UNIVERSITY OF CALIFORNIA, SAN DIEGO

Visual Attention from Dynamic Analysis of Head, Eyes and Salient Objects

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

in

Electrical Engineering
(Intelligent Systems, Robotics, and Control)

by

Ashish Tawari

Committee in charge:

Professor Mohan M. Trivedi, Chair
Professor Serge J. Belongie
Professor Garrison W. Cottrell
Professor Edwin L. Hutchins
Professor Bhaskar D. Rao
Professor Nuno M. Vasconcelos

2014
The dissertation of Ashish Tawari is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2014
DEDICATION

To my beloved parents.
The face is a picture of the mind with the eyes as its interpreter.

—Marcus Tullius Cicero
# TABLE OF CONTENTS

Signature Page ................................................................. iii
Dedication ........................................................................ iv
Epigraph ........................................................................... v
Table of Contents .............................................................. vi
List of Figures ..................................................................... viii
List of Tables ....................................................................... xi
Acknowledgements ............................................................ xii
Vita .................................................................................... xiv
Abstract of the Dissertation ..................................................... xvi

Chapter 1 Introduction ........................................................ 1
  1.1 Contributions and Outline ............................................. 2

Chapter 2 Focus of Visual Attention (FoViA) Estimation by Simultaneous Analysis of Viewer and View .......................... 6
  2.1 Introduction ................................................................ 6
  2.2 Related Research ....................................................... 7
  2.3 FoViA: Lookin-in Looking-out Framework ....................... 9
    2.3.1 Proposed Bayesian Model ..................................... 9
    2.3.2 Looking-out via Naturalistic Driving Data ................. 10
    2.3.3 Looking-in via Driver Observation: \( p(a_{z}/l_{z}, T_{i}) \) .... 13
  2.4 Concluding Remarks .................................................. 15

Chapter 3 Continuous Head Movement Estimator: A Distributed Camera Framework ................................. 17
  3.1 Introduction ................................................................ 17
  3.2 Related Research ........................................................ 19
  3.3 Issues and Challenges in Continuous and Robust Head Movement Analysis ................................. 23
  3.4 CoHMEt: Framework and Algorithms ......................... 25
    3.4.1 Facial Feature Detection and Tracking .................... 26
    3.4.2 Pose Estimation ................................................... 30
    3.4.3 Perspective Selection Procedure ............................ 31
  3.5 Testbed and Dataset .................................................... 31
  3.6 Experimental Evaluations and discussion ...................... 34
    3.6.1 On-road Performance Evaluation ............................ 35
  3.7 Concluding Remarks .................................................. 36

Chapter 4 Head, Eye and Iris Dynamic (HEIDy) Cue and Gaze Zone .................................................... 41
  4.1 Introduction ................................................................ 41
  4.2 Related Studies .......................................................... 41
  4.3 Proposed Gaze Zone Estimation Framework .................... 43
    4.3.1 Head Pose Feature Extraction ............................... 44
  4.4 HEIDy Feature Extraction ............................................ 45
    4.4.1 Iris Center Detection and Tracking ....................... 45
# 4.4 Gaze-Surrogate Estimation

- 4.4.2 Gaze-Surrogate Estimation ........................................ 47

# 4.5 Experimental Evaluation

- 4.5.1 Datasets and Annotation ........................................ 48
- 4.5.2 Results: Iris Tracking ........................................... 50
- 4.5.3 Results: Gaze Zone Estimation ................................. 51

# 4.6 Concluding Remarks .................................................. 57

## Chapter 5 Systems and Analysis ...................................... 59

- 5.1 Introduction .......................................................... 59
- 5.2 Attentional Target Detection System ............................ 59
  - 5.2.1 Related Research .................................................. 60
  - 5.2.2 Proposed Approach ............................................. 62
  - 5.2.3 Dataset ........................................................... 66
  - 5.2.4 Experimental Evaluation ...................................... 67
- 5.3 Driver Attention Guard ............................................. 69
  - 5.3.1 Testbed Design and Architecture ............................ 71
  - 5.3.2 Proposed Approach ............................................. 73
  - 5.3.3 Qualitative On-Road Evaluation ............................ 74
- 5.4 Concluding Remarks ................................................. 75

## Chapter 6 Conclusion .................................................... 78

## Appendix A Facial Expression Recognition by Cross Modal Analysis ........................................ 80

- A.1 Introduction and Motivation ....................................... 80
- A.2 Related Studies ...................................................... 81
- A.3 Audio-Visual Data Association Approach ....................... 85
  - A.3.1 Face tracking and alignment ................................. 85
  - A.3.2 Visual sequence analysis: A Bimodal approach .......... 86
  - A.3.3 Appearance feature extraction .............................. 88
  - A.3.4 Auditory feature extraction ................................. 88
- A.4 Experimental Analysis .............................................. 89
  - A.4.1 Audio-visual dataset .......................................... 89
  - A.4.2 Results and Discussion ....................................... 89
- A.5 Concluding Remarks ................................................. 93

## Bibliography ............................................................. 96
LIST OF FIGURES

Figure 2.1: A Bayesian model in the driving context: left hand side of the equation represents the probability that the point \( z \) is attended to given the feature value \( f_z \) and location \( l_z \) of the point. ................................. 11

Figure 2.2: Illustration of traffic pattern around ego-vehicle during safe driving: the detected vehicles (Top), the motion vectors (middle) and the trajectories of the tracked vehicles. ...................................................... 12

Figure 2.3: An example of a coarse gaze direction projected onto the image plane of the looking-out camera. .......................................................... 14

Figure 3.1: Head movements during a merge event. 3D model of a head illustrates observed facial feature from a fixed camera perspective and self-occlusion induced by large head movements. ................................. 19

Figure 3.2: 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 ............................................. 24

Figure 3.3: A block diagram of our proposed approach. It’s a two step approach: extracting features from an image and using corresponding features points on a 3D generic-face model to estimate head pose ............................................. 25

Figure 3.4: Illustration of the online learning process of estimating restricted face region in the image plane. .......................................................... 28

Figure 3.5: 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. .......... 28

Figure 3.6: Tracked facial feature/landmarks and their correspondences in 3D face image. Solid red circles are the points utilized for the head pose calculation. ............... 31

Figure 3.7: 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. ............................................. 32

Figure 3.8: 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. .......... 33

Figure 3.9: LISA-A experimental testbed equipped with and capable of time synchronized capture of camera array and multiple Inertial Measurement Units (IMUS) [156]. 34

Figure 3.10: (a) Distribution of head pose values in yaw rotation angle for (a) an entire test drive and (b) accumulation of selected events in the same test drive. Clearly, data from chosen events shows a more even distribution across yaw when compared to the entire drive. ............................................. 35

Figure 3.11: 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. .......................................................... 37

Figure 3.12: Error distribution with respect to the true head pose in yaw. The graphs reflects the first three error quartiles respective of true head pose in yaw for single perspective (1st column), 2-camera perspective (2nd column) and 3-camera perspective (3rd column) using CLM+POS and MPS+POS. .......... 38

Figure 3.13: 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. ............... 39
Figure 4.1: The examples of non-ellipse iris shape. ........................................ 43
Figure 4.2: Block diagram of the gaze zone estimation system. ....................... 44
Figure 4.3: The 10 initial estimates of iris centers. ...................................... 46
Figure 4.4: Eye ball image formulation: estimating $\beta$, gaze-angle with respect to head, from $\alpha$, $\theta$, $d_1$ and $d_2$ ......................................................... 48
Figure 4.5: Out-of-bag classification error as a function of number of trees. Dashed blue curve and solid red curve represent errors for static head pose alone and with dynamic head pose features respectively. ..................................... 49
Figure 4.6: LISA-A experimental testbed equipped with and capable of time synchronized capture of looking-in and looking-out cameras. ....................... 50
Figure 4.7: Graphical user interface for annotating gaze zone. A human expert uses all the contextual information (driver and surround views) for ground truth generation. Top portion shows the approximate regions in vehicle frame used for gaze annotation. ......................................................... 51
Figure 4.8: An illustration of patch size 8 (blue), 16 (red), 24 (green), 32 (pink), 40 (black), and 64 (purple). ................................................................. 52
Figure 4.9: Normalized error as a function of iterations for models with different patch sizes. Patch size 40 converges fast and provides better error performance. ................................................................. 52
Figure 4.10: Gaze estimation accuracy Vs time window ($W$ sec) for dynamic feature extraction. After 2 sec time window the improvement in accuracy is negligible. ......................................................... 53
Figure 4.11: Visualization of output by CoHMEt and Zone Estimation: Bottom left: input video feeds with multi-camera perspective; bottom right: Tracked head pose visualization; top left: 3 zones (left, front and right) visualization; top right: zone membership. ......................................................... 56
Figure 4.12: Challenges faced in estimating gaze direction when the driver looks at the speedometer. The top row is driver looking at speedometer. The bottom row is right before driver looks at speedometer. ......................................................... 57
Figure 4.13: Example results on the real-world driver dataset. The top row has five correct results, and the bottom row has five incorrect results. ......................................................... 57

Figure 5.1: Three dashboard images showing examples of the very unconstrained position and orientation the dashboard can have in the field of view. ............... 63
Figure 5.2: Detected lines in the image. Red lines are discarded due to too much of a skew to constitute the dashboard edge. ......................................................... 63
Figure 5.3: Confidence map (top) and its resulting mask (bottom). ..................... 64
Figure 5.4: Examples of pedestrian detection scenarios where the exclusion mask has been overlaid in red. ................................................................. 65
Figure 5.5: Top view diagram of the test bed setup. ........................................ 65
Figure 5.6: An example of annotated sequence from the time synchronized video. ................................................................. 67
Figure 5.7: Normalized error of surrogate gaze estimate on a continuous segment. ................................................................. 68
Figure 5.8: Visualization of the LILO attention result in a sequence where gaze switches from one salient location to other. The red box around the pedestrian illustrates the salient region, the Gaussian kernel with yellow center shows the gaze location in the driver’s field of view. The solid box is the pedestrian. ................................................................. 69
Figure 5.9: ADTF graphical user interface for synchronized data capture, processing and visualization. ................................................................. 72
Figure 5.10: Example of a time trace illustrating the development of attention buffer. ................................................................. 74
Figure 5.11: Controlled testing of Attention Guard system in Candlestick Park, San Francisco, CA. (a) Driver looking away from the forward driving direction and (b) ego-vehicle (white) coming to stop slowly while depending upon lead vehicle behavior. ................................................................. 75
Figure 5.12: Attention Guard: A functional illustration in a controlled experiment in a driving test track. The driver is interacting with infotainment system, not looking toward forward driving direction. The blue screen visualizes the surround context seen by the car.

Figure A.1: An example of spontaneous conversation between driver and passenger during a driving task. Film strip shows samples of five images equally spaced in the utterance. First half of the utterance contains the speech and later half the road noise. Notice, however, that facial features are.

Figure A.2: Overview of the proposed expression recognition system. Cross-Relevance feedback block provides the importance of the other modality at current time interval based on the analysis of its own modality. Frame-Relevance measure block can potentially utilize both Cross-Relevance feedback and.

Figure A.3: Reference shape derived from the database showing 66 landmark positions along with the ones in red which are used during image alignment process.

Figure A.4: (a) An example of tracked face and the landmarks, and (b) aligned face image obtained using reference shape during image alignment step.

Figure A.5: Signal processing involved in calculation of single image representation of the image sequence. The bottom curves are intensity (dotted green - associated with the left axis) and pitch contour (solid blue associated with right axis) along with voiced region depicted by the red cross. The middle plot is.

Figure A.6: Binary-class classification accuracy for all the possible 15 different combinations over six basic emotions: happy (ha), Sad (sa), Surprise (Su), Fear (Fe), Anger (An) and Disgust (Di). (a) Subject dependent analysis, average accuracy of the ten experiments corresponding to each 10-fold cross-validation.
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Selected studies on vision based head pose and dynamics estimation systems</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>which are already tested or have potential to work in automobile environment.</td>
<td></td>
</tr>
<tr>
<td>3.2</td>
<td>A list of events considered for evaluation, and its respective count and</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>number of frames.</td>
<td></td>
</tr>
<tr>
<td>3.3</td>
<td>On-road performance evaluations of the proposed CoHMET.</td>
<td>37</td>
</tr>
<tr>
<td>4.1</td>
<td>Static Feature: Confusion matrix for 8 gaze zones numbered as shown in</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>Figure 4.7.</td>
<td></td>
</tr>
<tr>
<td>4.2</td>
<td>Dynamic Feature: Confusion matrix for 8 gaze zones numbered as shown in</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>Figure 4.7.</td>
<td></td>
</tr>
<tr>
<td>4.3</td>
<td>HEIDy feature: Confusion matrix for 6 gaze zones numbered as shown in</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>Figure 4.7.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Notice the performance improvement in speedometer and other zones.</td>
<td></td>
</tr>
<tr>
<td>5.1</td>
<td>Dashboard masking cuts the false positive rate in half, without impacting</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>the detection performance too much.</td>
<td></td>
</tr>
<tr>
<td>5.2</td>
<td>Performance of the attention estimator. Pedestrian accuracy shows how many</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>pedestrian bounding boxes are correctly determined to be the attention</td>
<td></td>
</tr>
<tr>
<td></td>
<td>point for the driver. Mean and median error are measures of how far the</td>
<td></td>
</tr>
<tr>
<td></td>
<td>gaze-surrogate point is from the correct pedestrian bounding box.</td>
<td></td>
</tr>
<tr>
<td>A.1</td>
<td>Multi-class classification accuracy (in %): Confusion matrix for</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>leave-one-subject-out cross validation</td>
<td></td>
</tr>
<tr>
<td>A.2</td>
<td>Multi-class classification accuracy (in %): Confusion table for randomized</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>10 fold cross validation</td>
<td></td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

My deepest gratitude is to my advisor, Prof Mohan M. Trivedi, for his immense support, motivation, guidance and wisdom. In his association over past several years, I have come to know the greater purpose of research beyond personal intellectual gratification. The motivation to improve the lives of human being and the society at large was not this strong in the back of my mind when I started the graduate study. I have learned new meanings of dedication, enthusiasm, inspiration and perseverance from him. They all are invaluable to me.

I would like to thank my committee members, Prof. Serge Belongie, Prof. Gary Cottrell, Prof Edwin Hutchins, Prof. Bhaskar Rao, and Prof. Nuno Vasconcelos, for their time, expertise and guidance, which have significantly contributed to this work.

I thank my colleagues and friends in the CVRR laboratory for their constant support, encouragement, stimulating discussions and their precious time they devoted towards my pursuit. I’d especially like to acknowledge the collaborations and efforts of my co-authors, Dr. Cuong Tran, Dr. Sayanan Sivaraman, Sujitha Martin, Eshed Ohn-Bar, Andreas Mogelmose, Kuo Chen, and Prof. Trivedi.

I am thankful to my sponsors, the Powell Graduate Fellowship, the UC Discovery Program, the National Science Foundation, Audi and Volkswagen Electronics Research Laboratory.

Thanks to all my wonderful friends who brought much needed laughter and comfort throughout these long endeavoring years. Without your friendship, I would not have gone this far, at least not without losing my sanity.

Finally, Thanks to my family. I cannot be enough grateful for their care, understanding, and patience;
Specially to my parents, for their unconditional love and support;
To my brother who essentially set me on this path long long ago even before I could understand what career I wanted;
To my sisters whose support kept me going;
And last but not the least, to my love and fiancée whose sweet entrance to my life made this last phase of my Ph.D. even smoother and more rewarding.

Publication acknowledgements: Chapter 2 is in part a reprint of material currently being prepared for submission for publication to the IEEE Transactions on Intelligent Transportation Systems. Chapter 3 is in full a reprint of materials that are published in the IEEE Transactions on Intelligent Transportation Systems (2014), by Ashish Tawari, Sujitha Martin, and Mohan M. Trivedi, and in the IEEE international Conference on Image Analysis and Processing (2013), by Ashish Tawari and Mohan M. Trivedi. Chapter 4 is in part a reprint of materials that are published in the IEEE Intelligent Transportation Systems Conference (2014), by Ashish Tawari, Kuo Chen and Mohan M. Trivedi, and the IEEE Intelligent Vehicles Symposium (2014), by Ashish Tawari and Mohan M. Trivedi. Chapter 5 is in part a reprint of materials that are published
VITA

2006 B. Tech. in Electrical Engineering, Indian Institute of Technology, Bombay, India

2008-2014 Graduate Student Researcher, University of California, San Diego

2014 Ph. D. in Electrical Engineering (Intelligent Systems, Robotics, and Control), University of California, San Diego

PUBLICATIONS


Ashish Tawari and Mohan M. Trivedi, ”Robust and Continuous Estimation of Driver Gaze Zone by Dynamic Analysis of Multiple Face Videos,” IEEE Intelligent Vehicles Symposium, June 2014.


ABSTRACT OF THE DISSERTATION

Visual Attention from Dynamic Analysis of Head, Eyes and Salient Objects

by

Ashish Tawari

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

University of California, San Diego, 2014

Professor Mohan M. Trivedi, Chair

Face of a person conveys a wealth of information about his/her attentive state. Particularly, head and eyes have the potential to derive where and at what the person is looking. Since humans primarily attend to objects of interest, knowledge of salient objects in the surrounding region can help to accurately infer the focus of visual attention of the person. We present novel computational frameworks and systems to infer visual attention by analyzing dynamics of head, eyes and salient objects. We evaluate proposed systems in intelligent automobile spaces with an emphasis on accurate, robust and continuous performance in the naturalistic driving conditions.
Chapter 1

Introduction

In this age of technology, intelligent systems and smart devices are becoming more and more an integral part of the human’s daily life. Advancement in computation, both in terms of cost (getting lower) and speed (getting faster), has enabled new sensing technologies and networks to observe, understand, and interact with humans and their environment. Past several decades have witnessed significant progress in number of areas including communication, medicine, manufacturing to name a few to improve human society’s quality of living, safety and productivity.

One area where technology has still large unrealized potential is that of transportation and intelligent vehicles. A Global status report on road safety by the World Health Organization (WHO) [118] indicates that worldwide the total number of road traffic deaths remains unacceptably high at 1.24 million per year. In fact, there were 5.6 million police-reported motor vehicle crashes in the United States along in 2012, with over 33,000 fatalities. Driving task demands careful interaction of driver-vehicle-environment “ecosystem”. Any weak element of this system can lead to unpleasant consequences.

