbf{p}_i \mid \mathbf{p}_i \in (\mathbf{BR}(\mathbf{P}_{t-1}^*) + (-b, -b, +b, +b,))\} \quad (\text{A.8})$$

where m_{t-l}^* is the mixture chosen for the previous frame I_{t-1} , $\mathbf{BR}(\cdot)$ is defined in Eq. A.5 and b is the border width. Figure A.4 depicts the overall process. These optimizations decreased the processing time by at least 4 folds.

A.4.2 Pose Estimation

Given a 3D model of an object, POS (Pose from Orthography and Scaling) [20] finds the position and orientation of the camera coordinate with respect to the object reference frame. It minimizes the reprojection error using weak perspective transform. Given a point on 3D model, say M_i , and its measured projection in the image plane, say $p_i = (x_i, y_i)$, POS solves the following linear system of equations:

$$M_0 M_i \cdot \alpha \mathbf{i} = x_0 x_i \quad i = 1 \cdots N_c$$

$$M_0 M_i \cdot \alpha \mathbf{j} = y_0 y_i \quad i = 1 \cdots N_c$$

where, $M_0 M_i$ represents the vector from the reference point on the 3D model M_0 to M_i , α is the scale factor associated with the weak perspective projection and N_c is number of $3D - 2D$ point correspondences. The vectors \mathbf{i} and \mathbf{j} form the first two rows of the rotation matrix and

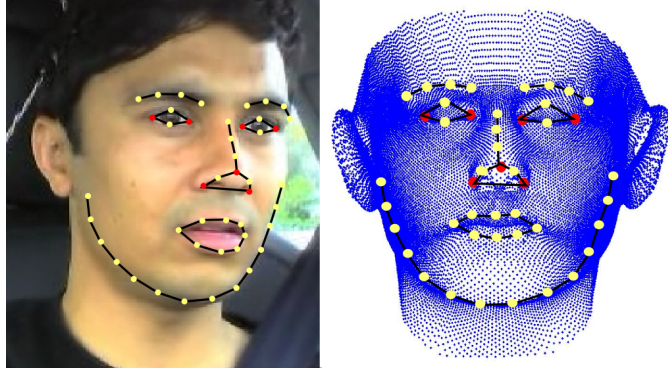


Figure A.5: Tracked facial feature/landmarks and their correspondences in 3D face image. Solid red circles are the points utilized for the head pose calculation.

the third row is given by the vector $\mathbf{k} = \mathbf{i} \times \mathbf{j}$, a cross product. Note, however, that while \mathbf{k} is perpendicular to \mathbf{i} and \mathbf{j} , vectors \mathbf{i} and \mathbf{j} aren't necessarily perpendicular due to noisy $3D - 2D$ point correspondences. Therefore, the rotation matrix is projected into the $SO(3)$ space by normalizing the magnitude of the eigenvalues.

To solve this system of equations, POS requires at least 4 points of correspondences in general positions. CLM and mixture of pictorial structures model, however, output more than 4 fiducial points. In our current implementation, we use the following fiducial points as they are less deformable: four eye corners, two nose corners and a nose tip. Figure A.5 shows these points (red solid circle) on a test image and its corresponding points on the 3D mean face model.

A.4.3 Perspective Selection Procedure

CoHMEt tracks head independently in each camera stream and their outputs are further analyzed to choose the best perspective and corresponding head pose. The block diagram in Figure A.6 illustrates this process for a general setup of N cameras, where the cameras are numbered in the increasing order from the leftmost position in the distributed camera array setup. In the proposed system, we utilize three cameras positioned to the left, front and right of the driver as seen in Figure A.8. The system is initialized with the front camera and during the tracking phase, transitions from one perspective to another is allowed based on the operating range $(\Omega_-^{N_c}, \Omega_+^{N_c})$ of the selected camera and yaw movement direction. When tracking is lost, due to either loss of facial point detection or rejection of the estimated points, re-initialization is performed using a scoring criteria. For the CLM based approach, the system is re-initialized with the perspective that has the highest symmetry score, where the symmetry of the face is computed using the detected facial landmarks. This ensures perspective close to frontal position is chosen. For the MPS based system, the score obtained during facial landmark detection (explained in Section A.4.1) is utilized. Figure A.7 illustrates this process as applied to the data from a naturalistic on-road driving.

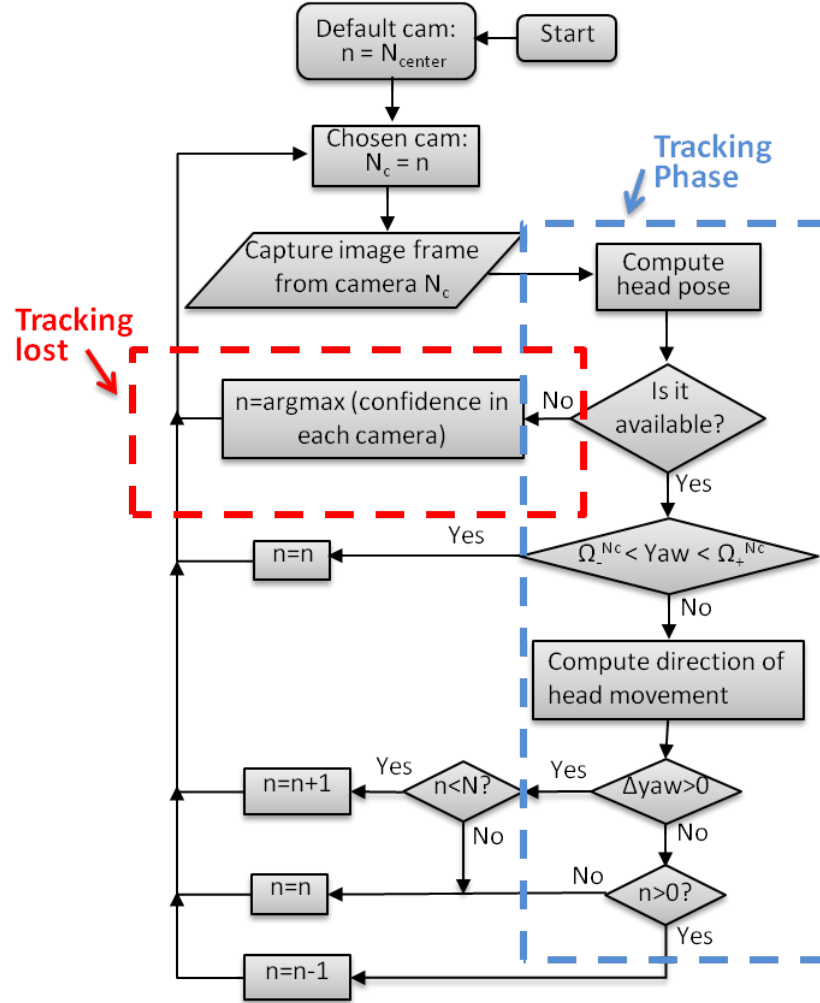


Figure A.6: Perspective selection approach. Tracking phase utilizes head pose and dynamics to switch between perspectives, while a scoring criterion during a lost track re-initializes with the highest score camera.

