UNIVERSITY OF CALIFORNIA, SAN DIEGO

Contextual Visual Object Recognition and Behavior Modeling for Human-Robot Interactivity

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

in

Electrical Engineering (Signal and Image Processing)

by

Eshed Ohn-Bar

Committee in charge:

Professor Mohan M. Trivedi, Chair
Professor Serge Belongie
Professor Garrison Cottrell
Professor Bhaskar Rao
Professor Nuno Vasconcelos

2017

The dissertation of Eshed Ohn-Bar is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2017

DEDICATION

To reducing car accidents and road traffic injuries.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Throughout this journey, I greatly appreciated the advice, support, and knowledge of my colleagues, friends, family, advisor, and members of the committee.

My passion for computer vision and robotics was ignited by taking a class with Mohan (Computer Vision and Multimodal System) in my first year as a Ph.D. student. Despite admitting to him of having no knowledge or experience in the field, he encouraged me to take this special class. Over the years, his continued encouragement and intense excitement has only increased my love for the field. The committee members, Prof. Serge Belongie, Prof. Garrison Cottrell, Prof. Bhaskar Rao, and Prof. Nuno Vasconcelos, have all had a similar contribution to my research and Ph.D. journey. They have shaped my view of the field throughout lectures, projects, and helpful discussions.

I was fortunate to have some great colleagues, including undergraduates, masters, fellow Ph.D. students in LISA, and supportive staff. I would like to thank academic staff, in particular Alice Carr, Jesse Martel, Mo Latimer, Gabrielle Coulousi, Crystal Liu, Karen Riggs-Saberton, and others who have always helped me with great attitude. I would like to thank Martha, who would often keep me company early mornings in the lab and perform the essential role of keeping our lab rooms clean, and the graduate advisors (Kacy Vega and Shana Slebioda) for their time and help.

I thank my colleagues, in particular Cuong Tran, Sayanan Sivaraman, Ashish Tawari, Sujitha Martin, Larry Ly, Kevan Yuen, Rakesh Rajaram, Akshay Rangesh, Sean Lee, Frankie Liu, Ravi Satzoda, Andreas Mogelmose, Miklas Kristoffersen, Jacob Dueholm, Alfredo Ramirez, and Nikhil Das, who immense help and great attitude were truly invaluable. Special thanks are required for the many hours donated by colleagues and friends who unconditionally volunteered to participate in experiments for me. I hope that one day I'll get to give back to each one of you in some way or another. I am also thankful for the support of our sponsors and industry partners (Toyota, KETI), who actively engaged in various elements of the research in our lab.

To my family (Mom, Dad, Ofek, Elifal, Beeri, Agam, Momo, Malia, Nilus), who supported me all along - you rock and I love you. To Richard - without your listening and support I would have probably quit the Ph.D. a long time ago.

Publication acknowledgements: Chapter 2 is in part a reprint of material that is published in the IEEE Transactions on Intelligent Vehicles (2016), by Eshed Ohn-Bar, and Mohan M. Trivedi. The dissertation author was the primary investigator and author of this paper.

Chapter 3 is in part a reprint of material that has been accepted for publication in the journal of Pattern Recognition (2016), by Eshed Ohn-Bar, and Mohan M. Trivedi. The dissertation author was the primary investigator and author of this paper.

Chapter 4 is in part a reprint of material that is published in the IEEE Transactions on Intelligent Transportation Systems (2014), by Eshed Ohn-Bar, and Mohan M. Trivedi. The dissertation author was the primary investigator and author of this paper.

Chapter 4 is in part a reprint of material that is published in the International Conference on

Pattern Recognition (2014), by Eshed Ohn-Bar, Sujitha Martin, Ashish Tawari, and Mohan M. Trivedi. The dissertation author was the primary investigator and author of this paper.

Chapter 4 is in part a reprint of material that is published in the IEEE Intelligent Transportation Systems Conference (2014), by Eshed Ohn-Bar, and Mohan M. Trivedi. The dissertation author was the primary investigator and author of this paper.

Chapter 5 is in part a reprint of material that is published in the journal of Computer Vision and Image Understanding (2015), by Eshed Ohn-Bar, Ashish Tawari, Sujitha Martin, and Mohan M. Trivedi. The dissertation author was the primary investigator and author of this paper.