Recent survey estimates that driver error is the critical reason for over 90% of the crashes [1]. This alludes ‘driver’ as a weak link. However, driver remains center of the driving task due to the enhanced and unique abilities of humans to safely react to novel unseen situations. Even during certain ‘autonomous’ driving mode such as adaptive-cruise-control mode, the driver is in-charge and responsible for the action taken by the vehicle, in a legal sense, and is required to take over the control, if needed. Hence, it’s important to understand what are the reasons for driver errors in order to address and mitigate their adverse effect.

Among different reasons attributed to the driver error, driver distraction (e.g phone usage, talking, eating), inattention (drowsiness, fatigue etc.) and inadequate surveillance are some of the prominent ones [1]. In fact, with more and more devices whether as parts of in-vehicle infotainment systems or other nomadic devices such as smartphones, GPS navigation system
etc., occupying space in the car cockpit, drivers are presented with increasing opportunities to be distracted. Although automakers are careful in designing in-vehicle systems keeping in mind safety concerns, they have no or very little control over other device usage by the driver. There exist laws banning certain distracting activities such as handheld cellphone usage, texting etc., but their enforcement are often not strict. Moreover, there are other naturally occurring distractions, e.g. checking speedometer at ‘wrong’ time, parent looking back to his/her child, minds wondering off, or sometimes looking in the ‘wrong’ direction etc. Hence, simply relying on mandatory laws is not sufficient, but it is important to assist driver with attention technologies to avoid unpleasant consequences of distracting activities.

Surveys on automotive collisions [128, 17] demonstrated that drivers were less likely (30%-43%) to cause an injury-related collision when they had one or more passengers who could alert him to unseen hazards. Consequently, there is a great potential for human-centric intelligent driver assistance systems (DASs) [162, 161, 65, 37] to alert or even guide [146] the driver briefly through the critical situations and improve road safety.

Even as human continues to be the the center of the driving activity, most of the DASs, specially those that are commercially deployed, ignore driver in the loop. They often just rely on the data from a number of embedded sensors observing road obstacles, ego vehicle’s state such lane position/deviation etc. to assess driver distraction and alert the driver of hazardous situation. Prior research suggest that systems that do not take into account the driver’s state, often produce a large percentage of alerts that are either unnecessary (where situation is judged dangerous by the system but not by the driver) or false (where situation does not correspond to a threat at all). High rate of such alerts would both annoy drivers and degrade their confidence that the system provides anything but nuisance alerts. This is a very important factor in human assistance systems design, as to reduce false positives for driver acceptance and trust of the system. This can be achieved by adapting to the driver’s current state as well as his/her preferences.

Therefore, there exist an window of opportunity to not only improve the system’s effectiveness to convey information to the driver properly and at the right time but also improve its ability and the scope of operation. Prior research have shown that driver monitoring using smart camera setup can improve early recognition of driver distraction and inattention. Cues such as driver head dynamics and eye gaze have been shown to reveal driver distraction and inattention. Monitoring driver behavior is, hence, becoming an increasingly important component of an Advanced Driver Assistance System (ADAS).

1.1 Contributions and Outline

In this dissertation, we present computational algorithms and systems to infer focus of visual attention of a human agent, in our case, the driver. Significant amount of work pertains
to improving state-of-the-art driver monitoring systems by novel analysis of the driver’s head and eye region. First, We layout realistic design criteria needed for the systems to work in a challenging outdoor environment such as driving. Our analyses show that many existing head pose tracking systems may demonstrate good performance for 95% of the time but they tend to fail for those 5% of the time when it is needed the most - specially when head moves fast and have large deviation from the frontal pose. We introduce novel metrics including success rate to assess the systems against the set realistic requirements. For quantitative evaluation, we acquired a novel database focused on driving events and maneuvers targeting large head movement away from the front driving direction during lane changes, turns etc. We introduce a distributed camera framework which showed significant improvement compared to existing systems in terms of accuracy, operational range as well as success rate. Our analyses show the need for such framework for successful real-world deployment of driver monitoring systems.

To study driver’s visual attention and conversely visual distraction, an important component is to infer the driver’s gaze direction. Towards this end, we develop computational framework for coarse (head) to fine (eye) level of analysis. Prior studies [183] on driver simulator have shown that even though precise eye gaze can better explain driver distraction, the head pose still provide a good estimate of the distraction level. Indeed, it’s the robustness requirement of the ADASs that have made head pose based driver monitoring systems popular. Underlying assumption is that head pose provides coarse and robust approximation of the precise eye gaze. Our first goal is to improve the coarse gaze estimation using ‘robust’ cues. Literature suggest that any visual search, requires a careful coordination between head and eye movements to ‘optimally’ guide through the search space [39, 177]. We hypothesize that how one arrives to a particular head pose, can predict how eye gaze might have moved.

Above finding led us to investigate the role of both, static head pose and head pose dynamics, for gaze zone estimation. Gaze zone estimation corresponds to inferring the driver’s eye gaze position to zones relevant to the driving tasks such as front, left/right side mirrors, speedometer, center console, etc. We find that dynamic information does provide significant improvement over static head pose. We, however, also find that the head pose alone based system is inherently limited specially for disambiguating between zones separated by subtle eye gaze differences. Hence, we propose gaze-surrogate features estimated from eye region via eyelid and iris analysis, keeping in mind robustness requirement. We develop a Bayesian framework for contexts/tasks based salient object analysis from surround perception and incorporate gaze direction to infer driver’s focus of attention and attentional target. We develop and implement real-time distraction detection algorithm incorporating surround situation criticality as well as driver’s preferences and the attentive state.

In our further pursuit of understanding driver state, we study affective state of a human agent. Driving in particular presents a context in which a user’s emotional state plays a significant role. Emotions have been found to affect cognitive style and performance. Even mildly positive
feeling can have a profound effect on the flexibility and efficiency of thinking and problem solving. Towards this end, we present a novel facial expression recognition system exploiting cross-modal data (audio) association. The ultimate goal in all of this work is to move towards better intelligent systems for human interaction and safety.

Among the contributions of this dissertation:

- Focus of visual attention (FoViA) estimation by joint analysis of driver and surround observation. (Ch. 2)
- Literature review of potential head pose based driver monitoring systems and a distributed camera framework for continuous head movement estimation. (Ch. 3)
- Systematic cost-benefit analysis of various camera configurations to maximize system performance. (Ch. 3)
- Robust iris detection algorithm - a iterative regression framework. (Ch. 4)
- Novel gaze surrogate measures via eyelid and iris analysis (Ch. 4)
- Gaze zone estimation system from head, eye and iris dynamic (HEIDy) features. (Ch. 4)
- Distraction detection algorithm for lane keep and brake assistance. (Ch. 5)
- Attended target detection algorithm. (Ch. 5)
- Novel facial expression recognition system exploiting cross-modal (audio) data association. (Appendix A)

The following chapter provides Bayesian formulation to estimate focus of visual attention in a looking-in-looking-out framework. We discuss the probabilistic formulation in the context of driving and raise important research questions for vision systems, particularly related to driver monitoring. Chapter 3 provides a literature review of vision based head pose and dynamics estimation systems which are already tested or have potential to work in an automobile environment. It further describes a novel head pose dataset collected from naturalistic on-road driving conditions in urban streets and freeways with the particular emphasis on events inducing spatially large head movements (e.g. merge, lane change etc.). It then details the distributed camera framework and performs a comprehensive comparative study of various state-of-the art facial landmark approaches for head pose estimation and evaluates the proposed system against the head pose dataset. We present two such solutions that additionally exploit constraints present in the driving context and video data to improve tracking accuracy as well as computation time. Furthermore, we conduct a thorough comparative study with different camera configurations.

Chapter 4 describes gaze zone estimation approach based on head, eye and iris dynamic (HEIDy) features. It presents an iris center detection algorithm using shape-free, local-patch
based regression model. It introduces gaze-surrogate features and investigate the efficacy of head pose and eye cues to determine driver gaze to different zones, which are relevant to the driving tasks. It further presents a systematic study to understand the incremental benefits of using head pose, its dynamics and eyes cues in estimating gaze zone. Chapter 5 presents an integrated looking-in (driver state) and looking-out (traffic state) analysis. In particular, it describes two systems: attended target detection system and distraction detection system. Finally, Chapter 6 provides concluding remarks.
Chapter 2

Focus of Visual Attention (FoViA) Estimation by Simultaneous Analysis of Viewer and View

2.1 Introduction

Visual stream of data is constantly bombarding our eyes ($10^8 - 10^9$ bits/second) [80, 68]. To make sense from this vast amount of data in real-time requires mechanisms to cleverly select important and relevant data for further processing. Visual attention, in general, refers to factors that influence such selection mechanisms.

In the context of driving, understanding driver attention or conversely, inattention or distraction is an important factor in improving road safety. To know at what or where the driver is looking and is attentive of is a key factor to the design of futuristic ADASs in terms of both comfort and safety. The knowledge that the driver has attended to a particular vehicle, pedestrian, traffic sign, temporary warning etc. can enable an ADAS to appropriately warn the driver without annoying him or her. If the driver is distracted from the forward driving direction, an appropriate alert can be given or the ego-vehicle may even be guided momentarily through the dangerous situation to avoid unpleasant consequences. It is important, therefore, for such a system to maintain a holistic understanding of the driver-vehicle-environment “ecosystem”.

In this chapter, we provide a Bayesian framework to model focus of visual attention (FoViA) which, in a principled way, combines both the view (the traffic state) and the viewer
(the driver state). The framework can also incorporate task information which is known to influence the selection of the focus of attention [107, 57, 178]. We discuss different components of the proposed framework, in particular, how they can be learned and inferred in the context of driving. In this process, we will raise several relevant research questions which will form the basis of the following chapters to develop accurate and robust algorithms and approaches for ADASs. In particular, we will discuss in great detail about issues, challenges and solutions related to understanding the driver’s attentive state by direct observation of the driver. In Chapter 5, we will explore different applications pertaining to the visual attention for driver assistance by joint analysis of the traffic and driver state.

At this point, we would like to emphasize that although we keep our focus in the driving context, developed algorithms and frameworks are general in their nature and can be deployed in other human-machine interfaces as well.

### 2.2 Related Research

Modeling visual attention has been a very active research area for the last several decades. Many different models exist that have made theoretical contributions as well as demonstrated practical applications in computer vision, graphics and robotics to name a few among other areas [90, 167, 106, 33]. We discuss some relevant and select literature that bring important concepts which are applicable and useful to the driving context. For a review on computational models of visual attention in general, we encourage readers to refer to [20, 49, 21, 16].

Literature suggests that any visual search, for the most part, is presumed to be guided by some combination of a goal- (top-down) and stimulus-driven (bottom-up) mechanisms, depending on the situation. Bottom-up cues are mainly based on characteristics of a visual scene (stimulus-driven)[69], whereas top-down cues (goal-driven) are determined by cognitive phenomena like knowledge [44, 180, 43], expectation [133, 8], reward [61, 135], and current goal [178, 107, 58].

Bottom-up saliency-based attention has been studied extensively [79, 70]. This attentional mechanism is also called exogenous, automatic, reflexive and is considered fast and involuntary. Saliency intuitively characterizes some parts or events in the visual field that stand out from their neighboring background. For example, flashing lights are salient, as are objects that pop out against their background (i.e. high relative contrast), such as retroreflective traffic signs, on set of a brake light from a leading vehicle etc. Salience of visual objects or events, however, can vary as a function of their locations on the retina. It is well known that retina has a high resolution central fovea and a low-resolution periphery. There are psychophysical studies that have documented detection thresholds for different types of stimuli at different degrees of eccentricity [6, 98]. Walker et al. [127] incorporate foveated representation with reducing resolution as distance increases from the center.

Top-down attention, on the other hand, is task-driven and is considered slow and involunt-
In a famous experiment, Yarbus [178] asked subjects to watch the same scene (a room with a family and an unexpected visitor entering the room) under different tasks, such as determine the ages of the people or just freely observer the scene etc., and found considerable differences in eye movements and fixations. What it reveals is that seeing is inextricably linked to the observer’s cognitive goals. For instance, subjects performing complex tasks such as making tea or sandwich, playing cricket, walking etc. were found to direct a majority of fixations toward task-relevant locations [57]. In the context of driving, drivers often look to forward roadway to gather relevant information for safe driving (e.g. while vehicle following) or look over their shoulder during lane change maneuver. task. Sodhi et al. [141] studied how secondary tasks while driving, such as adjusting the radio or answering a phone, affect eye movements.

The prevailing view is that bottom-up and top-down attentions are combined to direct our attentional behavior. SEEV, a model of scanning [171], incorporates bottom-up and top-down influences in four components (whose initials derive the name of the model): Salience & Effort (bottom-up) and Expectancy & Value (top-down). Salience and Value (importance or relevance of an information source to the task at hand) are as discussed above. Effort is an inhibitory component that discourages longer scans in time or space [135], e.g. moving attention between two physically far apart locations such as between forward roadway and radio console. Expectancy component accounts for the tendency of an observer to look more frequently towards information sources with higher bandwidth (or event rate), given the sources are relevant to the task [133]. The goal of such model is to predict how an observer or an operator would allocate his or her visual attention to different areas of interest e.g. estimate amount and distribution of time that a driver is looking away from forward roadway. For this, the model simply adds the four factors with appropriate weights and signs. Determining these weights, however, is a considerable challenge since the units of measurements of the four components differs significantly from each other. Often heuristics or some ‘common-sense’ logics are employed e.g. coding their values in ordinal scale based upon their ‘known’ importance for a particular application.

Zhang et al. [184] incorporate bottom-up and top-down influences in a Bayesian framework. In this framework, bottom-up saliency emerges naturally as the self-information of visual feature which is similar to other works as Olivia et al. [116], Torralba [159] and Bruce and Tsotsos [18]. The overall model incorporates the top down influences in terms of the knowledge of the target’s appearance or prior locations. Although the top-down knowledge can account for task specific information implicitly, the influence of task dependence is not explicit. In driving, there exist different concurrent tasks and we need to incorporate the task influences in a principled manner. Moreover, none of the above studies as such incorporate state of the viewer itself.

Inspired from above studies, we present a Bayesian model to infer the focus of visual attention (FoViA) by simultaneous observation of the view and the viewer. It differs from prior work in: (1) it incorporates tasks’ influences by explicitly modeling bottom-up and top-down probability density conditioned upon the tasks and (2) it learns and infer these components by
observing view (the surround traffic) and the viewer (the driver) both. We call this as looking-in looking-out framework because it requires observations from outside of the vehicle as well as from inside the cockpit. This is discussed next.

2.3 FoViA: Lookin-in Looking-out Framework

By analyzing salient regions in the field of view (bottom-up attention), one can estimate the focus of attention of the driver. However, in a complex environment such as driving, it is hard to say precisely where or at what the driver is looking, since eye fixations are often governed by goal-driven mechanisms (top-down attention model). We propose to estimate the driver’s focus of attention by simultaneously observing the driver and the driver’s field of view. In the following sections, we first describe the overall Bayesian model and then discuss how to learn and infer different components.

2.3.1 Proposed Bayesian Model

Let $z$ denote a point in the visual field. A point here is loosely defined. It can be associated with a space (e.g. a pixel location in the looking out camera) or other things, such as an object. We let the binary variable $a_z$ represent whether or not a point is attended to. Let $L$ denote the location space and $l_z$ corresponds to the location of the point $z$, so $l_z \in L$. Let $A^t_z$ be the overall saliency measure of the point $z$ at time $t$. It represents the probability of attending to the point $z$ given its feature $f_z$ and location $l_z$ as in Equation 2.1:

$$A^t_z = p^t_z(a_z = 1|f_z, l_z) \quad (2.1)$$

For the purpose of this work, let $T$ be a discrete random variable drawn from the space of all tasks/contexts, $T \in \mathcal{T} = \{T_1, T_2, \ldots, T_n\}$. Then,

$$A^t_z = \sum_{T_i} p^t_z(a_z = 1|f_z, l_z, T_i) \quad (2.2)$$

$$= \sum_{T_i} p^t_z(a_z = 1|f_z, l_z, T_i) P^t(T_i) \quad (2.3)$$

Let’s first further simplify the first component of the right-hand side of Equation 2.3 to understand the effect of bottom-up and top-down processes. For notational simplicity, we drop the superscript $t$, corresponding to time and subscript $z$ referring to the point. Using Bayes’ rule:
\[
p(a|f, l, T_i) = \frac{p(f, l|a, T_i)p(a|T_i)}{p(f, l|T_i)} \quad (2.4)
\]
\[
= \frac{p(f|a, l, T_i)p(l|a, T_i)p(a|T_i)}{p(f|l, T_i)p(l|T_i)} \quad (2.5)
\]
\[
= \frac{1}{p(f|l, T_i)}p(f|a, l, T_i)p(l|a, T_i)p(a|T_i) \quad (2.6)
\]
\[
= \frac{1}{p(f|l, T_i)} p(f|a, l, T_i) p(a|l, T_i) \quad (2.7)
\]

First term on the right side of equation 2.7 represents bottom-up influence as it only depends on the environment’s features observed at the point. The other two terms, on the other hand, involve some form of target knowledge as they favor the features and locations that are previously attended for a given task and hence are referred to as top-down factors.

The overall attention of the driver is not an instantaneous phenomenon but a temporal one. We define total attention (TA) at time \( t \) as:

\[
AT^t_z = \beta \sum_{k=0}^{\infty} (1 - \beta)^k A^t_{z-k} \quad (2.8)
\]
\[
AT^t_z = A^t_z \ast D(t; \beta) \quad (2.9)
\]

where, \( D(t; \beta) = \beta(1 - \beta)^t \) is a weighted average filter as a function of time step.

Next, we discuss how to learn and infer different components (looking-in and looking-out, see Figure 2.1) and what they signify in the context of driving.

### 2.3.2 Looking-out via Naturalistic Driving Data

We call this component as looking-out since it involves analysis of features from the environment. Looking-out component has both bottom-up and top-down factors as shown in Figure 2.1.