A.5 Experimental Evaluations and discussion

The CoHMEt framework is evaluated on naturalistic driving data. The data collection was focused and targeted around various maneuvers and events that cause large head turns (away from the driving direction) as they are of special interest for driver safety. By evaluating on these select events, we show the need and usefulness of a multi-perspective setup for continuous and reliable head tracking. Ground truth head pose is generated by mounting an inertial sensor on top of the driver’s head and retrieving the head rotations in pitch, yaw and roll angles.

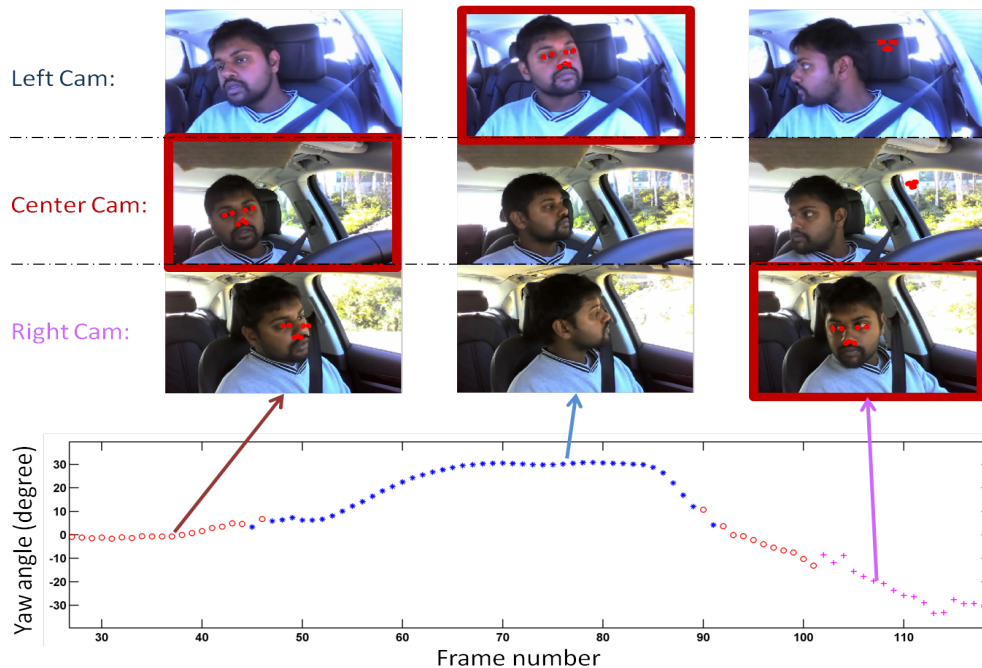


Figure A.7: Illustration of multiple perspective framework on a segment taken from a subject's naturalist on-road driving experiment. The horizontal axis represents frame number (w.r.t. left camera) and the vertical axis represents head rotations in the yaw rotation angle relative to car reference frame. The blue asterisks represent the left camera, the red circles represent the center camera and the magenta crosses represent the right camera. The plot shows the head scan by the driver from left to the right mirror starting from front pose. The evolution of the perspective selection is presented.

A.5.1 Testbed and Dataset

Data is collected from naturalistic, on-road driving using the LISA-A testbed as shown in Figure A.8. Three cameras are mounted facing the driver: one on the front windshield near the A-pillar, one on the front windshield and one near the rear view mirror. They capture face view in color video stream at 30fps and 640×360 pixel resolution.

In addition, the vehicle is instrumented with Inertial Motion Units (IMUs) with sensors placed on the divers head and fixed at the back of the car to track their respective motions. Sensor fusion of the IMUs' data provide precise ground truth head pose data for evaluation. Sensor fusion is required since the IMU attached to the driver's head is affected by the car's movement. To compensate this effect, an IMU rigidly fixed to the car is used to capture vehicle dynamics. The multiple IMUs involve calibrated accelerometer- and gyroscope-sensors. The IMU unit, however, has some drift associated with the gyroscope, a commonly known phenomenon. This is overcome by resetting angle calculation in the beginning of each event where initial orientation is provided by hand annotating the face image. Since, on average, each event lasts around 10 seconds, the drift during this period is practically non existent.

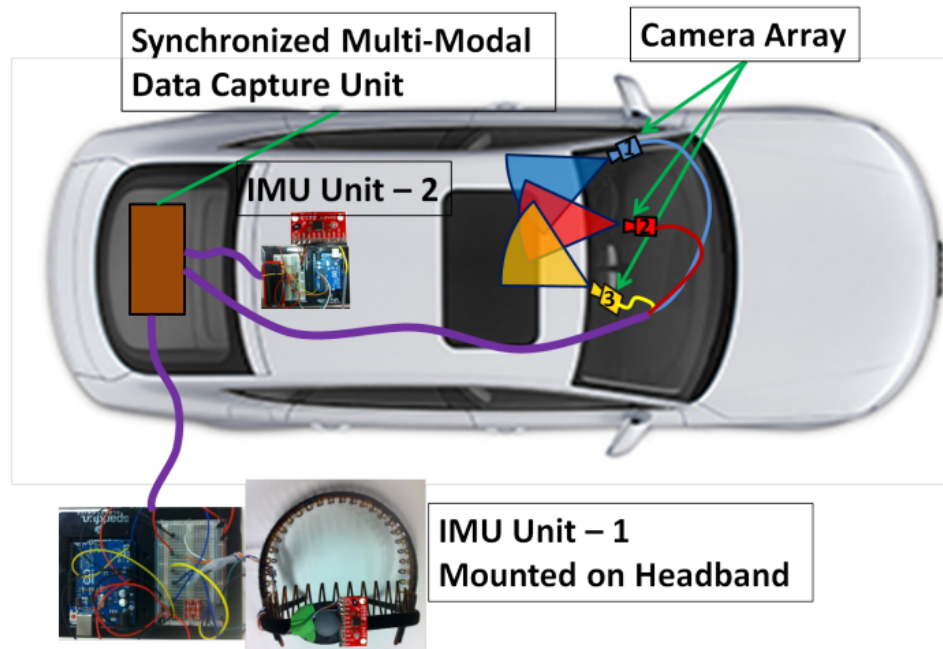


Figure A.8: LISA-A experimental testbed equipped with and capable of time synchronized capture of camera array and multiple Inertial Measurement Units (IMUS) [98].