Chapter 6 is in part a reprint of material that will be published in the journal of Pattern Recognition (2017), by Eshed Ohn-Bar, and Mohan M. Trivedi. The dissertation author was the primary investigator and author of this paper.

VITA

2010	B. S. in Mathematics, respectively, University of California, Los Angeles
2011	M. Ed. in Teaching, Urban Schools, and Social Justice, University of California, Los Angeles
2011-2017	Graduate Student Researcher, University of California, San Diego
2017	Ph. D. in Electrical Engineering (Signal and Image Processing), University of California, San Diego

PUBLICATIONS

Eshed Ohn-Bar and Mohan M. Trivedi, "Are All Objects Equal? Deep Spatio-Temporal Importance Prediction in Driving Videos", *Pattern Recognition*, 2017.

Rakesh Rajaram, Eshed Ohn-Bar and Mohan M. Trivedi, "Refining Deep Vehicle Detectors for Autonomous Driving", *under review, IEEE Transactions on Intelligent Vehicles*, 2016.

Eshed Ohn-Bar and Mohan M. Trivedi, "Looking at Humans in the Age of Self-Driving and Highly Automated Vehicles", *IEEE Transactions on Intelligent Vehicles, 1(1)*, 2016.

Eshed Ohn-Bar and Mohan M. Trivedi, "Multi-scale Volumes for Deep Object Detection and Localization", *Pattern Recognition*, 2016.

Rakesh N. Rajaram, Eshed Ohn-Bar and Mohan M. Trivedi, "Looking at Pedestrians at Different Scales: A Multiresolution Approach and Evaluations", *IEEE Transactions on Intelligent Transportation Systems*, 2016.

Akshay Rangesh, Eshed Ohn-Bar and Mohan M. Trivedi, "Long-term, Multi-Cue Tracking of Hands in Vehicles", *IEEE Transactions on Intelligent Transportation Systems, 17(5)*, 2016

Aida Khosroshahi, Eshed Ohn-Bar, and Mohan M. Trivedi, "Surround Vehicles Trajectory Analysis with Recurrent Neural Networks", *IEEE Intelligent Transportation Systems Conference*, 2016.

Rakesh N. Rajaram, Eshed Ohn-Bar, and Mohan M. Trivedi, "RefineNet: Iterative Refinement for Accurate Object Localization", *IEEE Intelligent Transportation Systems Conference*, 2016.

Rakesh N. Rajaram, Eshed Ohn-Bar, and Mohan M. Trivedi, "A Study of Vehicle Detector Generalization on US Highway", *IEEE Intelligent Transportation Systems Conference*, 2016.

Akshay Rangesh, Eshed Ohn-Bar, Kevan Yuen, and Mohan M. Trivedi, "Pedestrians and their Phones - Detecting Phone-based Activities of Pedestrians for Autonomous Vehicles", *IEEE Intelligent Transportation Systems Conference*, 2016.

Siddharth Siddharth, Akshay Rangesh, Eshed Ohn-Bar, and Mohan M. Trivedi, "Driver Hand Localization and Grasp Analysis: A Vision-based Real-time Approach", *IEEE Intelligent Transportation Systems Conference*, 2016.

Eshed Ohn-Bar and Mohan M. Trivedi, "What Makes an On-road Object Important?", *IEEE International Conference on Pattern Recognition*, 2016.

Eshed Ohn-Bar and Mohan M. Trivedi, "To Boost or Not to Boost? On the Limits of Boosted Trees for Object Detection", *IEEE International Conference on Pattern Recognition*, 2016.

Eshed Ohn-Bar and Mohan M. Trivedi, "Detection and Localization with Multi-scale Models", *IEEE International Conference on Pattern Recognition*, 2016.

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

Eshed Ohn-Bar and Mohan M. Trivedi, "Learning to Detect Vehicles by Clustering Appearance Patterns", *IEEE Transactions on Intelligent Transportation Systems, 16(5)*, 2015.

Eshed Ohn-Bar, Ashish Tawari, Sujitha Martin, and Mohan M. Trivedi, "On Surveillance for Safety Critical Events: In-Vehicle Video Networks for Predictive Driver Assistance Systems", *Computer Vision and Image Understanding, 134*, 2015.

Nikhil Das, Eshed Ohn-Bar, and Mohan M. Trivedi, "On Performance Evaluation of Driver Hand Detection Algorithms: Challenges, Dataset, and Metrics", *IEEE Intelligent Transportation Systems Conference*, 2015.