**Looking-out bottom-up factor:** \( p(f_z|l_z, T_i) \)

This factor refers to the probability of observing \( f_z \) given it’s location \( l_z \) and the tasks. Notice that the term, in Equation 2.7, favors occurrence of rare features as salient. In other words, if the probability of a feature \( f \) at point \( z \) is lower, it is more salient and it may attract the attention as abnormal event. For example, in the driving context, for safe driving there are expected position and motion pattern of surrounding objects and if any road user deviates e.g. unusually fast moving vehicle overtaking or cutting the ego vehicle, it will draw the notice of the
Figure 2.1: A Bayesian model in the driving context: left hand side of the equation represents the probability that the point \( z \) is attended given the feature value \( f_z \) and location \( l_z \) of the point. Right hand side of the equation has three factors: (1) looking-out bottom-up models the traffic state and emphasizes rare (or abnormal) events as salient, (2) looking-out top-down learns about the targets’ features that a driver pay attention to and reflects the driving style or the preference of the driver and (3) looking-in factor represents the probability of attending to the point \( z \) given no other information except its location \( l_z \). This is inferred by observing the driver and it incorporates the driver state. We call this the ‘saliency-filter’.

driver. In other words, this term accounts for any environmental distraction that can attract the attention of the driver.

To learn this bottom-up saliency, there are two key factors. One is the feature space, and the other is the probability distribution over the features. In the driving context, we propose to use motion features of surrounding objects. Figure 2.2 illustrates an example of trajectories which can describe how different vehicles in a freeway move through a visual scene. Notice that we work here at object levels i.e. \( z \) refers to an object such as a vehicle. This requires first the detection and tracking of objects, estimating the flow vector (position and velocity) and then an trajectory refers to a sequence of a flow vectors. This is applicable to other road users (e.g. pedestrian, bicyclist etc.) at other scenarios (e.g. urban streets, intersections, etc.) too. Object detection and tracking is beyond the scope of this work. We encourage reader to refer to Sivaraman and Trivedi [139]. For trajectory learning and anomaly (abnormal event) detection, please refer to [99, 100]. In learning the probability density, the assumption is that for most of the time driver drives safely, hence an anomalous or an abnormal event will be treated as outlier to the learned probability distribution.
Figure 2.2: Illustration of traffic pattern around ego-vehicle during safe driving: the detected vehicles (Top), the motion vectors (middle) and the trajectories of the tracked vehicles.
Looking-out top-down factor: $p(f_z/a_z, l_z, T_i)$

This factor refers to the probability of observing $f_z$ at its location $l_z$ given the task and the knowledge that the point $z$ is attended to. Notice that this term, in Equation 2.7, favors features values that are consistent with the knowledge of previously attended points. In other words, features that had attracted earlier will likely to draw the attention, i.e. will have higher probability value, when presented in the visual field again as important and relevant events. For example, in the driving context, if a lead vehicle brakes (not necessarily hard brake) can draw the attention of the driver, e.g. to keep a safe following distance, and hence is important salient event. This term then can be called to be learning driver preferences and style because in learning these features are conditioned upon the driver’s attentive state. The challenge, however, is how to know that the driver is attentive of an event or an object. One can ask the driver itself but this can be cumbersome or may even interfere with naturalistic driving behavior. Another way is to infer this from the action taken by the driver. Going back to our example, if the lead vehicle brakes then to maintain the following distance, driver may release the throttle pedal or even brake. This suggests that the driver was attentive of the lead vehicle’s braking event. In Chapter 5, we will see how these learned behaviors can be used for a ADAS which is conscious of the driver’s style. Several different modeling techniques could be used to model the driver’s style and preferences such as the autoregressive exogenous (PWARX) model [157], the Gaussian mixture model (GMM) [5], neural network [145], [117], and the hidden Markov model [13].

2.3.3 Looking-in via Driver Observation: $p(a_z/l_z, T_i)$

This factor represents the probability of attending to the point $z$ given no other information except its location $l_z$ for a given task $T_i$. If there exists no observation of the viewer (the driver), in general, this is assumed to be uniformly distributed (location invariance or non-informative prior assumption) or biased towards a particular direction with the prior knowledge associated with a specific task [38]. However, if direct observation of the driver is available, this can be inferred by the evidence rather than any assumption. The effect of this term is to weight the visual field by understanding the driver state, particularly the gaze direction. Figure 2.3 illustrates an example of such an effect. We refer to this as the saliency filter as it prunes the possibility of being attentive at a particular location in the visual field by knowledge the gaze direction.

Let $l_g$ be the gaze location of the driver and is drawn from the location space $L$. We model the saliency filter $p(a_z/l_z, T_i)$, as a Gaussian density function in $l_z$. To be specific, for 2-D location space $L$ (e.g. image pixel coordinate as in Figure 2.3)

$$p(a_z/l_z, T_i) = \frac{1}{\sqrt{(2\pi\sigma^2_{T_i})^2}} exp\left\{ -\frac{\|l_z - l_g\|^2}{2\sigma^2_{T_i}} \right\}$$

(2.10)
The variance $\sigma_{Ti}$ controls the coupling of attention to the gaze location. A higher value of $\sigma_{Ti}$ indicates loose coupling and vice versa. Varied levels of coupling between gaze location and attention is evident from the well-known phenomena of covert and overt attention [124]. Note that we modify the variance of the Gaussian to be task dependent $\sigma_{Ti}$. For example, when driver is turning around a curve, he or she actively looks to the inner edge (or more precisely the tangent point) [14] and hence suggest a tight coupling; on the other hand, for maintaining lane position, drivers may use peripheral vision and that they learn its use over time [144]. However, separation between gaze location (more precisely, fixation) and attention is often observed in simple tasks [124] or tasks that are learned and hence become easier over time [144]. Furthermore, it is not known yet how to measure covert attention. For our purposes, to evaluate the vision techniques, we will implicitly assume that attention is where the gaze is.

The question pertaining to computer vision is how do we estimate the gaze location $l_g$. In the ever changing and dynamic environment like driving, precise eye gaze estimation is a very challenging task or at times even not possible, for example, due to the occlusion caused by a sunglass. We will further motivate this in the following chapters, but here it suffices to say that instead, we estimate an approximate or coarse gaze location (say $\hat{l}_g$) for the robustness purposes. We call it as the gaze surrogate measure.

To deal with uncertainty in the gaze measurement, we take the output $\hat{l}_g$ of a vision-based gaze estimator and model uncertainty in the estimate with the Gaussian distribution:

**Figure 2.3:** An example of a coarse gaze direction projected onto the image plane of the looking-out camera.
where, $\sigma_m$ accounts for the measurement uncertainty. Then, it can be shown that saliency filter is given by Equation 2.12 as below:

$$p(l_g) = \frac{1}{\sqrt{(2\pi\sigma_n^2)^2}} \exp \left\{ -\frac{\|l_g - l_{\hat{g}}\|^2}{2\sigma_n^2} \right\}$$

where, $\hat{\sigma}_{T_i}^2 = \sigma_{T_i}^2 + \sigma_m^2$. Our purpose is to develop vision techniques to estimate surrogate measure. This is the subject matter of the following two chapters, where we discuss in details, proposing novel measures as well as algorithms by analyzing driver’s head and eye region.

### 2.4 Concluding Remarks

In this chapter, we proposed a probabilistic approach to estimate focus of visual attention by incorporating bottom-up and top-down factors as well as task influences. We presented looking-in looking-out framework to learn different components from naturalistic driving data. The looking-out bottom-up factor models the traffic-state and the looking-out top-down factor incorporates driver preferences and style. We introduced a ‘saliency-filter’ factor which models driver states by direct observation via looking-in vision sensors.

Our focus remains in the looking-in component as this is often ignored in current driver assistance systems. We firmly belief that futuristic ADASs would greatly benefit by incorporating driver behavior and adapting to the driver preferences and style. Particularly, the knowledge of gaze location/direction is a key to understanding attentive state of the driver. Precise gaze location estimation still remains a challenging problem specially in outdoor-settings like driving and at times it’s even infeasible due to the occlusion such as by eyewears e.g. sunglasses. We instead suggest to utilize gaze surrogate measure that can be estimated robustly and hopefully, it provides a good approximation to true gaze location/direction. There are several research questions: (1) what are such surrogate measures? (2) How can they be estimated robustly in real-world driving conditions and (3) what are realistic design requirements for such systems? (4) What if eye information is accessible by the vision sensor, can a better surrogate measure be conceived? (5) If yes, what is the incremental benefit? Next, two chapters provide a systematic study answering these questions.
Acknowledgements

Chapter 2 is in part a reprint of material currently being prepared for submission for publication to the IEEE Transactions on Intelligent Transportation Systems, by Ashish Tawari and Mohan M. Trivedi. The dissertation author is the primary investigator and author of the paper.
Chapter 3

Continuous Head Movement Estimator: A Distributed Camera Framework

3.1 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. Robustness requirement of ADAS has suggested the use of head dynamics. While precise gaze direction provides useful information, head pose and dynamics provide course gaze direction, which are often sufficient in a number of applications [84, 39]. Recent studies have used head motion along with lane position and vehicle dynamics to predict a driver’s intent to turn [22] and intent to change lanes [97]. 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 [40]. Significant amount of research has gone towards fatigue and attention monitoring using driver head dynamics [7, 12]. 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 different situations [9].

Automatic head dynamic analysis remains a challenging vision problem. Not only should head movement analyzer be robust to ever changing driving situation but it also needs to be continuously functional in a non-selective manner to gain driver’s trust. Specifically, such a
system 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 head pose relative to some initial configuration.

- **Fast:** The system must be able to estimate head pose 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).

Many state-of-the-art vision based head pose algorithms have taken the necessary steps to be automatic, fast and person invariant [102]. These systems have shown good performance 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 [94]. 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 3.1 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° 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 [95, 156]. Multi-camera systems exist in many other applications such as gesture recognition [66, 160], human body pose and activity recognition [59], 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
Figure 3.1: Head movements during a merge event. 3D model of a head illustrates observed facial feature from a fixed camera perspective and self-occlusion induced by large head movements.

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.

3.2 Related Research

Naturalistic driving presents unique challenges for vision based head dynamic estimation
Table 3.1: Selected studies on vision based head pose and dynamics estimation systems which are already tested or have potential to work in automobile environment.

<table>
<thead>
<tr>
<th>Research study</th>
<th>Objective</th>
<th>Methodology</th>
<th>Evaluations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhu and Ji ’04 [187]</td>
<td>Head pose estimation from uncalibrated monocular camera</td>
<td>Texture, IR Single Continuous</td>
<td>Operation 10fps Real-time Dataset In lab Metrics Unspecified</td>
</tr>
<tr>
<td>Guo et al. ’06 [55]</td>
<td>head pose estimation for driver surveillance</td>
<td>Texture Single 225 discrete poses</td>
<td>Real-time In-lab CCR</td>
</tr>
<tr>
<td>Wu and Trivedi ’08 [173]</td>
<td>Two-stage, coarse and fine, head pose estimation</td>
<td>hybrid Single 86 discrete poses</td>
<td>Not real-time In lab</td>
</tr>
<tr>
<td>Murphy-Chutorian et Trivedi ’10 [105]</td>
<td>Head pose estimation</td>
<td>Texture Single Continuous</td>
<td>Real-time Naturalistic driving Metrics MAE and STD Function of true angle: ME and STD</td>
</tr>
<tr>
<td>Lee et al. ’11 [84]</td>
<td>Gaze-zone estimation using head pose for forward collision warning</td>
<td>Yaw-Shape Pitch-texture Single</td>
<td>Real-time stationary vehicle Metrics RMS and SDF</td>
</tr>
<tr>
<td>Martin et al. ’12 [94]</td>
<td>Head pose estimation</td>
<td>Shape Single Continuous</td>
<td>Real-time Naturalistic driving Metrics STD and FR.</td>
</tr>
<tr>
<td>Proposed CoHMET</td>
<td>Continuous head dynamics estimation for spatially large head movements</td>
<td>Shape Multi Continuous</td>
<td>Real-time Naturalistic driving</td>
</tr>
</tbody>
</table>

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. |
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 [102].

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 [163], ellipsoidal [84] or mean 3D face [94]) to recover head pose. Smith et al. proposed several strategies using global motion and color statistics to detect and track both eyes, lip corners, and the bounding box of the face [140]. Based on these facial features, they estimated head orientation and gaze direction. However, the method cannot always find facial features when the driver wears eyeglasses, makes conversation or partial occlusion e.g. due to hands. Kaminski et al. analyzed the intensity, shape, and size properties to detect the pupils, nose bottom, and pupil glints [74]. Using these detected points along with geometric model of human face and eye, head orientation and gaze direction is estimated. However, the accuracy of the eye location significantly drops in the presence of large head movements, causing degradation in the 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 [113]. Along with these features, the authors used a cylindrical face model to find the driver’s yaw direction. Lee et. al. [84] 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 [50]. 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. [94], 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 one of our proposed
approaches 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 [55], 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. [10], 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 [187] 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 [173]. 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 [103], 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 3.1. 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.

### 3.3 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
Figure 3.2: 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.

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
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 3.2. In Figure 3.2, 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 3.2 show examples of the latter two scenarios with hand movements. Effects of lighting conditions are also highly dependent on the camera location. In Figure 3.2, 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.

3.4 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.

3.4.1 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 3.6. 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 [28, 29] and Pictorial Structure Matching proposed by Felzenszwalb and Huttenlocher [47]. 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 constrains the joint positioning of the landmarks.

A parametrized shape model to capture plausible deformations of landmark locations is given in Equation 3.1. This is also known as a point distribution model (PDM), a term coined by Cootes and Taylor [26]:

\[
p_i = s R_{2D}(\bar{p}_i + \Phi_i q) + t\]

(3.1)

where \(p_i = (x_i, y_i)\) is the 2D location of the \(i^{th}\) landmark in the image \(I\). \(\bar{p}_i\) is the 2D location of the \(i^{th}\) landmark of the mean shape and \(\Phi_i\) encodes the shape variations. Rigid parameters \(\theta_{rg} = \{s, R_{2D}, t\}\) with global scaling \(s\), in-plane rotation \(R_{2D}\) and translation \(t\), along with non-rigid parameters \(q\), represent parameters of the PDM.

Let’s define \(\theta = \{\theta_{rg}, q\}\). 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:

$$L(\theta\{l_i = 1\}_{i=1}^n, I) = p(\{l_i = 1\}_{i=1}^n, I) = \prod_{i=1}^n p(l_i = 1|\theta, I) \tag{3.2}$$

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 $p(l_i = 1|\theta, I)$ of the local detectors are approximated by various models, such as parametric representation - Gaussian density with diagonal covariance [25], full covariance [170], Gaussian mixture model [53], or nonparametric representation - kernel density estimate (KDE) [130]. In our current implementation, we chose KDE for it’s fast convergence property with good tracking ability [130]. It has shown its efficacy in other applications too, such as face expression recognition [155]. 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 $\theta_{rg}$ 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:

$$B^i = \begin{bmatrix} x^i_{min} & y^i_{min} & x^i_{max} & y^i_{max} \end{bmatrix}$$

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

$$H(x, y) = \sum_{B^i} U[x - x^i_{min}]U[y - y^i_{min}]$$

$$- U[x - x^i_{max}]U[y - y^i_{max}]$$

$$M(x, y) = U \left[ \frac{H(x, y)}{\gamma} - \alpha \right] \tag{3.4}$$

$$BR(P) = \{(\min x, \min y, \max x, \max y) \}$$

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
Figure 3.4: Illustration of the online learning process of estimating restricted face region in the image plane.

Figure 3.5: 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.

Rectangle $\text{BR}(\cdot)$ as defined in Eq. 3.5 is calculated using set of points $P = \{ p = (x, y) \mid M(x, y) > 0 \}$. We call this minimum bounding rectangle $B^*$, the restricted face region. Figure 3.4 depicts the overall process. Estimated facial landmark location $p_i$ within $B^*$ is considered admissible. Similarly, the probable size of face is proportional to the size of $B^*$. Finally, the rotation parameter is inferred from the estimated roll angle of the driver’s head. The estimate
value within $\pm 20^0$ 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 3.6.1.

**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 [48]. This is naturally represented by an undirected graph $G = (V, E)$, where the vertices $V = \{v_1, \cdots, v_n\}$ correspond to the $n$ parts, and for each pair of connected parts, there exists an edge $(v_i, v_j) \in E$. 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, $S_A(I_t, p_i, m)$, of placing the $i$th part (i.e. node $v_i$) at location $p_i = (x_i, y_i)$, and the likely deformation, $S_D(I_t, p_i, p_j, m)$, for each pair of connected parts. Optimization proceeds by maximizing over all mixtures:

$$S(I_t, P, m) = \sum_{i=0}^{n_m-1} S_A(I_t, p_i, m) + \sum_{(v_i, v_j) \in E} S_D(I_t, p_i, p_j, m)$$

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

where $\hat{m}_t \in M$ is a subset of all mixtures $M$ and $\hat{P}_t$ is a rectangular region of interest defined by Eq. 3.6 and Eq. 3.8 respectively.

In literature, the appearance similarity $S_A(I_t, p_i, \cdot)$ is modeled in various ways e.g. Gaussian derivative filter response around a point [47], feature based description such as Histogram of Gradient (HoG) [186], Haar-like feature [45] etc. $S_D(I_t, p_i, p_j, \cdot)$ is a distance function e.g. a Mahalanobis distance in some transformed space of $p_i$ and $p_j$. In our implementation, we use a discriminative, max-margin framework [186] to model the two scoring functions. Here, $G$ for each mixture is a tree and the optimization is performed efficiently with dynamic programming [47]. 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.
around the estimate from the previous frame:

\[ \hat{m}_t = m_{t-1}^* + \{-1, 0, 1\} \quad (3.6) \]

\[ m_{t-1}^* = \arg\max_{m \in \{\hat{m}_{t-1}\}} \left[ \max_{p \in \{\hat{P}_{t-1}\}} S(I_t, p, m) \right] \quad (3.7) \]

\[ \hat{P}_t = \{p_i | p_i \in (BR(P_{t-1}^*) + (-b, -b, +b, +b, +b))\} \quad (3.8) \]

where \(m_{t-1}^*\) is the mixture chosen for the previous frame \(I_{t-1}\), \(BR(\cdot)\) is defined in Eq. 3.5 and \(b\) is the border width. Figure 3.5 depicts the overall process. These optimizations decreased the processing time by at least 4 folds.