Using this testbed, multiple drivers were asked to drive naturally on local streets and freeways near UCSD. Approximately 60 minutes of data was collected in total and sunny weather conditions allowed for varying lighting conditions. Additionally, driving in an urban setting, the drivers passed through many stop signs and made multiple turns, and driving on the freeway allowed for multiple lane change occurrences, resulting in a dataset with wide spatial changes in head pose.

From the collected data, we select events when the driver is making right/left turns, right/left lane changes, stops at stop signs, and freeway merges. Table A.2 shows the events considered, the number of respective events analyzed during the total 60 minute drive containing all drivers, and the total number of frames accumulated for each event. The evaluations reported in the following section will be on these selected events. It is important to note that each event can induce more than one sequence of spatially wide head movements. Figure A.9 shows a typical histogram of yaw angle distribution during a test drive. It can be seen that while considering the entire drive, the driver is near frontal facing most of the time (Fig A.9a). However, yaw angle distribution is more spread out for the chosen events (Fig A.9b).

A.5.2 On-road Performance Evaluation

A series of experiments involving the naturalistic on-road data are conducted to characterize the performance of CoHMEt. Performance of CoHMEt, with 3-cameras and 2-cameras,

Table A.2: A list of events considered for evaluation, and its respective count and number of frames.

Events	No. of events	Total no. of frames
Right turns	27	7417
Left turns	13	4027
Stop sign	42	10853
Right Lane change	12	2565
Left Lane Change	15	2963
Merge	7	1778
Xing	3	628
All	119	22814

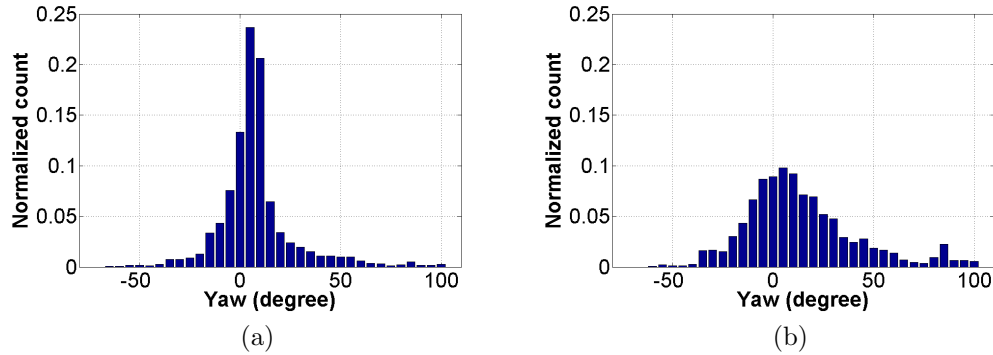


Figure A.9: Distribution of head pose values in yaw rotation angle for an entire test drive (a) and accumulation of selected events (b) in the same test drive. Clearly, data from chosen events shows a more even distribution across yaw when compared to the entire drive.

are compared against the performance of a single-view approach. Spatial distribution of cameras for a 3-, 2- and 1-camera view from a top-down perspective is illustrated in Figure A.10. For a quantitative evaluation over the database, three metrics are used: mean absolute error (MAE), standard deviation error (STD) and failure rate (percentage of the time when system’s output is unreliable). Head tracking is considered to be lost if the estimated head pose is not available or is more than 20° from the ground truth in either direction of the yaw rotation angle. Number of frames where head tracking is lost, normalized by the total number of frames over all events gives the failure rate.

As shown in Table A.5.2, the MPS+POS system shows a general trend of improvement in failure rate from 1-camera view to 2-camera view to 3-camera view. The best performance of 3.9% failure rate is achieved with the 3-cameras view compared to that of over 15% for the single view, a significant improvement. However, for the CLM+POS, the 2-camera view performed the worst. This is because CLM+POS requires near frontal pose for initialization and front camera is absent in the 2-camera view configuration. The 3-camera view with front camera again performed the best, dropping failure rate by a half compared to single front camera view. Hence, choice

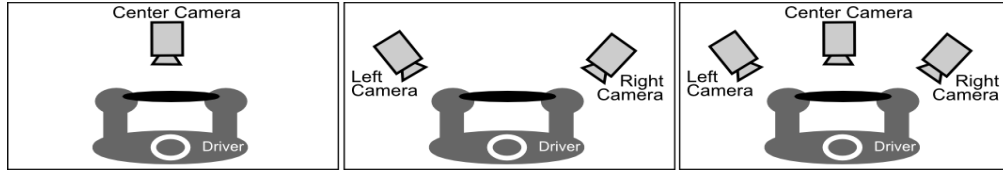


Figure A.10: Shows the setup of the single-camera view, 2-camera view and 3-camera view as discussed and compared for performance evaluation of the multi-view framework. The single-view setup is composed of the center camera only. The 2-camera view setup is composed of the left and right camera. The 3-camera view setup is composed of the left, center and right camera.

of the algorithms and the camera configurations are tightly coupled. This is further discussed below. From Table A.5.2, also notice that irrespective of the number of cameras (observing each column), MPS+POS algorithm outperforms CLM+POS approach. This can be attributed to the ability of the MPS formulation to incorporate global topological variation of the facial landmark due to different pose using different mixture components. The drawback of MPS, however, is computational complexity. In our experiments, using Intel 3.0 GHz CPU, MPS+POS without search space reduction, as explained in Section A.4.1, took ~ 13 seconds to process one frame and with search space reduction took ~ 3 seconds for one frame, a four fold improvement. CLM+POS, on the other hand, runs real-time with ~ 25 frames per second.