Rakesh N. Rajaram, Eshed Ohn-Bar, and Mohan M. Trivedi, "An Exploration of Why and When Pedestrian Detection Fails", *IEEE Intelligent Transportation Systems Conference*, 2015.

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

Eshed Ohn-Bar and Mohan M. Trivedi, "A Comparative Study of Color and Depth Features for Hand Gesture Recognition in Naturalistic Driving Settings", *IEEE Intelligent Vehicles Symposium*, 2015.

Eshed Ohn-Bar and Mohan M. Trivedi, "Can Appearance Patterns Improve Pedestrian Detection?", *IEEE Intelligent Vehicles Symposium*, 2015.

Eshed Ohn-Bar and Mohan M. Trivedi, "Hand Gesture Recognition in Real-Time for Automotive Interfaces: A Multimodal Vision-based Approach and Evaluations", *IEEE Transactions on Intelligent Transportation Systems, 15(6)*, 2014.

Eshed Ohn-Bar and Mohan M. Trivedi, "Beyond Just Keeping Hands on the Wheel: Towards Visual Interpretation of Driver Hand Motion Patterns", *IEEE Intelligent Transportation Systems Conference*, 2014.

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

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

Eshed Ohn-Bar and Mohan M. Trivedi, "Fast and Robust Object Detection Using Visual Subcategories", *IEEE Conference on Computer Vision and Pattern Recognition, Mobile Vision Workshop*, 2014.

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

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

Alfredo Ramirez and Eshed Ohn-Bar, "Go with the Flow: Improving Multi-View Vehicle Detection with Motion Cues", *IEEE International Conference on Pattern Recognition*, 2014.

Eshed Ohn-Bar, Sujitha Martin, and Mohan M. Trivedi, "Driver Hand Activity Analysis in Naturalistic Driving Studies: Issues, Algorithms and Experimental Studies", *Journal of Electronic Imaging: Special Section on Video Surveillance and Transportation Imaging Applications*, 2013.

Eshed Ohn-Bar and Mohan M. Trivedi, "Joint Angles Similarities and HOG^2 for Action Recognition", *IEEE Conference on Computer Vision and Pattern Recognition, Human Activity Understanding from 3D Data*, 2013.

Eshed Ohn-Bar and Mohan M. Trivedi, "The Power is in Your Hands: 3D Analysis of Hand Gestures in Naturalistic Video", *IEEE Conference on Computer Vision and Pattern Recognition, Analysis and Modeling of Faces and Gestures*, 2013.

Eshed Ohn-Bar, Sayanan Sivaraman, and Mohan M. Trivedi, "Partially Occluded Vehicle Recognition and Tracking in 3D", *IEEE Intelligent Vehicles Symposium*, 2013.

Eshed Ohn-Bar and Mohan M. Trivedi, "In-Vehicle Hand Activity Recognition Using Integration of Regions", *IEEE Intelligent Vehicles Symposium*, 2013.

Eshed Ohn-Bar and Mohan M. Trivedi, "Hand Gesture-based Visual User Interface for Infotainment", *Automotive User Interfaces and Interactive Vehicular Applications*, 2012.

ABSTRACT OF THE DISSERTATION

Contextual Visual Object Recognition and Behavior Modeling for Human-Robot Interactivity

by

Eshed Ohn-Bar

Doctor of Philosophy in Electrical Engineering (Signal and Image Processing)

University of California, San Diego, 2017

Professor Mohan M. Trivedi, Chair

Modeling spatio-temporal contextual information is fundamental in computer vision, with particular relevance to robotic intelligence and autonomous driving. We develop several frameworks for context modeling in image, video, multi-modal, and multi-cue data with applications to human-robot interactivity, in particular to the domain of intelligent vehicles. With the goal of developing contextual systems for interactivity, several key contributions are proposed: (1) A contextual framework for robust image-level scene understanding, including detection and localization of vehicles, pedestrians, and parts of humans (e.g. hands) in on-road setting, (2) A spatio-temporal, multi-modal, and multi-cue model which reasons over the complex interplay between the human (hand, head, and foot coordination), vehicle (speed, yaw-rate, etc.), and surround spatio-temporal context (agents, scene information) cues for understanding behavior and predicting activities, (3) A human-centric framework for object recognition and visual scene analysis, developed by studying a notion of object importance and relevance as measured in a spatio-temporal context of navigating a vehicle. The final contribution unifies the aforementioned

components of the thesis, including spatio-temporal object recognition, human perception modeling, and behavior and intent prediction into a single research task. Although the data and case studies in this work emphasize the safety-critical settings of navigating a vehicle, the contributions of this thesis are general and can therefore be applied to a wider array of applications involving human-machine interactivity.