### 3.4.2 Pose Estimation

Given a 3D model of an object, POS (Pose from Orthography and Scaling) [34] 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:

\[ M_0 M_i \cdot \alpha_i = x_0 x_i \quad i = 1 \cdots N_c \]

\[ M_0 M_i \cdot \alpha_j = y_0 y_i \quad i = 1 \cdots N_c \]

where, \(M_0 M_i\) represents the vector from the reference point on the 3D model \(M_0\) to \(M_i\), \(\alpha\) is the scale factor associated with the weak perspective projection and \(N_c\) is number of 3D – 2D point correspondences. The vectors \(i\) and \(j\) form the first two rows of the rotation matrix and the third row is given by the vector \(k = i \times j\), a cross product. Note, however, that while \(k\) is perpendicular to \(i\) and \(j\), vectors \(i\) and \(j\) 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 3.6 shows these points (red solid circle) on a test image and its corresponding points on the 3D mean face model.
3.4.3 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 Fig. 3.7 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 Fig. 3.9. 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 $(\Omega_{\text{N-}}, \Omega_{\text{N+}})$ 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 3.4.1) is utilized. Figure 3.8 illustrates this process as applied to the data from a naturalistic on-road driving.

3.5 Testbed and Dataset

Data is collected from naturalistic, on-road driving using the LISA-A testbed as shown in Figure 3.9. Three cameras are mounted facing the driver: one on 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 $640 \times 360$ 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
Figure 3.7: 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.

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.

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
Figure 3.8: 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.

Figure 3.8: 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.

conditions allowed for varying lighting conditions. Additionally, driving in an urban setting, the drivers passed through many stops 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 3.2 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 3.10 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 3.10a). However, yaw angle distribution is more spread out for the chosen events (Fig 3.10b).
Figure 3.9: LISA-A experimental testbed equipped with and capable of time synchronized capture of camera array and multiple Inertial Measurement Units (IMUS) [156].

Table 3.2: A list of events considered for evaluation, and its respective count and number of frames.

<table>
<thead>
<tr>
<th>Events</th>
<th>No. of events</th>
<th>Total no. of frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right turns</td>
<td>27</td>
<td>7417</td>
</tr>
<tr>
<td>Left turns</td>
<td>13</td>
<td>4027</td>
</tr>
<tr>
<td>Stop sign</td>
<td>42</td>
<td>10853</td>
</tr>
<tr>
<td>Right Lane change</td>
<td>12</td>
<td>2565</td>
</tr>
<tr>
<td>Left Lane Change</td>
<td>15</td>
<td>2963</td>
</tr>
<tr>
<td>Merge</td>
<td>7</td>
<td>1778</td>
</tr>
<tr>
<td>Xing</td>
<td>3</td>
<td>628</td>
</tr>
<tr>
<td>All</td>
<td>119</td>
<td>22814</td>
</tr>
</tbody>
</table>

3.6 Experimental Evaluations and discussion

The new CoHMEt framework, as introduced in this paper, 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.
3.6.1 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, 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 3.11. 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 3.3, 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 of the algorithms and the camera configurations are tightly coupled. This is further discussed below. From Table 3.3, 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 3.4.1, took $\sim 13$ seconds to process one frame and with search space reduction took $\sim 3$ seconds for one frame, a four fold improvement. CLM+POS, on the other hand, runs real-time with $\sim 25$ 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 [105] 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 3.12 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 $30^\circ$ 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 3.13 shows the quality of each 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.

3.7 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 pre-
Figure 3.11: 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.

Table 3.3: On-road performance evaluations of the proposed CoHMET.

<table>
<thead>
<tr>
<th>Method</th>
<th>Single camera view</th>
<th>CoHMET with 2-camera view</th>
<th>CoHMET with 3-camera view</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pitch (MAE, STD)</td>
<td>Roll (MAE, STD)</td>
<td>Roll (MAE, STD)</td>
</tr>
<tr>
<td>CLM+POS</td>
<td>(9.3°, 10.3°)</td>
<td>(8.6°, 12.2°)</td>
<td>(8.5°, 9.9°)</td>
</tr>
<tr>
<td>MPS+POS</td>
<td>(7.6°, 9.4°)</td>
<td>(8.2°, 6.4°)</td>
<td>(9.0°, 9.4°)</td>
</tr>
</tbody>
</table>
Figure 3.12: Error distribution with respect to the true head pose in yaw. The graphs reflect the first three error quartiles respectively of true head pose in yaw 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 3.13: 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.

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 [104] and multiple sensory cues [137].

Acknowledgements

Chapter 3 is in full a reprint of materials that are published in the IEEE Transactions on Intelligent Transportation Systems (2014), by Ashish Tawari, Sujitha Martin, and Mohan M. Trivedi, and in the IEEE international Conference on Image Analysis and Processing (2013), by Ashish Tawari and Mohan M. Trivedi. The dissertation author was the primary investigator and author of these papers.
This research was supported by research grants from the UC Discovery Program, Audi and VW Electronic Research Laboratory. The author is grateful for the advice and support of the LISA colleagues.
Chapter 4

Head, Eye and Iris Dynamic (HEIDy) Cue and Gaze Zone

4.1 Introduction

In this chapter, we look into the efficacies of gaze surrogate measures including head pose. However, a system using head pose alone is inherently limited. Towards this end, we propose novel gaze surrogate measures based on iris and eyelid region analysis. To evaluate the performance of these measures for estimating direction of attention, we study the gaze zone estimation problem. Here, gaze zone estimation refers to the localization of gaze direction towards the regions/zones relevant to the driving tasks such as front, left/right side mirrors, speedometer, center console, etc. Furthermore, we conduct systematic studies to understand the incremental benefits of different gaze surrogates and their dynamics in estimating gaze zone.

4.2 Related Studies

Driver distraction and inattention detection require monitoring driver’s head and eyes. Vision based systems are commonly used for this purpose as they provide non-contact and non-invasive solutions. We first present select relevant gaze zone related studies utilizing head pose and eye cues.

To determine head orientation, Ji and Yang [71] used an eigenspace algorithm to map pupil related feature to head pose space and further quantized the orientation into seven angles between $-45^\circ$ to $+45^\circ$. The pupil itself is obtained from bright and dark pupil image acquired using a specialized hardware setup. Gaze direction is then estimated using information about the head movement and relative position between pupil and glint, with the gaze direction quan-
tized into nine zones. The work showed promising results under different lighting conditions in laboratory settings. Its evaluation on on-road driving data is not available.

Facial feature based head pose estimation methods analyze geometric configuration of the features along with face model (e.g. cylindrical [163], ellipsoidal [84] or mean 3D face [94]) to recover head pose. Smith et al. analyzed global motion, and color and intensity statistics to track head and facial features such as eyes, lip corners, and the bounding box of the face [140]. From these tracked features, they estimated continuous head orientation and gaze direction. However, the method tracked these features separately and cannot always find them e.g. during partial occlusion due to hand, eye wears or during conversations. Kaminski et al. [74] detected pupils, pupil glints and nose bottom by analyzing the intensity, shape, and size properties. Using these properties, and a geometric model of human face and eye, continuous head orientation and gaze direction are estimated. However, the accuracy of the eye location significantly drops in the presence of large head movements, causing degradation in the performance for deviation from the frontal pose.

Remote eye tracking methods have shown good performance under constrained environments. However, such systems are highly susceptible to illumination changes, particularly in real-world driving scenarios. Many of the existing methods use active IR illumination techniques to obtain bright and dark pupil images which are further processed to detect and track iris and pupil. Their success strongly depends on the brightness and size of the pupils which are affected by several factors, including eye closure, eye occlusion due to the face rotation, external illumination interference, distance of the subject to the camera, and the intrinsic properties of the eyes (i.e., the bright pupil reflection tends to be darker for older people) [2, 109]. Due to these reasons, methods using corneal reflection with infrared illumination have been primarily used in an indoor setting [54] but are vulnerable to sunlight.

Passive light approaches relying on visible light are potentially better suited for outdoors. Shape- and appearance-based are two prominent approaches for eye detection. Shape based methods are based on the premise that the open eye is well described by its shape, which includes the iris and pupil contours and the exterior shape of the eye (eyelids). Simple ellipse [86] to more complex models e.g. incorporating eyelids shape parameters [179] have been proposed to model the eye shape. However, in naturalistic setting the shape of the eye can change significantly due to free head movement, facial expression, eye state (open/closing) etc., see Figure 4.1. Appearance models or image-template based models [168] detect and track eyes directly based on the photometric appearance as characterized by the color/intensity statistics of the eye and its surroundings. These methods usually demand large amount of training data with eye images of different subjects, under different orientations and illumination for reasonable performance. The existing methods, both shape- and appearance-based, are to a large extent applicable to near frontal view angles, fully open eyes, and under relatively constrained light conditions. For further details on different approaches of eye tracking and gaze estimation, we encourage readers
Recently, local patch based approaches have shown promising results for object detection, recognition and categorization. The local-feature based methods have also been applied to the detection of facial landmarks such as eye, nose and mouth corners [29, 175, 129, 186]. We formulate iris tracking problem in this framework (Section 4.4.1) without geometric shape-model and potentially provide better accuracy. Since features are calculated from local neighborhoods, they provide more robust performance to face pose and illumination changes than the holistic appearance based eye and iris detection approaches. Furthermore, from tracked eye landmarks and iris center, we propose ‘robust’ gaze-surrogate (Section 4.4.2) measures and show its efficacy for the driver’s gaze zone estimation. We compare the performance with a head pose alone system. Our analyses show promising results on the naturalistic driving dataset (Section 4.5.1). Also important to note is that the models for head pose through facial landmarks and iris tracking are learnt using separate datasets [52, 64]. This shows ability of the framework to generalize to different subjects.

4.3 Proposed Gaze Zone Estimation Framework

Figure 4.2 shows an overview of the proposed gaze zone estimation approach. It takes video streams as input, in our case, distributed camera feeds observing the driver. It extracts gaze-surrogate measures from the raw video and post-processes to form the feature vector which in turn, serves as an input to the gaze-zone-classifier. The classifier maps the feature space to one of the gaze-zones under consideration. It can also provide gaze zone membership, i.e. a probability that the driver’s gaze belongs to a given gaze-zone. We have discussed in detail about head pose estimation using CoHMET framework in Chapter 3. Here, we provide details about the features extracted from the head pose and its dynamic, which is used by a classifier to determine the zone-membership.
Next, we discuss the proposed novel gaze-surrogate measure using head, eye and iris dynamics (HEIDy). We describe in detail about the proposed iris detection and tracking approach. In Section 4.5, we present systematic comparative performance analysis of each of the proposed gaze surrogate measures from head pose, head dynamics and HEIDy cues.

### 4.3.1 Head Pose Feature Extraction

Literature suggest that in visual search head and eye gaze movements interact to each other in a coordinated manner to ‘optimally’ guide through the search space [4, 121, 39]. Such coordination between head and eye dynamics should be exploited in order to better predict the gaze zone using head pose cues.

It’s our hypothesis that how one arrives to a particular head pose, can predict how eye gaze might have moved. For this investigation, we examine the role of both, static head pose-and head pose dynamic features, for gaze zone estimation.

1. Static features: Current head pose angles, i.e. yaw, pitch and roll, form a part of the static features. The raw signal, however, has high frequency components causing tiny fluctuations. Hence, we pass it through a low pass filter, a simple moving average as in Eq. 4.1.

\[
y[n] = \frac{1}{L} \sum_{i=0}^{L-1} x[n - i]
\]  

(4.1)

where, $L$ is the filter length. Besides the current head pose, we also calculate center of the face, as the centroid of the two eyes and mouth region. This is an informative cue specially in the constraint sitting position of the driver. Thus, the total length of the static features is $3 + 2 = 5$.

2. Dynamic features: To capture the time trend, we extract several statistics of the time series. First, we smooth the raw data as in case of the static features. Then, from the windowed time series (through $W$ sec prior to the current time $t$), we extract: minimum value, position of the minimum, maximum value, position of the maximum, mean angle, mean angular velocity. The statistics are taken in all the three pose angles. Thus, the total length of the dynamic features is $3 * 6 = 18$. 
4.4 HEIDy Feature Extraction

In this section, we detail features extracted via Head, Eye and Iris Dynamics (HEIDy) analysis to infer driver’s gaze zone. First, we discuss local patch-based regression framework for robust iris tracking. We detail the training procedure and the iterative refinement steps to detect the iris centers jointly for both the eyes. Next, we discuss gaze-surrogate features estimated using eye landmarks and the iris centers.

4.4.1 Iris Center Detection and Tracking

The iris centers are detected by training a sequence of regression matrices $S = (R_1, ..., R_K)$. This sequence of regression matrices $S$ together with an initial estimate of the positions $p_0$ with respect to Image $I$ can be used in Algorithm 1 to obtain the positions of iris centers $p_K$ after desired $K$ iterations. At each iteration, the estimated positions of iris centers $p_k$ are computed by:

$$p_k = p_{k-1} + R_{k-1} \text{feat}(p_{k-1}, I)$$

where $\text{feat}(p_{k-1}, I)$ gives a feature vector (e.g., HoG descriptor [30] in our case) extracted at $p_{k-1}$ in image $I$. This can be viewed as a coarse to fine adjustment of the center locations as the successive model attempts to reduce the residual errors.

In Algorithm 2, the pseudo code describes the training procedure for obtaining a sequence of regression matrices $S = (R_1, ..., R_K)$. Assume that we are given a set of images $\{I^i\}$, their corresponding ground-truth iris centers $\{p^*_i\}$, and their corresponding initial estimate of iris centers $\{p^0_i\}$. When training a regression matrix $R_k$ at iteration $k$, we want to minimize the distance between the true iris center positions $p^*_i$ and the estimated positions $p^i_k$ in every image. $p^i_k$ represents the positions at iteration $k$ starting from the initial estimate of iris center positions $p^0_i$. The loss function at each iteration can be written as in Eq.4.3, and our goal is to minimize this loss function.

$$L = \sum_i \sum_{p^i_k} \| p^*_i - p^i_k - R_k \text{feat}(p^i_k, I^i) \|^2$$

The regression matrix $R_k$ which minimizes the loss function can be obtained by solving the linear least squares problem shown in Eq.4.4.

$$R_k = \arg \min_{R_k} L$$

Once the regression matrix $R_k$ is found, it is plugged into Eq.4.2 to compute the new $p^i_k$ which can be used in the next iteration. By repeating the same process for $K$ iterations, we obtain the sequence of regression matrices $S = (R_1, ..., R_K)$.

Before training an iris center detection model (a sequence of regression matrices), there
are important image processing steps needed to be done to ensure training quality and correctness. They are scaling, rotation, and histogram equalization. For scaling, an image is scaled to a size such that the distance between both eye centers equals to a fixed length. For rotation, an image is rotated in a way that the line going through both eye centers is parallel to the x-axis. Scaling and rotation ensure the size and the orientation of eyes to be approximately the same for every image. Furthermore, histogram equalization is applied only in the eye region of an image to enhance image contrast for a better low-level feature extraction.

Once images are registered via the above steps, we train each regression matrix $R_k$. For $k = 1$, we need initial estimate for the iris centers. It’s important to provide enough perturbations to the system to learn a ‘good’ model. We assign 10 initial estimates of the iris center to each eye in an image during training. The 10 initial positions are illustrated in Figure 4.3. They consist of eye centers (estimated from the tracked eye corners) and eight positions around the centers, and a ground-truth position. Inclusion of ground truth is to ensure regressor, $R_1$, does not deviate from actual position at the first step.

In order to improve iris center detection rate in videos, iris tracking is implemented. The tracking is done by using the detected positions from the previous frame as an initial estimate of the current frame. Since there is usually not much movement between two consecutive frames, it’s a very effective way for tracking and increasing iris center detection rate.

**Algorithm 1 Testing**

<table>
<thead>
<tr>
<th>Input:</th>
<th>Image I, initial estimates $p_0$ and regression matrices $S = (R_1, ..., R_K)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:</td>
<td>for $k = 1$ to $K$ do</td>
</tr>
<tr>
<td>2:</td>
<td>$h = feat(p_{k-1}, I)$</td>
</tr>
<tr>
<td>3:</td>
<td>$p_k = p_{k-1} + R_{k-1}h$</td>
</tr>
<tr>
<td>4:</td>
<td>end for</td>
</tr>
</tbody>
</table>

**Output:** $p_K$
Algorithm 2 Training

Input: Number of iterations $K$, data $I^i, p_i^k, p_i^*$ for $i = 1...N$

1: for $k = 1$ to $K$ do 
2: \[ h^i = \text{feat}(p_{k-1}^i, I^i) \]
3: \[ L = \sum_i \sum_{p_{k-1}^i} \|p_{k}^i - p_{k-1}^i - R_{k-1} h^i\|^2 \]
4: \[ R_{k-1} = \arg \min_{R_{k-1}} L \]
5: \[ p_{k}^i = p_{k-1}^i + R_{k-1} h^i \]
6: end for

Output: $S = (R_1, ..., R_K)$

4.4.2 Gaze-Surrogate Estimation

Given eye landmarks and detected iris centers, we compute two gaze-surrogate measurements: one to account for horizontal gaze movement and another for vertical gaze movement.

Vertical gaze movement with respect to head is inferred from the eye area. The eye area is computed using only the upper eyelid contour. This is inspired from the physiology of normal eye, where the upper eyelid naturally moves in the same direction as the vertical-gaze. Lower eyelid, however, does not provide any gaze relevant information. In fact, it may introduce noise e.g. upward movement of lower eyelid due to squinting or a facial expression. We believe this subtle difference is important to avoid unnecessary noise and hence improve the gaze estimate.

Horizontal gaze-direction is explained in our contemporary work by Tawari et al. [148]. We describe mathematical details here again for the sake of completion and the reader’s convenience. The horizontal gaze-direction $\beta$ with respect to head, see Figure 4.4, is estimated as a function of $\alpha$, angle subtended by an eye in the horizontal direction, head-pose (yaw) angle $\theta$, and the ratio of the distances of iris center from the detected corner of the eyes in the image plane. Equation 4.5-4.6 show the calculation steps.

\[
\frac{d_1}{d_2} = \frac{\cos(\theta - \alpha/2) - \cos(\theta - \beta)}{\cos(\theta - \beta) + \cos(180 - \theta - \alpha/2)} \quad (4.5)
\]

\[
\beta = \theta - \arccos\left(\frac{2}{d_1/d_2 + 1} \sin(\theta) \sin(\alpha/2) + \cos(\alpha/2 + \theta)\right) \quad (4.6)
\]

Since the raw eye-tracking data is noisy (due to blinking and tracking errors), we smooth angle $\beta$ with a median filter.
Figure 4.4: Eye ball image formulation: estimating $\beta$, gaze-angle with respect to head, from $\alpha$, $\theta$, $d_1$ and $d_2$

4.5 Experimental Evaluation

In this section, we report a series of experiments and results of overall gaze zone framework as well as intermediate modules. For zone classification, we use a random forest classifier in conjunction with proposed feature sets. Random forest has shown promising results in many machine learning applications. We also choose this classifier because of the ease of the ability to interpret the learned parameters, and the low number of tuning parameters. We used the random forest library available in Matlab. The only parameters we tune is the number of trees, as shown in Figure 4.5 for Head pose features based system. Similar out-of-bag error analysis is used to determine the number of trees in different experiments. Next, we introduce our database and the annotation tool box developed for fast and accurate annotation.