The MAE and STD for pitch, yaw and roll are relatively similar across different configurations and the placements of cameras. This is expected since a multiple camera system combines each camera independently and is bounded by the single view accuracy. While a direct comparison to other reported error rates in the literature may not be appropriate e.g due to different databases, to put in perspective, we refer to the results of the study by Murphy-Chutorian and Trivedi [71] evaluated on on-road data. The authors reported $MAE < 5^\circ$ in yaw angle when the system is initialized with ground truth. However, the fully automatic system, had $MAE > 10^\circ$ in yaw angle with large $STD \approx 17^\circ$. Our proposed framework, evaluated on the challenging naturalistic on-road dataset, has shown good results.

Next, we show in Figure A.11 the absolute yaw error statistics as a function of true yaw angle with respect to front camera. The figure shows first-, second- and third quartile of the errors associated with the respective yaw bins. It can be observed that the single camera system quickly loses tracks with high estimation error beyond 30° in either direction. While the multi-camera system is able to keep track over much wider span with better error statistics. Also, notice that higher errors are associated at the two extremes, which is due to decreased visibility of facial landmarks caused by self occlusion.

Finally, we conduct an experiment to study the operational range of the system and for a given choice of an algorithm, how to obtain ‘best’ camera placement. We chose the MPS+POS system for this experiment since the MPS formulation provides a facial feature detection score (higher the better), which we refer to here as quality. Figure A.12 shows the quality of each

Table A.3: On-road performance evaluations of the proposed CoHMET.

Method	Single camera view			CoHMET with 2-camera view			CoHMET with 3-camera view			
	Pitch	Yaw	Roll	Pitch	Yaw	Roll	Pitch	Yaw	Roll	Failure Rate
CLM+POS	(9.3°, 10.3°)	(6.9, 8.7°)	(3.4°, 5.5°)	(8.8°, 12.2°)	(5.7°, 7.5°)	(3.6°, 5.2°)	(8.5°, 9.9°)	(5.5°, 7.2°)	(2.7°, 4.0°)	11 %
MPS+POS	(7.6, 9.4°)	(8.2°, 6.4°)	(3.0°, 4.3°)	(8.6°, 9.7°)	(7.0°, 7.3°)	(3.8°, 5.4°)	(9.0°, 9.4°)	(5.9°, 6.9°)	(3.5°, 4.9°)	4%

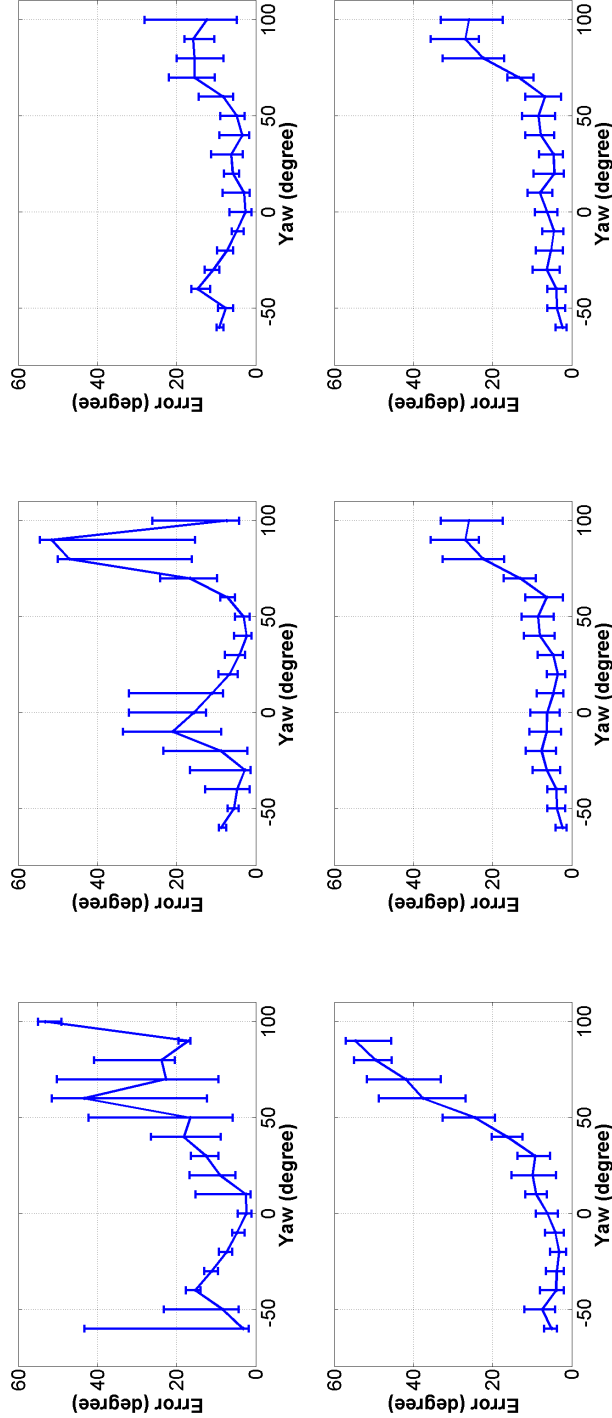


Figure A.11: Error distribution with respect to the true head pose in yaw. The graphs reflects the first three error quartiles for single perspective (1st column), 2-camera perspective (2nd column) and 3-camera perspective (3rd column) using CLM+POS (1st row) and MPS+POS (2nd row)

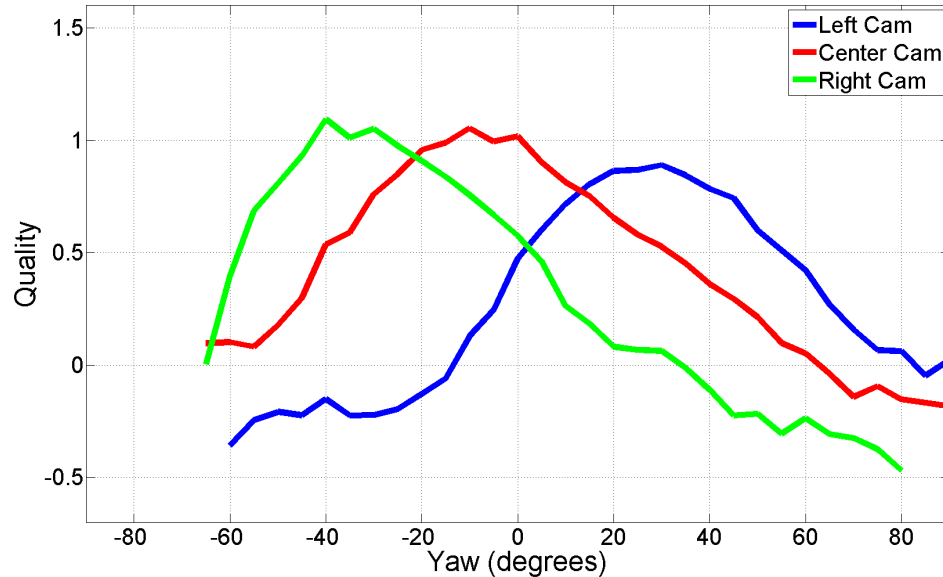


Figure A.12: Quality of head pose estimation from individual camera view with respect to head orientations in the yaw angle. A useful means of configuring camera positions to maximize operational range of the overall system.