Chapter 1

Contributions and Outline

The overarching aim of the thesis is developing machine learning and computer vision tools for human-robot interactivity, in particular for the application of intelligent vehicles. The thesis begins with and introduction and motivation in Chapter 2. Through an extensive survey, research challenges relevant to the theme of this thesis are identified. The following chapters are organized in terms of their semantic modeling level, starting from robust object detection and image-level contextual reasoning in Chapter 3, visual analysis of driver hand gestures in Chapter 4, and multi-cue driver behavior modeling in Chapter 5. The final research task, of spatio-temporal human-centric scene modeling in Chapter 6 leverages components from the previous chapters including object detection, spatio-temporal and multi-cue scene analysis, intent and activity prediction, perception modeling, and safe on-road navigation and planning. The thesis therefore begins (Chapter 2) and ends (Chapter 6) with a discussion on human-robot interactivity, while developing necessary research components in between chapters.

With the ultimate goal of developing better human-machine interactivity systems, we present the following contributions:

- We present an overview of a large corpus of related studies for studying human behavior and human-robot interactivity, particularly in the intelligent vehicles domain.
- We develop human-inspired algorithms for modeling visual contextual information. The proposed framework achieves improved generic object detection and localization performance.
- We develop a visual hand gesture understanding modules, capable of recognizing a variety of hand gestures in real-time. The dataset, methodology, and experimental insights allow for better design of human-robot interactivity systems and in-vehicle interfaces.
- We analyze the importance of multi-modal, multi-cue behavior understanding with a case study of modeling head, eye, and hand coordination. The study is further extended to a general, multi-cue fusion framework for understanding and predicting complex human activities and coordination. The framework shows promise on a real-world, driver action prediction task.

- We introduce a human-centric framework for object recognition by analyzing a notion of object importance, as measured in a spatio-temporal context of driving a vehicle. We find that various spatio-temporal cues are relevant for the importance classification task. Furthermore, we develop novel metrics in evaluating vision algorithms and dataset bias in applications where trust in automation is imperative and errors are costly.

Chapter 2

Introduction - Looking at Humans in the Age of Autonomous Robots

There is an unprecedented interest, activity, and excitement in the field of intelligent robots, and in particular of intelligent vehicles. In a great technological milestone, the culmination of research efforts of the past decades in a broad range of disciplines, including vehicle control, robotics, sensing, machine perception, navigation, mapping, machine learning, embedded systems, human-machine interactivity, and human factors, has realized practical and affordable systems for various automated features in automobiles [9]. This advancement is opening doors to possibilities only thought to be fictional a few decades ago.

Moving towards vehicles with higher autonomy opens new research avenues in dealing with learning, modeling, active control, perception of dynamic events, and novel architectures for distributed cognitive systems. Furthermore, these challenges must be addressed in a safety-time critical context. The exciting and expanding research frontiers raise additional questions regarding the ability of techniques to capture context in a holistic manner, handle many atypical scenarios and objects, perform analysis of fine-grained short-term and long-term activity information regarding observed agents, forecast activity events and make decisions while being surrounded by human agents, and interact with humans.

The aim of this chapter is to recognize the next set of research challenges to be addressed for achieving highly reliable, fail-safe, intelligent robots which can earn the trust of humans who would ultimately purchase and use these robots. This thesis studies the role of humans in the next generation of driver assistance and intelligent vehicles and robots in general. Understanding, modeling, and predicting human agents are discussed in three domains where humans and highly automated or self-driving vehicles interact: 1) inside the vehicle cabin, 2) around the vehicle, and 3) inside surrounding vehicles.