4.5.1 Datasets and Annotation

Here, we describe two databases used for the development and evaluation of the iris center detection algorithm and the gaze zone estimation framework.

We trained the iris center detection model using images from the Labeled-Faces-in-the-Wild dataset [64]. LFW is a face recognition database which contains more than 13,000 images of faces collected from the web, and 1,680 of people have two or more distinct images in the dataset. We randomly picked a total of 330 images from the dataset and manually annotated the iris centers for all the images. 270 of the images are used for training, and the rest of the 60
images are used for testing.

The iris center detection model is also applied to our naturalistic driving dataset to evaluate their efficacy for gaze zone estimation in real-world data.

For gaze zone estimation, the driving data is collected from naturalistic, on-road driving using the LISA-A testbed as shown in Figure 4.6. There are two looking-in cameras facing the driver for recording the driver’s head and eye movements, and one looking-out camera for recording the front road condition. The two looking-in cameras are mounted on the A-pillar and near the rear view mirror. They capture the face view in color video stream at 30fps and 640 × 360 pixel resolution. From the collected data, we mined sections where driver is making different maneuvers including right/left turns, right/left lane changes, stops at stop signs, and freeway merges for ground truth generation.

The above dataset is annotated for the gaze zone ground truth. This annotation effort focuses on annotating the driver’s gaze zone at a given moment. A graphical user interface is developed to assist a human expert to efficiently label the video frames. The expert is presented with both the driver looking camera views as well as the outside looking camera view to provide full contextual information, and is asked to utilize all the information including eye data to annotate the ‘best’ perceived gaze zone region. The annotation toolbox displays a vehicle interior front-view image, two looking-in camera views, and one looking-out camera view (see Figure 4.7).
Figure 4.6: LISA-A experimental testbed equipped with and capable of time synchronized capture of looking-in and looking-out cameras.

The looking-in and looking-out camera views are for assisting a human expert to determine where the driver is looking, and the vehicle interior image is an annotating plane to perform labeling. A human expert would look at the three available camera views to determine which zone the driver is looking at and draw a box within the zone on the vehicle interior image, such as the green box seen in Figure 4.7. We have collected over 2,000 annotations from this naturalistic driving data, and these annotations are used in the gaze zone estimation experiments.

4.5.2 Results: Iris Tracking

For determining a good patch size to extract features around each position $p_k$, a total of six different iris center detection models are trained for six different patch sizes, which are 8, 16, 24, 32, 40, and 64. Figure 4.9 presents normalized error in training set as a function of iterations. The normalized error of the models at each iteration was calculated by computing the difference between the detected iris center positions $p_k$ and the annotated ground-truth positions $p_*$, normalized by the distance between the two corners of the eye. Patch size 40 produced the best results. No significant improvement beyond iteration 2 is observed, hence we choose $K = 2$. Next, we evaluated the model performance in the testing set of 60 images. The average normalized error obtained on testing set is 0.0532.

After verifying the iris center detection model on the LFW dataset, the next step of the evaluation is in applying the model on our naturalistic driving data to determine how well it performs in real-world conditions. First, a qualitative evaluation is performed by human experts manually inspecting over 1,000 frames with different driver head and eye poses. The analysis shows promising results. Figure 8 shows the images of the 5 good and the 5 bad examples from the naturalistic driving data based on human inspection. Second, the quantitative evaluation of
Figure 4.7: Graphical user interface for annotating gaze zone. A human expert uses all the contextual information (driver and surround views) for ground truth generation. Top portion shows the approximate regions in vehicle frame used for gaze annotation.

the iris detection is performed based on gaze experiment as discussed in the following section.

4.5.3 Results: Gaze Zone Estimation

In this section, we conduct two experiments to estimate driver’s gaze zone. First experiment is with head pose information alone, and second experiment is with both head and eye cues. For each experiment, the naturalistic driving data is divided into several short time segments where the driver possibly gazed into different zones. We call these segments as events. Each event is randomly assigned to one of ten folds. Each fold contains multiple events, but there is no duplicated events among folds. This is to ensure that evaluation is performed on temporally separated frames. We provide performance results for randomized tenfold cross validation. The system is trained on nine folds at a time and tested on the remaining fold. This procedure is repeated ten times, each time leaving out a different fold.
Figure 4.8: An illustration of patch size 8 (blue), 16 (red), 24 (green), 32 (pink), 40 (black), and 64 (purple).

Figure 4.9: Normalized error as a function of iterations for models with different patch sizes. Patch size 40 converges fast and provides better error performance.
Gaze zone using head pose and dynamic features

We have divided driver’s view into eight different gaze zones related to driving tasks. The zones are labeled as following, 1 - left shoulder, 2 - left window rear-view mirror, 3 - front windshield, 4 - speedometer, 5 - center rear-view mirror, 6 - center console (infotainment panel), 7 - right window rear-view mirror and 8 - right shoulder (see Figure 4.7).

First, we examine the effects of time window size $W$ chosen for the dynamic features on the performance of the classifier. Figure 4.10 shows that as the time window increases the accuracy improves. Window size greater than 2 seconds does not provide any further significant improvement. Hence, we chose and fixed this window size for the next set of experiments.

We compare the classification performance using static-alone and static-cum-dynamic features. Table 4.1 and Table 4.2 show the confusion matrices and the classification accuracies for static and static-cum-dynamic features respectively. It can be seen that dynamic features significantly improve the classification accuracies. A closer look at the confusion matrices suggests that much of the confusion exists in adjacent gaze zones, and front (zone #3) and speedometer zones (zone #4) are the most confused ones. Figure 4.11 shows working of the overall system with probabilistic output for the three zones (Right, Front and Left) prediction, using dynamic features. Notice that static head pose (bottom right), without eye information, can be misleading, i.e. head pose alone (without dynamic information) would suggest non-frontal gaze zone.

To investigate the importance of the different dynamic features, we perform a permuta-
Table 4.1: Static Feature: Confusion matrix for 8 gaze zones numbered as shown in Figure 4.7.

<table>
<thead>
<tr>
<th>True Gaze Zone</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>85.9</td>
<td>13.2</td>
<td>0.9</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>2</td>
<td>1.1</td>
<td>87.5</td>
<td>11.2</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>3</td>
<td>0.4</td>
<td>5.3</td>
<td>88.3</td>
<td>1.3</td>
<td>1.9</td>
<td>0.6</td>
<td>1.8</td>
<td>0.3</td>
</tr>
<tr>
<td>4</td>
<td>0.0</td>
<td>0.8</td>
<td>42.4</td>
<td>49.6</td>
<td>4.0</td>
<td>3.2</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>5</td>
<td>0.0</td>
<td>0.0</td>
<td>12.4</td>
<td>0.0</td>
<td>80.6</td>
<td>4.1</td>
<td>2.8</td>
<td>0.0</td>
</tr>
<tr>
<td>6</td>
<td>0.0</td>
<td>0.0</td>
<td>9.3</td>
<td>0.0</td>
<td>5.8</td>
<td>83.3</td>
<td>1.3</td>
<td>0.0</td>
</tr>
<tr>
<td>7</td>
<td>0.0</td>
<td>0.0</td>
<td>2.6</td>
<td>0.0</td>
<td>6.4</td>
<td>2.0</td>
<td>87.1</td>
<td>1.9</td>
</tr>
<tr>
<td>8</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.1</td>
<td>0.0</td>
<td>25.8</td>
<td>73.0</td>
</tr>
</tbody>
</table>

Unweighted Accuracy = 79.4%
Weighted Accuracy = 85.7%

The idea is that if the variable is not important, then randomly permuting the values of variable in the out-of-bag cases will not degrade prediction accuracy. Our findings are, first, the dynamic information of the yaw angle is found the most important and that of roll angle the least important. This is expected since the yaw angle shows the largest variability for the different gaze zones. Second, among the statistics, we find the position of the extrema (min and max) and mean angular velocity of yaw and pitch to be more important than other statistics. This suggests that the direction of arrival to a particular pose can better indicate where one might be looking.

While dynamic features significantly improve the system performance, certain zones are inherently confusing such as front and speedometer, as subtle eye movement is sufficient to guide the eye gaze between them. The next experiment incorporates gaze surrogate features from eye region.

Gaze zone using HEIDy features

For this study, we excluded zones 1 (left shoulder) and 8 (right shoulder), since for these zones, usually, iris information is not available because of self-occlusion due to large deviation of
Table 4.2: Dynamic Feature: Confusion matrix for 8 gaze zones numbered as shown in Figure 4.7.

<table>
<thead>
<tr>
<th>True Gaze Zone</th>
<th>Recognized Gaze Zone</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2  3  4  5  6  7  8</td>
</tr>
<tr>
<td>1</td>
<td>94.4 5.2 0.5 0.0 0.0 0.0 0.0</td>
</tr>
<tr>
<td>2</td>
<td>0.8 95.7 3.3 0.2 0.1 0.0 0.0</td>
</tr>
<tr>
<td>3</td>
<td>0.2 2.3 93.9 1.4 1.1 0.5 0.5</td>
</tr>
<tr>
<td>4</td>
<td>0.0 0.8 45.6 48.0 2.4 3.2 0.0</td>
</tr>
<tr>
<td>5</td>
<td>0.0 0.0 5.7 0.0 90.9 2.3 1.1</td>
</tr>
<tr>
<td>6</td>
<td>0.0 0.0 6.1 0.0 3.3 90.4 0.3</td>
</tr>
<tr>
<td>7</td>
<td>0.0 0.0 1.4 0.0 1.9 0.3 96.1</td>
</tr>
<tr>
<td>8</td>
<td>0.0 0.0 0.0 0.0 2.2 0.0 1.1</td>
</tr>
</tbody>
</table>

Unweighted Accuracy = 88.2%
Weighted Accuracy = 93%

the head from the frontal camera pose. Also, these zones are well recognized by head pose features alone as seen earlier. Here, we present quantitative assessments of performance improvements (if any) achieved by the proposed simple and robust HEIDy features.

The result using HEIDy feature is presented in Table 4.3. Clearly, the system with eye features significantly outperforms the head pose alone system by 9% improvement in per-class classification accuracy for the six zones. Notice, the center-console (zone #6) and the center rear-view mirror (zone #5) are well separated with no false-detection between them. This can be attributed to the effectiveness of the vertical-gaze-surrogate feature. Also, note that the zone #3 (front windshield) performance for this system is much better than that for the previous system, where the front zone is confused with the adjacent zones. This can be attributed to the effectiveness of the horizontal-gaze surrogate feature. The greatest improvement is noticed for the zone #4, the speedometer, as expected. Yet, this is the zone with the lowest accuracy and the most confusion with the front zone. This suggests the inherent difficulty associated with separating these two zones. A closer inspection on frame by frame results, though, are inspiring as the adjacent frames are often classified correctly. Hence, a video segment level analysis may further improve the result.
Figure 4.11: Visualization of output by CoHMEt and Zone Estimation: Bottom left: input video feeds with multi-camera perspective; bottom right: Tracked head pose visualization; top left: 3 zones (left, front and right) visualization; top right: zone membership. Notice, static head pose alone (without eye and dynamic information) can be misleading in determining gaze zone.

Table 4.3: HEIDy feature: Confusion matrix for 6 gaze zones numbered as shown in Figure 4.7. Notice the performance improvement in speedometer and other zones.

<table>
<thead>
<tr>
<th>True Gaze Zone</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td></td>
<td>96.8</td>
<td>3.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>3</td>
<td>0.0</td>
<td></td>
<td>98.2</td>
<td>0.0</td>
<td>1.8</td>
<td>0.0</td>
</tr>
<tr>
<td>4</td>
<td>0.0</td>
<td>20.8</td>
<td>79.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>5</td>
<td>0.0</td>
<td>3.6</td>
<td>0.0</td>
<td>91.7</td>
<td>0.0</td>
<td>4.8</td>
</tr>
<tr>
<td>6</td>
<td>0.0</td>
<td>0.0</td>
<td>1.2</td>
<td>0.0</td>
<td>98.8</td>
<td>0.0</td>
</tr>
<tr>
<td>7</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>3.0</td>
<td>0.0</td>
<td>97.0</td>
</tr>
</tbody>
</table>

Unweighted Accuracy = 93.6%
Weighted Accuracy = 94.9%
Figure 4.12: Challenges faced in estimating gaze direction when the driver looks at the speedometer. The top row is driver looking at speedometer. The bottom row is right before driver looks at speedometer.

Figure 4.13: Example results on the real-world driver dataset. The top row has five correct results, and the bottom row has five incorrect results.

4.6 Concluding Remarks

Driver’s gaze estimation is an important component of the driver monitoring systems. We motivated that the coarse gaze direction is sufficient in a number of applications. We presented a computational framework to estimate coarse gaze direction and infer gaze zones using head and eye cues. Our emphasis is on the practical realization of a robust and reliable system. Towards this end, we presented a novel iris tracking algorithm with no geometric shape assumptions. Since the features are computed in the local neighborhood, the method enjoys the benefits as in other local patch-based methods in other applications over holistic appearance based approaches. We further proposed gaze surrogate measurements to estimate horizontal and vertical eye movements. Our analyses for gaze zone estimation show very promising results with significant improvement over head pose alone system.

In current implementation, we only use video sequence for iris tracking. In future, we will extent proposed framework to utilize video sequence to also incorporate gaze-dynamics in inferring gaze zones as well as other driver activities [111].
Acknowledgements

Chapter 4 is in part a reprint of materials that are published in the IEEE Intelligent Transportation Systems Conference (2014), by Ashish Tawari, Kuo Chen and Mohan M. Trivedi, and the IEEE Intelligent Vehicles Symposium (2014), by Ashish Tawari and Mohan M. Trivedi. The dissertation author is the primary investigator and author of these papers.

This research was supported by research grants from UC Discovery Program, Audi and VW Electronic Research Laboratory for their support and collaboration. The author would like to thank LISA colleagues for their valuable assistance.
Chapter 5

Systems and Analysis

5.1 Introduction

So far, we have discussed robust approaches to analyze driver’s head and eye to accurately determine the attention direction. This mainly pertains to the looking-in component of the proposed FoViA framework, as discussed in Chapter 2. In this chapter, we will discuss systems incorporating looking-out component and we provide joint analyses and evaluations. For most practical applications, just knowing raw attention direction is not sufficient but instead we would like to know whom or what the drivers are looking at, and conversely, whom or what they are ignoring. We motivated earlier that ADASs with such inference making capability can not only appropriately warn the driver of unseen hazard but also reduce nuisance alerts and improve the driver’s trust in the system. Similarly, if the driver is distracted from the forward driving direction, an appropriate alert or even active assistance such as, automatic brake or lane keep, can be provided. For the design of such futuristic ADASs, we study two particular systems pertaining to the visual attention: (1) Attentional target detection system and (2) Attention Guard - a holistic distraction detection system. Following sections describe the two systems.

5.2 Attentional Target Detection System

The goal of this system is to determine the target attended to by the driver. For this, it utilizes proposed looking-in looking-out framework by simultaneously observing the driver and the driver’s field of view. We propose to measure coarse eye position and combine the salience of the scene to understand what object the driver is focused on at any given moment. Note that we are not proposing a precise gaze tracking approach, but rather a coarse gaze direction combined with the analysis of scene salience to determine important objects of interest - in our case pedestrians.
Our interest in pedestrians comes from the fact that in 2011, pedestrian deaths accounted for 14 percent of all traffic fatalities in motor vehicle traffic crashes in the United States. Almost three-fourths (73%) of pedestrian fatalities occurred in an urban setting versus a rural setting. 88% of pedestrian fatalities occurred during normal weather conditions (clear/cloudy), compared to rain, snow and foggy conditions. By knowing which pedestrians the driver has and has not seen, measures against collisions can be taken more accurately. While our main focus is on pedestrians, the framework can easily accommodate any object of interest or even a low-level saliency model to estimate the focus of attention.

To study this system in naturalistic driving conditions, we needed a testbed which can observe the driver and the driver’s field of view simultaneously. To observe the driver’s large field of view (as driver’s can move their head to look around), we also needed a camera setup with a large field of view (almost 180\degree panoramic view) or a system that can move with the driver’s head to provide his/her field of view. We chose the later setup of a head mounted camera. This provides first-person or egocentric view. This is not a unique choice, in fact, a panoramic view camera system or any setup which can detect and track objects around the vehicle (e.g. lidar or radar system) can work. We chose egocentric view because of its ease of availability.

First-person or ego-centric vision is not a new field; however, in recent times they have become more popular and practical thanks to the technological advances made in lightweight, wearable, egocentric cameras. The GoPro camera, for instance, can be mounted on helmets and is popular in a lot of sports such as biking, surfing, and skiing. The Microsoft SenseCam can be worn around the neck and has enough video storage to capture an entire day for the idea of “life logging”. Cognitive scientists like to use first-person cameras attached to glasses (often in combination with eye trackers such as Tobii or SMI) to study visual attention in naturalistic environments. Most recently, emerging products like Google Glass have begun to make the first attempts to bring the idea of wearable, egocentric cameras into the mainstream.

Ego-centric vision attempts to understand human behavior by acquiring information on what the person is looking at [75]. It employs videos/images from head mounted cameras. The ego-centric vision, however, comes with its own challenges. We give an overview of relevant related work in Section 5.2.1 and explain the methods for determining gaze and detecting pedestrians in Section 5.2.2. In Section 5.2.3 we briefly review our captured data, and Section 5.2.4 shows our results.

5.2.1 Related Research

Use of wearable cameras is not new [131]. In the last decade, gaze tracking systems such as [35], Tobii and SMI have made mobile gaze tracking in real life settings possible. More recently, the advances in hardware technology have made their usage more common in the computer vision community [143, 77, 126, 46, 87]. These systems are often used successfully in laboratory
or controlled lighting conditions. Their use in complex environments is limited due to lengthy
calibration, motion, illumination changes and, in case of driving, possible hindrance to the driver’s
front- or side-view. We discuss select work in activity recognition and gaze-behavior related
research areas which are relevant in our current and larger interest in studying driver intent and
behavior in real-world driving.