of the three cameras as a function of true yaw angle. Given a desired level of quality, this can provide the operational range of a camera and how cameras should be placed with respect to each other to maximize the operational range of the overall system.

A.6 Concluding Remarks

Robust systems for observing the driver behavior will play a key role in the development of IDAS. Analyzing the driver’s head movement is becoming an increasingly important aspect of such systems, since it is a strong indicator of the driver’s field of view, current focus of attention as well as intent. In a driving environment, the driver is prone to make large spatial head movements during maneuvers such as lane changes, right/left turns. During these crucial moments, it is important to continuously and reliably track the head of the driver. Moreover, the system needs to perform uninterrupted with high accuracy to be accepted and trusted by the driver.

In this chapter, we proposed the CoHMEt to address the above design criteria. We presented two approaches of facial feature tracking to compute head pose and conducted systematic comparative studies with different configurations of multiple-camera systems. The best system could reliably track head movement over 96% of the time. The evaluations are performed over the naturalistic real-world driving dataset, which is a must as they present the actual scenario.

Towards this end, we collected a unique and novel dataset of naturalistic driving with distributed cameras. The dataset targets spatially large head turns (away from the driving direction) during different maneuvers (e.g. merge, lane change) on urban streets and freeways. Going forward, we will pursue a unified framework to combine the two approaches to improve computational cost without sacrificing failure rate and head pose error. Finally, CoHMEt framework can also be adapted in other “Intelligent Environments” with multiple participants [70] and multiple sensory cues [92].

A.7 Acknowledgments

Appendix A is in full a reprint of material that is published in the IEEE Transactions on Intelligent Transportation Systems (2014), by Ashish Tawari, Sujitha Martin and Mohan M. Trivedi. The dissertation author was one of the primary investigator and author of this paper.

Bibliography

- [1] Traffic safety facts. *Department of Transportation-National Highway Traffic Safety Administration*, 2014.
- [2] Prachi Agrawal and PJ Narayanan. Person de-identification in videos. *Circuits and Systems for Video Technology, IEEE Transactions on*, 2011.
- [3] Christer Ahlstrom, Katja Kircher, and Albert Kircher. A gaze-based driver distraction warning system and its effect on visual behavior. *Intelligent Transportation Systems, IEEE Transactions on*, 2013.
- [4] Kellie J Archer and Ryan V Kimes. Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52(4):2249–2260, 2008.
- [5] Simon Baker, Iain Matthews, Jing Xiao, Ralph Gross, Takeo Kanade, and Takahiro Ishikawa. Real-time non-rigid driver head tracking for driver mental state estimation. In *11th World Congress on Intelligent Transportation Systems*, 2004.
- [6] Tobias Bär, Denys Linke, Dennis Nienhüser, and J. Marius Zöllner. Seen and missed traffic objects: A traffic object-specific awareness estimation. In *IEEE Intelligent Vehicles Symposium*, pages 31–37, June 2013.
- [7] Tobias Bär, Jan Felix Reuter, and J. Marius Zöllner. Driver head pose and gaze estimation based on multi-template icp 3-d point cloud alignment. In *IEEE Conference on Intelligent Transportation Systems (ITSC)*, 2012.
- [8] Jorge P. Batista. A real-time driver visual attention monitoring system. In *Proceedings of the Second Iberian conference on Pattern Recognition and Image Analysis - Volume Part I*. Springer-Verlag, 2005.
- [9] K. Bengler, K. Dietmayer, B. Farber, M. Maurer, C. Stiller, and H. Winner. Three decades of driver assistance systems: Review and future perspectives. *IEEE Intelligent Transportation Systems Magazine*, 2014.
- [10] Luis Miguel Bergasa, Jesús Nuevo, Miguel A Sotelo, Rafael Barea, and María Elena Lopez. Real-time system for monitoring driver vigilance. *Intelligent Transportation Systems, IEEE Transactions on*, 2006.
- [11] Stewart A Birrell and Mark Fowkes. Glance behaviours when using an in-vehicle smart driving aid: A real-world, on-road driving study. *Transportation research part F: traffic psychology and behaviour*, 2014.
- [12] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *IEEE Intl. Conf. Computer Vision*, 2013.

- [13] Shinko Yuanhsien Cheng and Mohan M. Trivedi. Turn-intent analysis using body pose for intelligent driver assistance. *IEEE Pervasive Computing*, 2006.
- [14] Shinko Yuanhsien Cheng and Mohan Manubhai Trivedi. Turn-intent analysis using body pose for intelligent driver assistance. *Pervasive Computing, IEEE*, 2006.
- [15] Robert T Collins, Ralph Gross, and Jianbo Shi. Silhouette-based human identification from body shape and gait. In *Automatic Face and fGesture Recognition, 2002. Fifth IEEE International Conference on*, 2002.
- [16] Timothy F Cootes and Christopher J Taylor. Active shape models - smart snakes. In *In British Machine Vision Conference*, pages 266–275. 1992.
- [17] D. Cristinacce and T. F. Cootes. A comparison of shape constrained facial feature detectors. In *In 6th International Conference on Automatic Face and Gesture Recognition*, 2004.
- [18] David Cristinacce and Timothy F Cootes. Feature detection and tracking with constrained local models. In *BMVC*, 2006.
- [19] David Cristinacce and Timothy F. Cootes. Feature detection and tracking with constrained local models. In *Proceedings of the British Machine Vision Conference*, 2006.
- [20] Daniel F Dementhon and Larry S Davis. Model-based object pose in 25 lines of code. *International journal of computer vision*, 1995.
- [21] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona. Fast feature pyramids for object detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2014.
- [22] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2012.
- [23] Anup Doshi, Shinko Yuanhsien Cheng, and Mohan M Trivedi. A novel active heads-up display for driver assistance. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 2009.
- [24] Anup Doshi, Brendan T Morris, and Mohan M Trivedi. On-road prediction of driver’s intent with multimodal sensory cues. *Pervasive Computing, IEEE*, 2011.
- [25] Anup Doshi and Mohan M Trivedi. On the roles of eye gaze and head dynamics in predicting driver’s intent to change lanes. *Intelligent Transportation Systems, IEEE Transactions on*, 2009.
- [26] Anup Doshi and Mohan M Trivedi. Attention estimation by simultaneous observation of viewer and view. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010.
- [27] Anup Doshi and Mohan M Trivedi. Head and eye gaze dynamics during visual attention shifts in complex environments. *Journal of vision*, 2012.
- [28] Anup Doshi and Mohan M. Trivedi. Head and eye gaze dynamics during visual attention shifts in complex environments. *Journal of Vision*, 2012.
- [29] Frédéric Dufaux and Touradj Ebrahimi. A framework for the validation of privacy protection solutions in video surveillance. In *Multimedia and Expo (ICME), IEEE International Conference on*. IEEE, 2010.