Ogaki et al. [110], using an inside-out camera system, combined eye-motion from an
inside looking camera and global motion from an outside one to recognize indoor office activities.
The authors suggest that joint cues from inside looking and outside looking cameras perform the
best across different users. Doshi and Trivedi [38] introduced a similar system, but primarily for
vehicular use. Pirsiavash and Ramanan [123] detected indoor apartment activities of daily living
in first person camera view. They used object-centric action models which perform much better
than low-level interest points based ones to recognize activities. They show that using ground-
truth object labels in the action models significantly improve recognition performance. This
suggests that recognizing objects of interest is key to recognizing tasks/activities in naturalistic
settings.

Gaze allocation models are usually derived from static picture viewing studies. Many
of the existing works are based on the computation of image salience [67] using low-level image
features such as color contrast or motion to provide a good explanation of how humans orient
their attention. However, these models fail for many aspects of picture viewing and natural task
performance. Borji et al. [15] observe that object-level information can better predict fixation
locations than low-level saliency models. Judd et al. [73] show that incorporating top-down
image semantics such as faces and cars improves saliency estimation in images.

Inspired by the above findings, we present a driver’s visual attention model using inside
and outside looking camera views. In particular, we propose a model to determine coarse gaze
direction and combine it with an object based saliency map to determine the allocated attention
of the driver. Note that our interest lies in ‘higher-level’ semantic information about the driver
attention and not ‘low-level’ precise gaze measurement. Our proposed framework circumvents
the precise gaze estimation problem by utilizing a saliency map to achieve robust performance.
Precise eye gaze from remote cameras is difficult not only due to low resolution of the eye region,
but also due to large head turns, self occlusion, illumination changes and hard shadows existing in
an ever changing dynamic driving environment. To deal with large head turns and self occlusion,
we propose to use a distributed camera system to monitor the driver.

We evaluate the proposed framework using a novel naturalistic driving data set using
multiple cameras monitoring the driver and the outside environment. We use the head mounted
camera from Google Glass to capture the driver’s field of view. This particular device did not
provide the ability to automatically synchronize footage with other cameras at per frame level.
However, the ease, quick setup time (wearing and pressing capture button) as well as clean and
uncluttered face view still makes the device a good choice. To obtain frame level synchronization,
we mount an outside looking camera on the ego-vehicle which in turn is synchronized to the rest of the systems. Details on our synchronization strategy is provided in Section 5.2.3. A head mounted camera provides the ability to capture not only the driver’s outside field of view but also inside cockpit-view. In this work, we focus on the analysis of the outside view using the head mounted camera. This view poses unique challenges as discussed later.

5.2.2 Proposed Approach

To infer the driver’s attention, we are interested in knowing what object, in our case which pedestrian, the driver is looking at. There are two steps involved: first, estimating where driver is looking and second, detecting objects of interest in his/her field of view.

Please refer to the previous chapter for estimating the gaze direction. In our current analysis, we focus on horizontal gaze variation $\beta$ (see Equation 4.6) that is the most volatile and exercised direction by the driver to gain the knowledge of the environment. Below, we describe the second step of salient region detection and then attended object determination by combining the gaze-direction and the detected salient objects.

Salient Object Detection

Salience of an area of interest is determined based on whether the area contains the pedestrian or not. This is object based saliency, which has shown to be a better measure than low level visual feature for certain tasks [44]. For this, we require a pedestrian detector. Using first person view presents a number of interesting challenges compared to a stationary car-mounted camera. The major challenge is to determine the region of interest in which to look for pedestrians. With a stationary camera, it is either mounted so there are no obstructions in its view of the road, or it is mounted so any obstructions can easily be masked out manually.

This is not the case for first person view, where the perspective constantly changes and there is no way of setting up a constant mask. This section introduces an algorithm to automatically mask out the dashboard and other unwanted areas.

The pedestrian detection module in this system is based on the classic HOG-SVM detection presented by Dalal and Triggs in [31]. It is trained on the Inria person dataset from the same paper. The pedestrian detection itself is simply a module in the full system, and it could be swapped with other approaches without issues.

The most important part of the interior mask is the dashboard mask. The dashboard can take up just the bottom of the image, the majority of the image, or not be present at all (fig. 5.1) and the algorithm must handle all of those situations. We detect the distinct line between the windshield and dashboard and build from that:

1. Smooth out the input image with a Gaussian blur to even out noise.
2. Detect edges using the Canny edge detector [19].

3. Determine the major lines in the image using the generalized Hough transform [42].

4. Filter the lines by angle to include only near-horizontal lines.

5. Build a confidence map of the dashboard.

Figure 5.2 shows sample output of step 4. Green lines are those that are horizontal enough to be considered in the dashboard map, red lines are ignored due to their extreme angles.

For each detected line, a polygon is drawn, which masks out all of the image below the line. These masks are combined and result in a single-frame dashboard map. To counter noisy line detections, a cumulative confidence map is introduced.

The cumulative confidence map is created by adding 1 to all pixels in the map covered by the current single-frame map and subtracting 1 from all pixels not covered by the current single-frame map. Areas that are detected in several subsequent frames will grow to a high
confident, but after a while of no detections, the confidence will fall and eventually the mask disappears. Examples of confidence maps and masks are in Figure 5.3.

The use of the cumulative map is governed by two parameters, $\kappa$ and $\lambda$. $\kappa$ is the mask threshold. Any pixel in the confidence map with a value higher than $\kappa$ is considered part of the dashboard map. In this implementation $\kappa = 2$. This parameter controls how confident the system must be in a given pixel to include it in the mask. $\lambda$ is the upper limit of confidence values. For a very high $\lambda$ value, the confidence can grow very high, thus resulting in a long delay before the pixel goes below $\kappa$. $\lambda$ defines how long the memory of the system is. In this implementation $\lambda = 10$.

Apart from filtering out the dashboard, we detect and filter out black blobs large enough that they can only be part of the vehicle interior. We also discard pedestrian bounding boxes larger than 40% of the frame height.

**Attended Object Determination**

This step combines gaze-direction ($\beta$) and the salient object detected to determine which object the driver is attended to. This requires mapping from gaze-direction with respect to the head, $\beta$ to pixel position in the external looking camera image. Equation 5.1 shows the mapping function to determine the $x$-position (i.e. the yaw-direction) in the image plane.
Figure 5.4: Examples of pedestrian detection scenarios where the exclusion mask has been overlaid in red.

\[ P_x(\beta; C_x, M_x, \phi) = C_x - M_x \ast \frac{\sin(\beta)}{\cos(\phi + \beta)} \]  

(5.1)

where \( \phi \) is the angle between the external camera image-plane and eye-image plane, \( C_x \) is the pixel position when looking straight \((\beta = 0)\), and \( M_x \) is a multiplication-factor, determining the change in pixel position with change in gaze direction. A calibration step with the user’s cooperation (by asking them to look in particular directions) can be performed to determine the parameters. Since the device is not firmly fixed to the head and can move during usage, we ideally need to perform calibration again. However, for our purposes, we found that as long as the camera is not rotated (along the vertical-axis allowed by the device for adjusting the display), it did not degrade the performance during normal usage. A Gaussian kernel around this location is combined with the detected object based image saliency to infer the allocated attention location. This leads to the attended object as the closest object detected around the gaze-location.
5.2.3 Dataset

Data is collected from naturalistic on-road driving using a vehicular test bed, which is equipped with one Google Glass and three GigE cameras as shown in Figure 5.5. The Google Glass is worn by the driver to give a first-person perspective. Data is captured from Google Glass at 50 frames per second with a resolution of 1280 x 720 pixels, and stored internally on the device. Of the three GigE cameras, one is mounted to the left of the driver near the A-pillar on the windshield looking at the driver, and two are mounted to the right of driver near the rear-view mirror on the windshield - one looking at the driver and one looking outside. A multi-perspective two camera approach is adopted to look at the driver because it increases the operational range when the driver makes spatially large head movements [147]. Data from the GigE cameras is captured at a resolution of 960 x 1280 pixels and is stored on a laptop with time stamps of millisecond precision. This allows for time synchronized videos.

In order to synchronize the first-person video with the videos looking at the driver, synchronization points are annotated using the first-person view and the outside front view. The criteria used in choosing these synchronization points include naturally occurring changes in traffic lights and artificially introduced momentary but periodic bursts of light (e.g. LED lights mounted to be visible in both first person view and outside front view). Then, assuming constant frame rate in the first-person video, linear interpolation is used to synchronize the first-person video with videos looking at the driver.

Using this test bed, multiple drivers were asked to drive on local streets. Approximately 40 minutes of data was collected in total, where the drivers passed through many stop signs, traffic signals and pedestrian crossings. In this paper, we are interested in events where the vehicular test bed is near or at these intersections, because these times are especially rich with visual interaction between driver and pedestrians. To evaluate our proposed attention system, two sets of ground truth labels are created via manual annotation on interesting event segments in the driving sequences. First, we manually annotated 410 frames and 1413 pedestrians, as seen in the first person perspective camera, with bounding boxes when either their face is visible or a significant portion of their body is visible. Second, we manually annotated 300 frames of where the driver is looking in the first person view - in particular, we annotated possible pedestrian candidate(s) as shown in Figure 5.6. This is accomplished by carefully looking at the driver’s head and eye movements in the time synchronized videos with significant utilization of temporal and spatial context. For example, by looking at the driver’s gaze over a time period, we are able to zero-in on particular pedestrians within a larger group. Annotating what the driver is looking at is especially challenging, and we have attempted to address this by obtaining consensus from multiple experts.
Figure 5.6: An example of annotated sequence from the time synchronized video.

Table 5.1: Dashboard masking cuts the false positive rate in half, without impacting the detection performance too much.

<table>
<thead>
<tr>
<th></th>
<th>False positives per frame (FPFF)</th>
<th>Detection rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-filtered (baseline)</td>
<td>2.94</td>
<td>0.27</td>
</tr>
<tr>
<td>With dashboard filter</td>
<td>1.45</td>
<td>0.21</td>
</tr>
</tbody>
</table>

5.2.4 Experimental Evaluation

In this section we discuss the results of an experimental evaluation over several hundred frames of manually labeled data. There are two main contributions to evaluate: the impact of the dashboard masking and the attention estimation performance. In this section, both will be tested separately and then in combination.

Dashboard masking cuts the number of false positives in half with a low impact on the detection rate, as shown in table 5.1. This paper is not about pedestrian detection as such, but the detection rates have been included to demonstrate that the masking does not impact them negatively in a significant way. The test set (1413 annotated pedestrians over 410 frames) is very challenging with articulated pedestrians and heavy occlusions, and while the detection numbers are low from an absolute point of view, the attention estimation still works well, as we shall see below.

The attention estimation has been tested on the same sequence with manually annotated pedestrians. Ground truth for the attentional location also was determined manually. Table 5.2 shows the accuracy of the proposed system given a perfect pedestrian detector, as well as the combined system. As points of comparison, we also include results of a simple attention estimator using only head pose - the center-bias based solution. This places the focus of attention on the central field of view of the driver’s head.

The gaze-surrogate estimation significantly outperforms the baseline. The extra information gained by monitoring the eye gaze on top of the head pose gives rise to much better accuracy. The full system gives the correct object of attention in nearly half the cases and with
Table 5.2: Performance of the attention estimator. Pedestrian accuracy shows how many pedestrian bounding boxes are correctly determined to be the attention point for the driver. Mean and median error are measures of how far the gaze-surrogate point is from the correct pedestrian bounding box.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Mean gaze error (in pixels)</th>
<th>Median gaze error (in pixels)</th>
<th>Attended pedestrian accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Center-bias based (baseline)</td>
<td>148.3</td>
<td>127.0</td>
<td>55.9</td>
</tr>
<tr>
<td>Proposed</td>
<td>54.1</td>
<td>32.2</td>
<td>79.4</td>
</tr>
</tbody>
</table>

Figure 5.7: Normalized error of surrogate gaze estimate on a continuous segment.
a relatively low median error. The attention estimation works better with a perfect pedestrian detector, enhancing the accuracy by 172% from 46.0% to 79.4%. Since it relies on detected pedestrian bounding boxes, the system will inevitably give the wrong output if the correct pedestrian is not detected - in that case the attention will simply be associated with the nearest detected bounding box. Implementing a perfect pedestrian detector is outside the scope of this paper, but the entire system would work with a different and better detector. It is very likely that tracking of pedestrians could improve the detection system by compensating for missed detections, but it is also worth to note that due to the, at times, rather extreme ego-motion of the driver’s head, this is not a trivial task.

Figure 5.8 shows the pixel error of the gaze-surrogate detection over a full test sequence and the system is almost universally better than the baseline, except in the few situations where the subject of attention is right in the middle of the field-of-view, where the baseline system is better by sheer coincidence.

5.3 Driver Attention Guard

Driver Attention Guard is a holistic distraction detection system. Distraction, like attention, does not have a single universally accepted definition. Many studies and authors have
put forth their versions which highlight different aspects of distraction [125, 81, 136]. Though, what is consistent among them is that distraction is a subset of inattention. For our work, we accept the following definition by Lee et al. [83]:

“Driver distraction is the diversion of attention away from activities critical for safe driving toward a competing activity.”

What is clear from this definition and also almost agreed universally is that distraction can significantly deteriorate driver performance. The above definition, by stating the presence of a competing activity, tries to separate other causes of inattention such as due to cognitive states (e.g. fatigue, drowsiness) that lead to diminished capacity to attend to the roadway. We encourage readers to refer to Appendix A for approaches to recognize affective states of a person. Such approaches related to the facial expression analysis can be applied for inferring driver’s cognitive states. Our current focus, though, is on distraction as defined above.

Diversion of attention away from driving direction can reduce the driver’s surround awareness (e.g. missed events [62], [85]) and impair decision making (e.g. increased reaction times [41], [60],[82]) or deteriorate performance of the driver (e.g. deteriorated control of the vehicle [27, 41, 62]), leading to an increased risk of crashes or near crashes. Single long glance and accumulated glance duration (caused by repeated short glances over a duration of a secondary task) both are associated with higher safety risk [166, 36, 78, 134], hence a distraction detection algorithm should take into account both such behaviors.

Ahlstrom et al. [3] investigated a distraction warning system which utilizes eye-gaze and when not available, head pose to detect the distraction event. A rule based system is used to regulate a fixed time length (2 sec) ‘attention buffer’ - an indicator of driver’s attentive state and the trigger for a warning. Although the study does not provide generalizable conclusions due to limited number of participants, it shows a positive trend towards improved visual behavior. The system, however, does not take the surround context into consideration. Simply relying on glance behavior without taking into account the surround context can lead to nuisance alerts. In fact, many other existing systems, such as forward collision warning (FCW) or lane departure warning (LDW), only take into account the immediate, unexpected conflicts. The major drawback of such systems are high rate of nuisance alerts. Moreover, situation criticality and hence the usefulness of the alerts are inherently subjective task that is likely to vary among drivers. Thus, it is needed for such systems to adapt to not only traffic situation and driver state but also driver preferences. Driver preferences (such as vehicle following distance, pedal profile etc.) becomes even more important for active assistance systems.

Based on above observations, a holistic integration of driver, vehicle and surround object states is proposed in the design of the Attention Guard system. Note that the purpose of this system is not for inattention detection caused due to fatigue or sleep but distraction detection due to e.g. momentary lapse of attention to forward driving direction at ‘wrong’ time or repeated
engagement with the in-vehicle systems such as infotainment etc. Also, our focus is on urban environments as they presents more vulnerable scenarios to driver distraction.

5.3.1 Testbed Design and Architecture

In order to develop the future Urban Intelligent Assist system, a complete vehicular context is required. This complete vehicular context includes the driver, vehicle surround and vehicle state. At the backbone of this research is the creation of a human-centered intelligent vehicle that captures this complete contextual data and interacts with its occupants to assist the driver in critical (stressful) situations that exist during urban driving.

The challenges in the design of such testbed lie in selection of sensor modalities and sensor placements. From commercial point of view, they need to be unobtrusive and should not compromise aesthetics. From the research point of view, the data need to be collected from on-road naturalistic driving conditions. Since they present the actual real-world scenarios and set realistic requirements for the development of a robust and reliable system. Hence the sensor placement should not alter normal driving behavior, for example, sensor obstructing driver’s view for safe maneuvering.

Hardware Architecture

Built on a 2011 Audi A8, the automotive testbed has been outfitted with extensive auxiliary sensing for the research and development of advanced driver assistance technologies. The goal of the testbed buildup is to provide a near-panoramic sensing field of view for experimental data capture. Figure 3.9 shows the different sensing modal as detailed below:

1. Internal Vision: Three cameras are placed monitoring driving head and foot behavior - two cameras on the A-pillar and near the rear view mirror observing drive’s head movements and another camera pointing down observing foot movements. They capture face view in color video stream at 30fps and foot view monochrome infrared stream at 25fps.

2. External Vision: For looking out at the road, the UIA experimental testbed features a single forward-looking camera, captured at 25Hz. In this study we use the camera for lane marker detection and lane tracking.

3. Radar: For tracking vehicles on the sides of the ego-vehicle, we employ two medium-range radars, which have been installed behind the rear-side panels on either side of the vehicle. The radars are able to detect and track vehicles as they overtake the ego vehicle on either side.

4. Lidar: The UIA testbed features two lidar sensors, one facing forward and one facing backward. We use these sensors for detecting and tracking vehicles, as well as detecting
obstacles such as guardrails and curbs. The lidars provide high fidelity sensor information, and are able to estimate parameters such as vehicle length, width, and orientation, as well as position and velocity.

5. GPS and Vehicle Dynamic Sensors: Vehicle state - steering and pedal profile are measured through CAN bus interface and for the precise vehicle localization a GPS with inertial measurement unit is installed in the back of the vehicle trunk.

Currently, the experimental testbed features robust computation in the form of a dedicated PC for development, which taps all the available data from the on-board vehicle systems. Sensor data from the radars and lidars are fused into a single object list, with complete object tracking and re-identification handled by sensor-fusion.

**Software architecture**

Large amount of data from above mentioned multi-modal multi-sensory network coming through different channels USB, CAN, FlexRay and Ethernet bus system, requires seamless synchronized capture. We utilize EB Assist ADTF\(^1\) (Automotive Data Time Triggered Framework) software development tool. ADTF provides a graphical user interface to deploy and connect various software components as shown in Figure 5.9. Besides this the framework offers tools for real-time data playback, data handling, processing and visualization in the lab as well in the test car.

\(^1\)http://automotive.elektrobit.com/home/driver-assistance-software/eb-assist-adtf.html
5.3.2 Proposed Approach

There are following three components of the system, continuously monitoring inside driver state and outside environment state along with ego vehicle parameters:

1. CoHMEt (Continuous Head Movement Estimator): A calibrated multi-camera setup is used for reliable and continuous head monitoring. Each perspective is processed independently to estimate head pose and the outputs are fused at later stage, as discussed in Chapter 3. Please refer to the chapter for algorithmic details and optimal camera position determination procedure.

2. Gaze Zone Estimation: Head pose and dynamics are then used to estimate driver’s gaze zone, as discussed in the previous Chapter 4, in particular, to determine whether the driver is paying attention to the front or not. Furthermore, a confidence measure to determine the quality of the estimate, is calculated based on the head pose tracking accuracy. This measure allow the system to determine whether or not to engage the assistance. We chose head pose based system as head pose estimation was robust and also less susceptible to eye glasses etc. One could easily design a system which could utilize HEIDy cues (as discussed in the previous chapter) and fall back to head pose alone if the confidence of the former is below desired value. We leave this as future work. In our experience, we found head pose based system was quite effective for the binary choices forward looking or not.

3. Attention Buffer Modulation: Attention buffer reflects driver’s attentive state in the driving task. Number of studies have shown that higher percentage of eyes-off-road duration, either due to single long glances or accumulated glance duration, is directly proportional to the increased likelihood of the occurrence of safety critical events. Hence, we use the total eyes-off-road duration (the time duration driver not looking road ahead) as a measure of distraction. We define a attention buffer which when becomes empty or 0%, driver is considered distracted. The buffer starts with 100% (or full buffer) and it is depleted when the driver’s gaze zone is not towards the forward driving direction. The depletion rate $m$ is dependent upon the front vehicle (if any) dynamics. Equations 5.2-5.3 provide the depletion rate $m$ of the buffer. When there is no front vehicle the buffer is decreased such that the driver has 2 sec worth of attention buffer. When the driver looks to the driving direction the buffer is increased with the same rate. The choice of 2 sec worth attention buffer is based on the suggestion of Alliance of Automobile Manufacturers (AAM) in designing HMI. When the distraction is detected, an assistance is provided by maintaining the ego-vehicle position within lane markings and following distance to the lead vehicle. Figure 5.10 illustrates an example of a possible buffer development scenario.
Figure 5.10: Example of a time trace illustrating the development of attention buffer.

\[ TTA = \frac{x - \Delta x}{u} - \frac{u}{2 \times a} \]  \hspace{1cm} (5.2)

\[ m = \frac{B}{TTA} \]  \hspace{1cm} (5.3)

Where, \( B \) is the current buffer level, \( TTA \) refers to time-to-activation, \( x \) is the headway distance (i.e. relative position of the front vehicle), \( u \) is the relative of the front vehicle, \( \Delta x \) and \( a \), are the driver’s preferred vehicle following distance and expected acceleration respectively. The last two quantities are learned for a particular driver (see the looking-out top-down component of FoViA framework discussed in the Chapter 2). This is an important consideration since the situation criticality is essentially traffic state depended as assessed by the driver itself. Hence, a assistance system must also take this into account. Also, by incorporating the driver preference and style, our hope is to make the system and the driver response (if he/she were alert) non-differentiable. This is in the direction of positive HMI, where driver is not overwhelmed by the frequent warnings but only when absolutely needed and/or assisted with a smooth active control.

5.3.3 Qualitative On-Road Evaluation

For testing the full functionality of the Attention Guard system, we require driver to be distracted, looking away from the driving direction. For obvious reasons, this is not done on public roads. A pilot test, however, is successfully conducted in a controlled test track environment equipped with lane markings and a lead vehicle to simulate front vehicle behavior. Figure 5.11 shows a snapshot of testing phase, when driver is not paying attention to the driving direction. Figure 5.12 shows what are seen by the vehicle sensors - the inside (driver) and the outside context (lead vehicle behavior). The Attention Guard system recognizes the distracted driver,
Figure 5.11: Controlled testing of Attention Guard system in Candlestick Park, San Francisco, CA. (a) Driver looking away from the forward driving direction and (b) ego-vehicle (white) coming to stop slowly while depending upon lead vehicle behavior.

the attention buffer going to zero (red signal), and activates the lane keep assistance and adaptive cruise control system for steering and braking assistance, needed for the distracted driver.

5.4 Concluding Remarks

We have introduced a novel approach to analyzing the attention state of a human subject, given cameras focused on the subject and their environment. In particular, we are interested in analyzing the focus of attention of a human driver. We presented a looking-in and looking-out framework combining gaze surrogate and object based saliency to determine the focus of attention. We evaluated our system in a naturalistic real-world driving data set with no scripted
Figure 5.12: Attention Guard: A functional illustration in a controlled experiment in a driving test track. The driver is interacting with infotainment system, not looking toward forward driving direction. The blue screen visualizes the surround context seen by the car - a lead vehicle (in red) and the ego-vehicle (in black). The depleted attention buffer (red signal) shows the distracted driver state with high confidence (blue signal).

experiments. This made the data set very challenging, but realistic. We showed that by combining driver state (using face analysis), we significantly improve the performance over a baseline system based on image saliency with center bias alone. The proposed framework circumvents the precise gaze estimation problem (a very challenging task in a real-world environment like driving) and hence, provide a robust approach for driver’s focus of attention estimation.

We introduced a distraction detection algorithm that adapts to the traffic situation as well as driver state and also, incorporates driver preferences (vehicle following distance and pedal profile). Further testing in different driving conditions, weather conditions as well as for longer period are needed to make the proposed prototype system, a commercial product.

To further increase the scope of such systems in understanding in-vehicle driver activities, we recommend to include other body parts analysis e.g. hand in conjunction with head analysis [93]. We also suggest to include driver behavior analysis for early intent prediction [112] for the future assistance and recommendation system.

Acknowledgements

Chapter 5 is in part a reprint of materials that are published in the IEEE Intelligent Transportation Systems Conference (2104), by Ashish Tawari, Andreas Mogelmose, Sujitha Martin, Thomas Moeslund and Mohan M. Trivedi, and the IEEE Intelligent Vehicles Symposium (2014), by Ashish Tawari, Sayanan Sivaraman, Mohan M. Trivedi, Trevor Shannon and Mario Tippelhofer. The dissertation author was the primary investigator and author of these papers.

The support for this research was graciously provided by the UC Discovery Program,
Audi and VW Electronic Research Laboratory. The author is grateful for the advice and support of the LISA colleagues.
Chapter 6

Conclusion

Robust systems for observing driver behavior will play a key role in the development of advanced driver assistance systems. In combination with environmental sensors, cars can be designed with the ability to supplement the drivers awareness, preempting and preventing hazardous situations. Analyzing the driver’s head and eye 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 which are informative of driver’s intent and other secondary tasks.

In this work, we presented a framework to infer driver’s visual focus of attention, and conversely, visual distraction based on head, eye and iris dynamics. In a driving environment, the driver is prone to make large spatial head movements during maneuvers such as lane changes, right/left turns etc. During these crucial moments, it is important to continuously and reliably track the head of the driver. We showed that many existing approaches for head pose estimation may work for 95% times but fails for those 5% corresponding to the large head deviation from the front driving direction which are of special interest to driver safety. We proposed a distributed camera framework which significantly outperformed monocular system. The system we propose satisfies all of our design criteria including wide operational range for continuous functioning, lighting- and person- invariant and partial occlusion tolerant for robust performance.

We argued that while precise gaze direction is desirable, coarse gaze direction is often sufficient for higher level semantic analysis pertaining to visual attention including distraction detection and attentional target determination. We showed that there exists tighter co-ordination between head pose and eye gaze during driving based on the analysis of the gaze zone estimation system utilizing head pose alone features. We further found that head pose dynamics are better indicator of coarse gaze direction to different zones relevant to the driving tasks.

Robustness requirement of the ADASs have made the use of head pose more practical and popular. However, eye gaze remains to be an important cue. Keeping in mind robustness requirement, we proposed gaze surrogate measures based on iris and eyelid region analysis. These
cues can be extracted robustly and accurately, and they provide further refinement to head pose based coarse direction. We showed that confusion between adjacent zones is significantly reduced by using proposed gaze surrogate features.

With the ability to robustly infer coarse gaze direction, we then incorporated environment analysis of salient objects to estimate focus of visual attention. In a Bayesian framework, we proposed task/context dependent saliency model. We learned saliency model using the traffic object patterns from the naturalistic driving data based on motion cues. We evaluated the system by its ability to determine the object of attention using coarse gaze (looking-in) and salient objects (looking-out) analysis. With our ability to understand visual focus of attention of the driver, we designed distraction detection system, Attention Guard, incorporating driver's state, driver's style and preferences, and surround traffic state. In a qualitative analysis of controlled field test, we showed the efficacy of the proposed real-time system in brake assistance. Further testing in different driving conditions, weather conditions, longer period as well as more participants are needed to make a qualitative analysis, we leave this as a part of future works.

The scope of the proposed systems is not limited to ADASs. We believe the systems and cues developed in this work can provide fundamental building blocks for variety of research areas, including human-factor, cognitive and psychological sciences to assist researchers in human behavior studies. However, significant challenge still remains in uncontrolled settings with constrained resources, and we encourage researchers to improve and extend the approaches described in this work to develop next generation technologies and systems for human interaction and safety.
Appendix A

Facial Expression Recognition by Cross Modal Analysis

A.1 Introduction and Motivation

Affective state plays a fundamental role in human interactions, influencing cognition, perception and even rational decision making. This fact has inspired the research field of “affective computing” which aims at enabling computers to recognize, interpret and simulate affects [122]. Such systems can contribute to human computer communication and to applications such as learning environment, entertainment, customer service, computer games, security/surveillance, educational software as well as in safety critical application such as driver monitoring [105, 40]. To make human-computer interaction (HCI) more natural and friendly, it would be beneficial to give computers the ability to recognize affects the same way a human does. Since speech and vision are the primary senses for human expression and perception, significant research effort has been focused on developing intelligent systems with audio and video interfaces [138].

Multimodal systems, specifically with audio and visual modalities, have shown several interesting interactions between the two modalities. For example, audio-visual speech recognition (AVSR), also known as automatic lipreading, or speechreading [108] aims at improving automatic speech recognition by exploring the visual modality of the speaker’s mouth region. Not surprisingly, it has outperformed audio alone ASR system particularly in noisy conditions. Similarly, the well known perceptual phenomenon, McGurk effect [89], which demonstrates an interaction between hearing and vision in speech perception. Furthermore, Munhall et. al. [101] suggests that rhythmic head movements are correlated with the pitch and amplitude of speaker’s voice and that visual information can improve speech intelligibility by 100% over that possible using auditory information only.
Figure A.1: An example of spontaneous conversation between driver and passenger during a driving task. Film strip shows samples of five images equally spaced in the utterance. First half of the utterance contains the speech and later half the road noise. Notice, however, that facial features are more expressive after speech content while head dynamics is concomitant with the speech [151].

In the field of affect recognition, there have been number of efforts to exploit audio-visual information as well and our framework can utilize these methods. However, above examples, where visual modality improves audio alone system, are motivated us to ask the fundamental question of how does audio modality influence visual perception, in particular, for the task of facial expression recognition. It is evident that speech generation influences facial expression. Also, for expression recognition the coupling between these two modalities is not so tight unlike the case in audio-visual speech recognition task.

Towards this end, we present a novel facial expression recognition framework using bi-modal information. Our framework explicitly models the cross-modality data correlation while allowing them to be treated as asynchronous streams. To recognize the key emotion of an image sequence, the proposed framework seeks to summarize the emotion using one single image derived from hundreds of frames contained in the video. We also show that the framework can improve the recognition performance while significantly reducing the computational cost by avoiding redundant or insignificant frame processing using auditory information.

A.2 Related Studies

Our long term goal is to study the cross-modal influence of the audio-visual data streams on each other for the affect recognition task. In this study, however, our focus is on face expres-
sion recognition. Hence we first discuss some of the representative works for facial expression recognition and then move our discussion on existing audio-visual affect recognition approaches to highlight the challenges lies in the integration of the two modalities. For an overview of audio only, visual only and audio-visual affect recognition, readers are encouraged to study a recent survey by Zeng et al. [182].

Because of the importance of face in emotion expression and perception, most of the vision-based affect recognition studies focus on facial expression analysis. A large amount of existing facial expression recognizers employ various pattern recognition approaches and are based on 2D spatiotemporal facial features: geometric features or appearance based features. Geometric-based approaches track the facial geometry information over time and classify expressions based on the deformation of facial feature [96]. Chang et al. [63] defined a set of points as the facial contour feature, and an Active Shape Model (ASM) is learned in a low dimensional space. Lucey et al. [88] employed Active Appearance Model (AAM)-derived representation while Valtar, Patras, and Pantic [164] tracked 20 fiducial facial points on raw video using a particle filter.

On the other hand, appearance-based approaches emphasize on describing the appearance of facial features and their dynamics. Zhao and Pietikaninen [185] employed the dynamic Local Binary Pattern (LBP) which is able to extract information along the time axis. Bartlett et al. [11] used a bank of Gabor wavelet filter to decompose the facial texture. More recently, Wu et al. [174] utilized Gabor Motion Energy Filters which is also able to capture the spatial-temporal information. Yang and Bhanu [176] created a single good image representation from a visual sequence by first registering the face image to a reference image using dense SIFT flow algorithm and extract appearance feature using Local Phase Quantization (LPQ). The method has provided the best overall emotion recognition performance till date for the GEMEP-FERA benchmark [165]. This can be derived as one of the special cases in our framework. It is important to mention that precise registration of frames is an important step otherwise single representation of image sequence using all the frames could suffer from large deviation of head pose.

Cohen et al. [23] performed expression classification in video sequence using temporal and static modeling by Naive-Bayes based (‘static’) and HMM based (‘dynamic’) classifiers respectively. Static classifiers outperformed dynamic ones. It is argued that dynamic classifiers are more complex, therefore they require more training samples and many more parameters to learn compared with the static approach. Author suggests that dynamic classifiers are more suited for person-dependent systems due to their higher sensitivity not only to changes in appearance of expressions among different individuals, but also to the differences in temporal patterns. Static classifiers are easier to train and implement, but when used on a continuous video sequence, they can be unreliable especially for frames that are not at the peak of an expression. This brings an important aspect of how to obtain a better and robust representation of an expression from video sequences. Can multimodality help in this regard? We seek to answer these questions.
Figure A.2: Overview of the proposed expression recognition system. Cross-Relevance feedback block provides the importance of the other modality at current time interval based on the analysis of its own modality. Frame-Relevance measure block can potentially utilize both Cross-Relevance feedback and its own modality analysis to finally assign the importance to the current frame. The solid blocks and connections shows the active components in the current implementation.

As far as automatic facial affect recognition is concerned, most of the existing efforts studied the expressions of the six basic emotions (happy, Sad, Surprise, Fear, Anger and Disgust) due to their universal properties and the availability of the relevant training and test material (e.g., [76]). These emotions are often deliberate and exaggerated displays [158]. The deliberate behavior, however, differs in visual appearance, audio profile, and timing from spontaneously occurring behavior [24, 132]. This has led the research field to new trends: analysis of spontaneous affective behavior and development of multimodal analysis. Multimodal analysis helps to improve the performance in challenging naturalistic setting during spontaneous behavior. Combining complementary information from the two streams can help improve the recognition performance. However, the two modalities are not tightly coupled in spontaneous naturalistic behavior as depicted in Figure A.1 [151]. Moreover, speech generation affects the facial expression dynamics. In following paragraphs, we present some of the works which address these two issues. In particular, how they derive various visual representations for visual channel as well as how they model asynchrony in the two streams.

One of the challenging tasks of the visual tracking systems is to deal with changes in the shape of the mouth caused due to speech. In order to deal with this situation, Datcu et al. [32] proposed a data fusion technique where they rely only on the visual data in the silent phase of the video sequence and the fused audiovisual data during non-silent segments. The visual modality during non-silent segments only focused on the upper half of the facial region to eliminate the effects caused by changes in the shape of the mouth. However, the results show that full face
based model performs superior than partial face. Hence an alternative strategy is require to filter out the influence of phonemes.

Wang et al. [169] proposed a relatively inexpensive computational method for visual based emotion recognition which selected a single key frame from each audio-visual sequence to represent the emotion present in the entire sequence. The criterion for selecting the key frames from the audio-visual sequences was based on the heuristic that peak emotions are displayed at the maximum audio intensities. The visual features are extracted from these key frames using Gabor wavelets. Acoustic features is then combine with derived visual feature at feature level data fusion scheme for the classification task. However, choosing one single frame from visual sequence is very restrictive and the same is clear from the performance of their visual alone system.

An important audio visual fusion scheme which aim at making use of the correlation between audio and visual data streams and relaxing the requirement of synchronization of these streams, is that of model-level fusion. Zeng et al. [181] presented a Multistream Fused HMM to build an optimal connection among multiple streams from audio and visual channels according to the maximum entropy and the maximum mutual information criterion. Author, however, considered tightly coupled HMMs. Song et al. [142] proposed an approach for multimodal emotion recognition which was specifically focused on temporal analysis of three sets of features: ‘audio only features’, ‘visual only features’ (upper half of facial region) and ‘visual speech features’ (lower half of facial region) using a triple HMM, i.e. one HMM for each of the information modes. This model was proposed to deal with state asynchrony of the audio-visual features while maintaining the original correlation of these features over time. On the other extreme is the model that allows complete asynchrony between the streams. This is, however, infeasible due to the exponential increase in the number of state combinations possible due to the asynchrony.

Our contribution in this paper is two folds: first, we explicitly model the correlation between the two streams while allowing them to be treated as asynchronous streams; second, we assign importance to a particular frame and there by avoiding extreme treatment (all the frames or just a single frame). More importantly this is accomplished by incorporating cross-modal models developed at the first step. The idea is that the analysis of the sequential changes can be beneficial for the facial expression recognition, however, the onset and the offset of the facial dynamics are hard to detect using visual alone modalities. Hence most of the efforts often tries to classify every frame and take a majority voting in the end to come up with single expression class. If the near apex frame or a set of more representative frames can be picked up based on multimodal data, to represent an entire segment, we can restrict noisy/redundant sequential facial feature deformations to negatively influence the recognition performance, and hence describe emotions in a reliable manner. Initial findings based on the mentioned proposition was reported in [154]. Here, we provide further in depth analysis by statistically substantiating claims and compare multi-class classification performances with existing literature.
A.3 Audio-Visual Data Association Approach

Figure A.2 sketches an overview of the proposed recognition system. The salient feature of our framework is the introduction of cross-modal relevance feedback and the frame relevance measure blocks. In our present work, we have limited our discussion to facial expression recognition using visual features alone. Hence classification module only utilizes visual features. An audio-visual classification framework, however, can easily be devised. Important point to note is that unlike any standard fusion schemes (early- model-level- or late-fusion), the proposed method attempts to improve signal representation at the first place hence by reducing error propagation which in general is harder to deal at later stages. Another simplification made in this work is to utilize only cross-modal feedback. This is to highlight importance of cross-modal information feedback in this context.