- [30] Sujitha Martin Eshed Ohn-Bar, Ashish Tawari, and Mohan Trivedi. Head, eye, and hand patterns for driver activity recognition. In *IEEE International Conference on Pattern Recognition*. Citeseer, 2014.
- [31] M. Everingham, J. Sivic, and A. Zisserman. “Hello! My name is... Buffy” – automatic naming of characters in TV video. In *British Machine Vision Conference*, 2006.
- [32] Sachin Sudhakar Farfade, Mohammad J Saberian, and Li-Jia Li. Multi-view face detection using deep convolutional neural networks. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 2015.
- [33] Pedro F Felzenszwalb and Daniel P Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 2005.
- [34] Douglas A Fidaleo, Hoang-Anh Nguyen, and Mohan Trivedi. The networked sensor tapestry (nest): a privacy enhanced software architecture for interactive analysis of data in video-sensor networks. In *Proceedings of the ACM workshop on Video surveillance & sensor networks*. ACM, 2004.
- [35] Martin A Fischler and Robert A Elschlager. The representation and matching of pictorial structures. *Computers, IEEE Transactions on*, 1973.
- [36] Luke Fletcher, Nicholas Apostoloff, Lars Petersson, and Alexander Zelinsky. Vision in and out of vehicles. *Intelligent Systems, IEEE*, 2003.
- [37] Luke Fletcher and Alexander Zelinsky. Driver inattention detection based on eye gazeroad event correlation. *The international journal of robotics research*, 2009.
- [38] Arturo Flores and Serge Belongie. In *Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE Computer Society Conference on*. IEEE.
- [39] Lex Fridman, Philipp Langhans, Joonbum Lee, and Bryan Reimer. Driver gaze region estimation without using eye movement. *IEEE Intelligent Systems*, 2016, Accepted.
- [40] Lex Fridman, Joonbum Lee, Bryan Reimer, and Trent Victor. Owl and lizard: Patterns of head pose and eye pose in driver gaze classification. *IET Computer Vision*, 2016, In Print.
- [41] Andrea Frome, German Cheung, Ahmad Abdulkader, Marco Zennaro, Bo Wu, Alessandro Bissacco, Hartwig Adam, Hartmut Neven, and Luc Vincent. Large-scale privacy protection in google street view. In *Computer Vision, IEEE 12th International Conference on*. IEEE, 2009.
- [42] X. Fu, X. Guan, E. Peli, H. Liu, and G. Luo. Automatic calibration method for driver’s head orientation in natural driving environment. *IEEE Transactions on Intelligent Transportation Systems*, 2013.
- [43] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 2013.
- [44] Christian Gold, Daniel Damböck, Lutz Lorenz, and Klaus Bengler. take over! how long does it take to get the driver back into the loop? In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. SAGE Publications, 2013.
- [45] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 2010.

- [46] Leon Gu and Takeo Kanade. A generative shape regularization model for robust face alignment. In *Proceedings of the 10th European Conference on Computer Vision: Part I, ECCV '08*, pages 413–426, 2008.
- [47] Elias Daniel Guestrin and Moshe Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Trans. Biomed. Engineering*, 53, 2006.
- [48] Zhibo Guo, Huajun Liu, Qiong Wang, and Jingyu Yang. A fast algorithm face detection and head pose estimation for driver assistant system. In *Signal Processing, 2006 8th International Conference on*. IEEE, 2006.
- [49] Dan Witzner Hansen and Qiang Ji. In the eye of the beholder: A survey of models for eyes and gaze. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2010.
- [50] M.B. Holte, Cuong Tran, M.M. Trivedi, and T.B. Moeslund. Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments. *IEEE Journal of Selected Topics in Signal Processing*, 2012.
- [51] Kohsia S. Huang and Mohan M. Trivedi. Robust real-time detection, tracking, and pose estimation of faces in video streams. In *Proceedings of the 17th International Conference on Pattern Recognition*, 2004.
- [52] Ashesh Jain, Hema S Koppula, Bharad Raghavan, Shane Soh, and Ashutosh Saxena. Car that knows before you do: Anticipating maneuvers via learning temporal driving models. *ICCV*, 2015.
- [53] Vidit Jain and Erik Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.
- [54] Sheila G Klauer, Thomas A Dingus, Vicki L Neale, Jeremy D Sudweeks, David J Ramsey, et al. The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data. 2006.
- [55] Christopher K. Kovach and Ralph Adolphs. Investigating attention in complex visual search. *Vision Research*, 2015.
- [56] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 2012.
- [57] Sung Joo Lee, Jaek Jo, Ho Gi Jung, Kang Ryoung Park, and Jaihie Kim. Real-time gaze estimator based on driver’s head orientation for forward collision warning system. *Intelligent Transportation Systems, IEEE Transactions on*, 2011.
- [58] Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua. A convolutional neural network cascade for face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- [59] Nanxiang Li and Carlos Busso. Detecting drivers’ mirror-checking actions and its application to maneuver and secondary task recognition. *IEEE Transactions on Intelligent Transportation Systems*, 17(4):980–992, 2016.
- [60] Yuan Liao, Shengbo Eben Li, Wenjun Wang, Ying Wang, Guofa Li, and Bo Cheng. Detection of driver cognitive distraction: A comparison study of stop-controlled intersection and speed-limited highway. *Transactions on Intelligent Transportation Systems*, 2016.
- [61] Xiangpeng Liu and Axel Graeser. Robust face detection with eyes occluded by the shadow from dazzling avoidance system. In *Intelligent Transportation Systems (ITSC), IEEE Conference on*. IEEE, 2015.