A detailed approach to summarize the visual expression information into a single image representation is presented in following sections. We show that just by incorporating cross-modal information a significant reduction in computation cost can be achieved. On the other hand by avoiding spurious frames for further processing and there by reducing unwanted influence (degradation in recognition performance), classification accuracy can be improved as discussed in section A.4.2.

A.3.1 Face tracking and alignment

In recent times, model-based techniques have been extensively used in nonrigid deformable object fitting. We use the Constraint Local Model (CLM) [130] for face tracking in the image sequences. CLM utilizes parameterized shape model to capture plausible deformation of landmark locations. It predicts the locations of the landmarks using a group of landmark detectors. In [130], the response map of these detectors is represented non-parametrically and the landmarks’ locations are optimized via subspace constrained meanshifts while enforcing their joint motion via shape model. It fits well to various poses. We used a person independent model which was trained on the Multi-PIE database [130]. The fitting process on an image $I_{m,n}$ provides a row vector $P^{(m,n)}$ for each sequence $m$ and frame $n$ containing $l = 66$ detected landmark positions

$$P^{(m,n)} = [x_1, y_1; x_2, y_2; \cdots; x_l, y_l]$$

The detected landmark is normalized by appropriate scaling, rotation and translation to make center of eyes 200 pixel apart and line joining the two centers horizontal. We denote the normalized shape vector as $P_{N}^{(m,n)}$. Furthermore, a reference shape is calculated using Eq. A.1.

$$P_{ref} = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{N_m} \sum_{n=1}^{N_m} P_{N}^{(m,n)}$$ (A.1)
Figure A.3: Reference shape derived from the database showing 66 landmark positions along with the ones in red which are used during image alignment process.

where $N_m$ is total number of frames in sequence $m$ and $M$ is the total number of image sequences. Given this reference shape $P^{ref}$, image $I^{(m,n)}$ is aligned using affine transform to obtain the aligned image $I^{(m,n)}_{align}$. For alignment, we only considered the points which are relatively stable to track corresponding to the eyebrows, eyes, nose and mouth regions. Figure A.3 shows the reference shape obtained for the database and the points used for image alignment. An example of automatically tracked face and the aligned face is illustrated in Figure A.4.

A.3.2 Visual sequence analysis: A Bimodal approach

Our goal in audio-visual sequence analysis is to provide a segment level classification. A video sequence, however, consists of hundreds of frames and the question is how to intelligently utilize all or a subset of frames to obtain a single image representation. For this, we propose to derive a weighted mean image $I^{rep}_m$ for the sequence $m$ which hopefully is representative of emotional content of the segment. As shown in Eq. A.2, aligned face image $I^{(m,n)}_{align}$ is weighted by the relevance measure $w(n)$ derived from audio signal analysis.

$$I^{rep}_m = \sum_{n=1}^{N_m} w(n) I^{(m,n)}_{align}$$  (A.2)

where $\sum_{n=1}^{N_m} w(n) = 1$.

We propose two rule based approaches to assign value of relevance measure. The first approach utilizes all the frames in the active video sequence, hence discarding any prosodic
Figure A.4: (a) An example of tracked face and the landmarks, and (b) aligned face image obtained using reference shape during image alignment step.

Figure A.5: Signal processing involved in calculation of single image representation of the image sequence. The bottom curves are intensity (dotted green - associated with the left axis) and pitch contour (solid blue associated with right axis) along with voiced region depicted by the red cross. The middle plot is the speech signal showing the segments chosen by the two schemes weighted mean in green and mean in red box. Finally the image sequence is shown next. All the plots and image sequence have the same time axis. (a) Weighted mean image derived for the expression class of happy and (b) mean image derived for the expression class of happy.

information available in audio stream. This is accomplished by assigning uniform weights $w(.)$ to all the frames. For fair comparison with the second approach, the active video sequence does not include preceding and trailing silence period otherwise it introduces irrelevant frames where
subject may not even be looking to the camera in the utilized database. We call the resultant image as the ‘mean image’.

The second approach uses prosodic information related to pitch and energy contour to choose only certain frames for the calculation of image $I_{rep}$. In particular, we use four sub-segments of the given video segment: two corresponding to start and end of the speech segment and two corresponding to maximum intensity and maximum pitch value. Each sub-segment is $200ms$ long, centered around mentioned events. All the select frames are assigned the same weight to derive the single image representation. We call the resultant image as ‘weighted mean image’. Figure A.5 shows signal processing involved in a typical example of the two approaches and their weighted mean and mean image output.

A.3.3 Appearance feature extraction

Originally proposed for texture analysis, the Local Binary Pattern (LBP) family of descriptors (LBP [114], LBP-TOP [185], LPQ [115] and LPQ-TOP [72]), in recent years, have been extensively used for static and temporal facial expression analysis, and face recognition. We use the blur insensitive LPQ (Local Phase Quantization) appearance descriptor proposed by Ojanisivu et al. [115] as the feature for facial expression analysis. LPQ is based on computing the short-term Fourier transform on local image window. At each pixel the local Fourier coefficients are computed for four frequency points: $[\alpha 0]$, $[0 \alpha]$, $[\alpha \alpha]$ and $[\alpha -\alpha]$, where $\alpha$ is sufficiently small number. We use $\alpha = 1/3$ in our experiment. Then phase information is recovered using binary scalar quantization of the signs of the real and imaginary part of each coefficient. The resultant eight bit binary coefficients are then represented as integers using binary coding. Finally, a histogram of these integer values from all image positions is composed and used as a 256-dimensional feature vector. We also use de-correlation process to eliminate the dependency of the neighboring pixels before quantization.

In our experiment, we resize the aligned face images to $200 \times 200$ and further divided into non-overlapping tiles of $10 \times 10$ to extract local pattern. Thus the LPQ feature vector is of dimension $256 \times 10 \times 10 = 25600$.

A.3.4 Auditory feature extraction

In our prior work [152, 150], we have used prosodic and spectral features to model emotional states. We used subset of these features for cross-modal relevance calculation in the proposed framework. In particular, the pitch and intensity (energy) contours are used to derive weights $w(n)$ for the $n^{th}$ frame in visual stream as described in Section A.3.2.

For pitch contour calculation, we used the auto-correlation algorithm similar to [120]. The input speech signal is divided into overlapping frames with shift intervals (difference between the starting point of consecutive frames) of $10ms$. Each frame is of $60ms$ long to be able to
span 3 periods of minimum pitch value (in our case 50Hz). Pitch candidate over each frame is calculated and a dynamic programming technique is used to get the final pitch contour. Log-energy coefficients are calculated using 30ms frames with shift interval of 10ms. Figure A.5 shows the interpolated pitch contour and voiced segment as well as the intensity contour.

A.4 Experimental Analysis

A.4.1 Audio-visual dataset

In our experiments, we used the audio-visual affective database eNTERFACE’05 [92]. It contains the six archetypal emotions: happiness (ha), sadness (sa), surprise (su), anger (an), disgust (di) and fear (fe). 42 subjects were asked to react to six different situations. The subjects were given five different answers to react to these situations. However, they were not given any instruction on how to express their emotions. Two human experts judged whether the reaction expressed the emotion in an unambiguous manner. If not, it was discarded. The database is collected in English language. Among the 42 subjects, 81% were men and remaining 19% were women. 31% of the total set wore glasses, while 17% of the subjects had a beard. The database is captured in a controlled recording environment.

A.4.2 Results and Discussion

In this section, we present results for two classification tasks: the first one involves binary-class classification experiments and the second involves multi-class classification experiments. The latter helps us to conduct comparative study with other publications available in the literature. While, the purpose of binary classification task is to bring forth the importance of bi-modal data association in facial expression recognition using visual sequence data. Also, binary classification analysis helps us gain better insight on, specifically, the impact of our proposed framework and generally, the inherent confusion between two classes. It is also worthy to note that many multi-class classification strategies inherently involve multiple binary classification (in our case too) and their performance is often ignored from discussion. Hence, we present the same in the following paragraphs.

We perform binary classification using Support Vector Machine (SVM) with linear kernel and default parameters available in Matlab implementation. We have 15 binary classification tasks corresponding to every possible pair of six expression classes. We present both, subject dependent and independent analyses.

For subject dependent analysis, we utilize 10 fold cross validation strategy. That is the database is randomly divided into 10 folds in stratified manner so that they contain approximately the same proportions of labels as the original database. The system is trained on 9 folds and tested on the left out fold. This is repeated 10 times each time leaving out a different fold. In the
end, we obtain a classification accuracy. We repeated the above procedure 10 times generating 10 accuracy figures for each of binary classification task. Mean accuracy is reported in Figure A.6a. For subject independent analysis, we employ Leave-One-Subject-Out (LOSO) cross validation strategy. That is the system is trained using the data associated with all the subjects but one and tested on the left out subject. This is repeated until every subject is kept as test subject. Figure A.6b shows average accuracy over all the subjects.

Firstly, it is clear that the use of single image representation can provide high recognition accuracy. In 10 fold cross validation, the best result is obtained for the Happy/Anger binary classification with accuracy over 95%. Certain classes are, however, more confusing in visual domain like the Sad/Fear or Sad/Surprise with recognition accuracy around 78%. As expected, subject independent results show lower performance but follow similar trend. Later in this section, we compare results for multi-class classification with others’ and show the superiority of the proposed framework. It is important to point out though that we have not used any tuning of SVM parameters nor have we used any feature selection technique which often improves the performance greatly. Our focus here is to compare the usefulness of auditory cross-modal feedback for frame selection which is also evident from the results.

Furthermore, to determine the statistical significance of the results, we performed an analysis of variance (ANOVA) to eliminate the effects of random fold selection on the results. The data under test are the binary classification accuracies over ten, 10-fold cross validation experiments for the two different schemes as described in the section A.3.2. The analysis allows judgment on whether the results were significantly different between mean image representation and weighted mean image representation. Figure A.6a shows the significance test scores for the 15 binary classifiers.

Using this analysis, we can conclude, in a statistically significant way ($p < 0.05$), that, exploiting audio association, in most cases, improves the classification accuracy. The best improvement of over 10% is obtained for binary classification task of Sad/Fear when the auditory signal is utilized for the frame selection step. Except in two cases Surprise/Disgust and Anger/Disgust, trend is opposite with accuracy lowered by $\sim 2\%$. A closer look on the results suggests that emotion classes Fear and Happy have shown the most improvements. On the other hand, emotion classes Disgust and Sad may have not been benefited. This can be due to our rule based weight assignment for these particular emotion classes. Particularly, for Sad class having low arousal profile, regions corresponding to high intensity and pitch may not provide representative frames. This encourages us to learn such bimodal association automatically from audio-visual data which we will pursue in our future efforts.

Also, notice that audio assisted approach utilizes maximum of $4 \times 200ms = 800ms$ worth of visual data corresponding to the four segments as described in Section A.3.2 while using all the frames on an average requires $2.5sec$ worth of visual frame processing. Hence using cross-modal information improves the visual computation cost by factor of $\sim 3$. 
Finally, we perform multi-class classification and compare the results with that of previously published ones. Both the proposed schemes show similar performances. We have already discussed the benefits of *Weighted-mean image* earlier; hence, we present the results using the same. For multi-class classification, we again used SVMs as classifiers. A linear kernel, pairwise multi-class discrimination and Sequential Minimal Optimization learning [172] are used. In literature, often, average unweighted accuracy is reported since it’s a better metric than weighted accuracy, specially in case of unbalanced dataset. Hence, we report and compare unweighted accuracy. However, since the database in use has approximately equal number of instances per class, unweighted and weighted accuracy are approximately same. For fair comparison, we have chosen the studies which have used the same database. We evaluated both subject independent and dependent methodologies.

For subject independent evaluation, we use Leave-One-Subject-Out (LOSO) strategy to ensure strict speaker independence. Table A.1 shows the confusion matrix for six class classification task. Unweighted average accuracy of over 47% is obtained. Paleari et al. [119] reported the best average accuracy of ∼32% for visual alone system using SVM and Neural Networks (NN) classifiers. In fact, the best audio-visual system using the Bayesian fusion showed 43% average accuracy.

<table>
<thead>
<tr>
<th>Reference Emotion</th>
<th>Recognized Emotion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>An</td>
</tr>
<tr>
<td>An (Anger)</td>
<td>49.8</td>
</tr>
<tr>
<td>Di (Disgust)</td>
<td>6.5</td>
</tr>
<tr>
<td>Fe (Fear)</td>
<td>10.7</td>
</tr>
<tr>
<td>Ha (Happy)</td>
<td>4.7</td>
</tr>
<tr>
<td>Sa (Sad)</td>
<td>19.1</td>
</tr>
<tr>
<td>Su (Surprise)</td>
<td>18.1</td>
</tr>
</tbody>
</table>

Unweighted Accuracy = 47.6%
Figure A.6: Binary-class classification accuracy for all the possible 15 different combinations over six basic emotions: happy (ha), Sad (sa), Surprise (Su), Fear (Fe), Anger (An) and Disgust (Di). (a) Subject dependent analysis, average accuracy of the ten experiments corresponding to each 10-fold cross-validation procedure and statistical significance test (one-way ANOVA). A $p$ value of less than 0.05 indicates that the two populations under test have significantly different means. Each point shows its mean and one standard deviation bar plot. Signs $^{**}$, $^*$, $^?$ and $^{	ext{ns}}$ corresponds to $p < 0.01$, $0.01 < p < 0.05$, $0.05 < p < 0.15$ and $0.15 < p$ or ‘not significant’ respectively. (b) Subject independent analysis using Leave-One-Subject-Out cross validation. Accuracy. A direct comparison, however, is difficult since their evaluation criterion is based on
60% training and 40% testing. We chose LOSO strategy for the ease of reproducibility and can serve as baseline for future reference.

For subject dependent analysis, Mansoorizadeh and Charkari [91] used ten fold cross validation using single subject data and reported recognition performance averaged over all the subjects. The average accuracy of 37% is achieved using visual along system while best bimodal system has 70% accuracy. Authors used SMV as classifiers. Similar procedure using our proposed framework achieves over 64% of average accuracy. However, given the small size of the database (5 emotion instances per subject per class), we suggest to perform a 10 fold randomized cross validation using all subjects data. Table A.2 shows the confusion matrix for six class classification task. The improved average accuracy (from 47%) of over 62% suggests the subject dependency on randomly chosen training and testing instances. Our visual alone system is significantly better than that of [91]; while lower than that of bimodal system, as expected.

In another study, by Gajšek et al. [51] speaker dependent information is decoupled from emotion instance at feature extraction stage. Their video based subsystem showed average accuracy of 54.6%. Authors used SVM as classifier. It’s not clear whether 5 fold cross validation used in evaluation procedure has different subjects in training and testing subsets. While there exist other studies that used the same database, they are either not clear on their evaluation criteria (subject dependent/independent) or many of them used just a part of the database or yet others do not report their single modal subsystem performance. We utilized the whole database and presented both subject dependent and independent performances. Based on the comparative study, our proposed framework has shown promising results.

A.5 Concluding Remarks

Automatic analysis of human affective behavior have been extensively studied in past several decades. Facial expression recognition systems, in particular, have matured to a level where automatic detection of small number of expressions in posed and controlled displays can be done with reasonably high accuracy. Detecting these expressions in less constrained settings during spontaneous behavior, however, is still a challenging problem [153]. In recent years, increasing number of efforts have been made to collect spontaneous behavior data in multiple modalities [149]. The research shift towards this direction suggests to utilize the multimodal data analysis approaches.

In this work, we presented a novel approach which explicitly models the cross-modal data association. We then investigated two different rule based data association approach. Our results showed that use of audio data could improve the recognition performance in terms of computation cost (since in general visual processing is costlier than audio processing) as well as recognition accuracy. Unlike various data fusion strategies, our approach attempts to better represent signal at feature extraction level by weighting frames by its importance based on cross-
Table A.2: Multi-class classification accuracy (in %):
Confusion table for randomized 10 fold cross validation

<table>
<thead>
<tr>
<th>Reference Emotion</th>
<th>Recognized Emotion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>An</td>
</tr>
<tr>
<td>An (Anger)</td>
<td>65.2</td>
</tr>
<tr>
<td>Di (Disgust)</td>
<td>6.5</td>
</tr>
<tr>
<td>Fe (Fear)</td>
<td>7.9</td>
</tr>
<tr>
<td>Ha (Happy)</td>
<td>0.5</td>
</tr>
<tr>
<td>Sa (Sad)</td>
<td>9.7</td>
</tr>
<tr>
<td>Su (Surprise)</td>
<td>10.2</td>
</tr>
</tbody>
</table>

Unweighted Accuracy = 62.9%

relevance feedback for the task at hand, in this case facial expression recognition.

We reported one-to-one binary classification results as well as that of multi-class classification. Best improvement in recognition accuracy for the binary classification task was over 10% and was statistically significant ($p < 0.01$). However, for some expression recognition task recognition accuracy was lowered by $\sim 2\%$ which suggests that the rules used might not be suitable for the particular emotion class (in our experiment disgust class did not show much of improvement). Comparative study of multi-class classification results show significant improvement over previously published approaches for visual subsystem for both subject dependent and independent analyses.

In our future efforts, we will explore data driven approach to learn cross-modal relevance measure. The ability of our framework to incorporate these models in early stages of signal processing has a great potential for robust recognition performance. We will also incorporate audio modality in classification module for the design of fully automatic audio-visual affect recognition system.
Acknowledgements

Appendix A is in full a reprint of material that is published in the IEEE Transactions on Multimedia (2013), by Ashish Tawari and Mohan M. Trivedi. The dissertation author is the primary investigator and author of the paper.

This research was supported by research grants from the National Science Foundation, the UC Discovery Program and Volkswagen. The author is grateful for the advice and support of colleagues from the CVRR laboratory.
Bibliography


99


101


[126] Xiaofeng Ren and Chunhui Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3137–3144, June 2010.