- [62] Sujitha Martin, Eshed Ohn-Bar, Ashish Tawari, and Mohan Manubhai Trivedi. Understanding head and hand activities and coordination in naturalistic driving videos. In *Intelligent Vehicles Symposium Proceedings*. IEEE, 2014.
- [63] Sujitha Martin, Eshed Ohn-Bar, and Mohan M Trivedi. Automatic critical event extraction and semantic interpretation by looking-inside. In *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2015.
- [64] Sujitha Martin, Eshed Ohn-Bar, Kevan Yuen, Rakesh Rajaram, and Mohan M. Trivedi. Vision for intelligent vehicles and application. <http://cvrr.ucsd.edu/vivachallenge>. Accessed: 2016-08-25.
- [65] Sujitha Martin, Ashish Tawari, Erik Murphy-Chutorian, Shinko Y Cheng, and Mohan Trivedi. On the design and evaluation of robust head pose for visual user interfaces: algorithms, databases, and comparisons. In *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, 2012.
- [66] Sujitha Martin, Ashish Tawari, and Mohan M. Trivedi. Monitoring head dynamics for driver assistance systems: A multi-perspective approach. In *IEEE International Conference on Intelligent Transportation Systems*, 2013.
- [67] Joel C McCall, David P Wipf, Mohan M Trivedi, and Bhaskar D Rao. Lane change intent analysis using robust operators and sparse bayesian learning. *Intelligent Transportation Systems, IEEE Transactions on*, 2007.
- [68] E. Murphy-Chutorian and M.M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.
- [69] Erik Murphy-Chutorian, Anup Doshi, and Mohan M. Trivedi. Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation. In *IEEE Conference on Intelligent Transportation Systems*, 2007.
- [70] Erik Murphy-Chutorian and Mohan M. Trivedi. 3d tracking and dynamic analysis of human head movements and attentional targets. In *Second ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*, 2008.
- [71] Erik Murphy-Chutorian and Mohan M. Trivedi. Head Pose Estimation and Augmented Reality Tracking: An Integrated System and Evaluation for Monitoring Driver Awareness. *IEEE Transactions on Intelligent Transportation Systems*, 2010.
- [72] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head pose estimation and augmented reality tracking: An integrated system and evaluation for monitoring driver awareness. *Intelligent Transportation Systems, IEEE Transactions on*, 2010.
- [73] Angelo Nodari, Marco Vanetti, and Ignazio Gallo. Digital privacy: Replacing pedestrians from google street view images. In *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012.
- [74] E. Ohn-Bar and M. M. Trivedi. Looking at humans in the age of self-driving and highly automated vehicles. *IEEE Transactions on Intelligent Vehicles*, 2016.
- [75] Eshed Ohn-Bar, Ashish Tawari, Sujitha Martin, and Mohan M Trivedi. On surveillance for safety critical events: In-vehicle video networks for predictive driver assistance systems. *Computer Vision and Image Understanding*, 2015.

- [76] Kenichi Ohue, Yukinori Yamada, Shigeyasu Uozumi, Setsuo Tokoro, Akira Hattori, and Takeshi Hayashi. Development of a new pre-crash safety system. In *SAE 2006 World Congress & Exhibition*, SAE Technical Paper 2006-01-1461, 2006.
- [77] C. Olaverri-Monreal, C. Lehsing, N. Trubswetter, C.A. Schepp, and K. Bengler. In-vehicle displays: Driving information prioritization and visualization. In *Intelligent Vehicles Symposium (IV)*, IEEE, 2013.
- [78] Ralph Oyini Mbouna, Seong G Kong, and Myung-Geun Chun. Visual analysis of eye state and head pose for driver alertness monitoring. *Intelligent Transportation Systems, IEEE Transactions on*, 2013.
- [79] J. Paone, D. Bolme, R. Ferrell, D. Aykac, and T. Karnowski. Baseline face detection, head pose estimation, and coarse direction detection for facial data in the shrp-2 naturalistic driving study. In *Intelligent Vehicles Symposium (IV)*, IEEE, 2015.
- [80] Daniel Parks, Ali Borji, and Laurent Itti. Augmented saliency model using automatic 3d head pose detection and learned gaze following in natural scenes. *Vision Research*, 2015.
- [81] Zheng Pei, Song Zhenghe, and Zhou Yiming. Perclos-based recognition algorithms of motor driver fatigue. *Journal-China Agricultural University*, 2002.
- [82] Pietro Perona and Jitendra Malik. Scale-space and edge detection using anisotropic diffusion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 1990.
- [83] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *IEEE Conf. Computer Vision and Pattern Recognition*, 2012.
- [84] Faisal Z Qureshi. Object-video streams for preserving privacy in video surveillance. In *Advanced Video and Signal Based Surveillance, Sixth IEEE International Conference on*. IEEE, 2009.
- [85] A. Rangesh, E. Ohn-Bar, and M. Trivedi. Hidden hands: Tracking hands with an occlusion aware tracker. In *IEEE Conf. Computer Vision and Pattern Recognition Workshops-HANDS*, 2016.
- [86] Akshay Rangesh, Eshed Ohn-Bar, and Mohan Manubhai Trivedi. Long-term, multi-cue tracking of hands in vehicles. *Transactions on Intelligent Transportation Systems (in press)*, in press, 2016.
- [87] C. Rodemer, H. Winner, and R. Kastner. Predicting the drivers turn intentions at urban intersections using context-based indicators. In *Intelligent Vehicles Symposium (IV)*, IEEE, 2015.
- [88] J3016 S. International. Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems, 2014.
- [89] J.M. Saragih, S. Lucey, and J.F. Cohn. Face alignment through subspace constrained mean-shifts. In *Computer Vision, 2009 IEEE 12th International Conference on*, 2009.
- [90] Jeremy Schiff, Marci Meingast, Deirdre K Mulligan, Shankar Sastry, and Ken Goldberg. Respectful cameras: Detecting visual markers in real-time to address privacy concerns. In *Protecting Privacy in Video Surveillance*. Springer, 2009.
- [91] Rajinda Senaratne, David Hardy, Bill Vanderaa, and Saman Halgamuge. Driver fatigue detection by fusing multiple cues. In *Advances in Neural Networks-ISNN 2007*. Springer, 2007.

- [92] Shankar T. Shivappa, Mohan M. Trivedi, and Bhaskar D. Rao. Audiovisual information fusion in human-computer interfaces and intelligent environments: A survey. *Proceedings of the IEEE*, 2010.
- [93] Greg Slabaugh. Computing euler angles from a rotation matrix.
- [94] Ashish Tawari, Sujitha Martin, and Mohan M Trivedi. Privacy preserving approach and filters for camera based intelligent driver assistance and vehicle safety. In *UCSD Technology Transfer Office*, number SD 2014-163.
- [95] Ashish Tawari, Sujitha Martin, and Mohan M. Trivedi. Continuous head movement estimator for driver assistance: Issues, algorithms and on-road evaluations. *Intelligent Transportation Systems, IEEE Transactions on*, 2014.
- [96] Ashish Tawari, Andreas Møgelmoose, Sujitha Martin, Thomas B Moeslund, and Mohan M Trivedi. Attention estimation by simultaneous analysis of viewer and view. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2014.
- [97] Ashish Tawari, Sayanan Sivaraman, Mohan Manubhai Trivedi, Trevor Shannon, and Mario Toppelhofer. Looking-in and looking-out vision for urban intelligent assistance: Estimation of driver attentive state and dynamic surround for safe merging and braking. In *IEEE Intelligent Vehicles Symposium*. IEEE, 2014.
- [98] Ashish Tawari and Mohan M Trivedi. Head dynamic analysis: A multi-view framework. In *New Trends in Image Analysis and Processing-ICIAP*. Springer, 2013.
- [99] Ashish Tawari and Mohan M. Trivedi. Face expression recognition by cross modal data association. *IEEE Transactions on Multimedia*, 2013.
- [100] Ashish Tawari and Mohan M Trivedi. Robust and continuous driver gaze estimation by dynamic analysis of multiple face videos. In *IEEE Intelligent Vehicles Symposium*, (2014).
- [101] T Taylor, AK Pradhan, G Divekar, M Romoser, J Muttart, R Gomez, A Pollatsek, and DL Fisher. The view from the road: The contribution of on-road glance-monitoring technologies to understanding driver behavior. *Accident Analysis & Prevention*, 2013.
- [102] Renran Tian, Lingxi Li, Mingye Chen, Yaobin Chen, and G.J. Witt. Studying the effects of driver distraction and traffic density on the probability of crash and near-crash events in naturalistic driving environment. *Intelligent Transportation Systems, IEEE Transactions on*, 2013.
- [103] Cuong Tran and M.M. Trivedi. 3-d posture and gesture recognition for interactivity in smart spaces. *IEEE Transactions on Industrial Informatics*, 2012.
- [104] Cuong Tran and Mohan Manubhai Trivedi. Driver assistance for “keeping hands on the wheel and eyes on the road”. In *Vehicular Electronics and Safety (ICVES), IEEE International Conference on*, 2009.
- [105] Cuong Tran and Mohan Manubhai Trivedi. 3-d posture and gesture recognition for interactivity in smart spaces. *Industrial Informatics, IEEE Transactions on*, 2012.
- [106] Mohan M Trivedi and Shinko Y. Cheng. Holistic sensing and active displays for intelligent driver support systems. *Computer*, 2007.
- [107] Mohan Manubhai Trivedi, Shinko Yuanhsien Cheng, Edwin Malcolm Clayton Childers, and Stephen Justin Krotosky. Occupant posture analysis with stereo and thermal infrared video: Algorithms and experimental evaluation. *Vehicular Technology, IEEE Transactions on*, 2004.

- [108] Mohan Manubhai Trivedi, Tarak Gandhi, and Joel McCall. Looking-in and looking-out of a vehicle: Computer-vision-based enhanced vehicle safety. *Intelligent Transportation Systems, IEEE Transactions on*, 2007.
- [109] R. Valenti, N. Sebe, and T. Gevers. Combining head pose and eye location information for gaze estimation. *IEEE Transactions on Image Processing*, 21(2):802–815, 2012.
- [110] R. Vasudevan, V. Shia, Yiqi Gao, R. Cervera-Navarro, R. Bajcsy, and F. Borrelli. Safe semi-autonomous control with enhanced driver modeling. In *American Control Conference (ACC)*, 2012.
- [111] Francisco Vicente, Zehua Huang, Xuehan Xiong, Fernando De la Torre, Wende Zhang, and Dan Levi. Driver gaze tracking and eyes off the road detection system. IEEE.
- [112] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 2004.
- [113] B. Wall. Digitizing facial features fails to prevent identification of plaintiff. USA Today, March 1996.
- [114] Qiong Wang, Jingyu Yang, Mingwu Ren, and Yujie Zheng. Driver fatigue detection: a survey. In *Intelligent Control and Automation, 2006. WCICA 2006. The Sixth World Congress on*. IEEE, 2006.
- [115] Yang Wang, Simon Lucey, and Jeffrey Cohn. Enforcing convexity for improved alignment with constrained local models. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.
- [116] Junwen Wu and Mohan M. Trivedi. A two-stage head pose estimation framework and evaluation. *Pattern Recognition*, 2008.
- [117] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. CVPR, 2013.
- [118] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. From facial parts responses to face detection: A deep learning approach. In *International Conference on Computer Vision (ICCV)*, 2015.
- [119] Jang-Hee Yoo, Doosung Hwang, and Mark S Nixon. Gender classification in human gait using support vector machine. In *Advanced concepts for intelligent vision systems*. Springer, 2005.
- [120] Kevan Yuen, Sujitha Martin, and Mohan Trivedi. On looking at faces in a vehicle with deep networks. In *19th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2016.
- [121] Kevan Yuen, Sujitha Martin, and Mohan Trivedi. On looking at faces in an automobile: Issues, algorithms and evaluation on naturalistic driving dataset. In *IEEE International Conference on Pattern Recognition*. Citeseer, 2016.
- [122] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [123] Zhiwei Zhu and Qiang Ji. Real time 3d face pose tracking from an uncalibrated camera. In *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on*. IEEE, 2004.