It is clear that automobile industry has made a firm commitment to support developments towards what can be seen as “disruptive” transformation of automobiles driven by human drivers to intelligent robots who transport humans on the roads. What will then be the role of humans in such a rapidly approaching future? Would they seat as passive occupants, who fully trust their vehicles? Would there

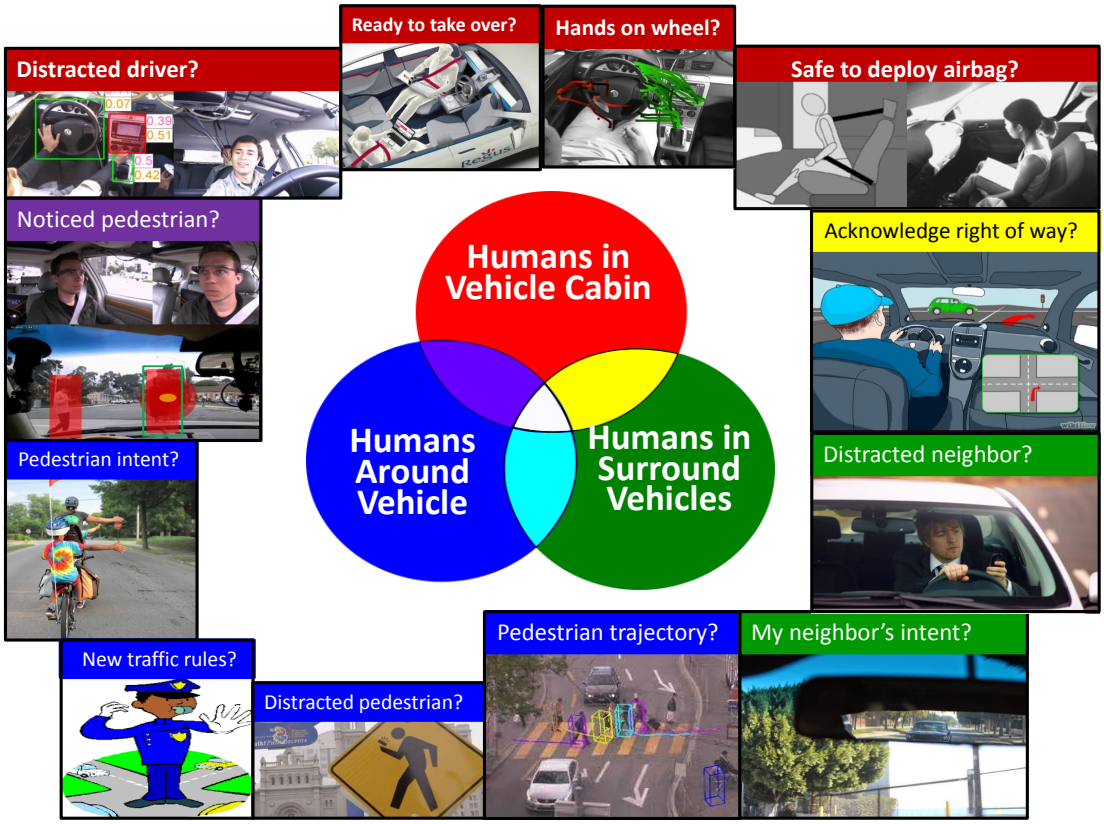


Figure 2.1: Intricate roles of humans to be considered in the development of highly automated and self-driving vehicles. For a safe and comfortable ride, intelligent vehicles must observe, understand, model, infer, and predict behavior of occupants inside the vehicle cabin, pedestrians around the vehicle, and humans in surrounding vehicles.

be a need for humans to "take over" control in some situations either triggered by the need perceived by the autonomous vehicle or desired by someone in the cabin? How should these autonomous vehicles interact with humans outside the vehicle (either as drivers of non-autonomous vehicles, pedestrians, emergency workers, etc.)? Because the future of intelligent vehicles lies in the collaboration of two intelligent systems, one robot and another human, this study aims to present core research ideas as they relate to humans in and around vehicles. In this collaboration of human and robot, the need for intelligent vehicles to observe, understand, model, infer and anticipate human behavior is necessary now more than ever.

There is an unprecedented interest, activity, and excitement in the field of intelligent vehicles. In a great technological milestone, the culmination of research efforts of the past decades in a broad range of disciplines, including vehicle control, robotics, sensing, machine perception, navigation, mapping, machine learning, embedded systems, human-machine interactivity, and human factors, has realized practical and affordable systems for various automated features in automobiles [9]. This advancement is opening doors to possibilities only thought to be fictional a few decades ago. The aim of this work is to recognize the next set of research challenges required to be addressed for achieving highly reliable,

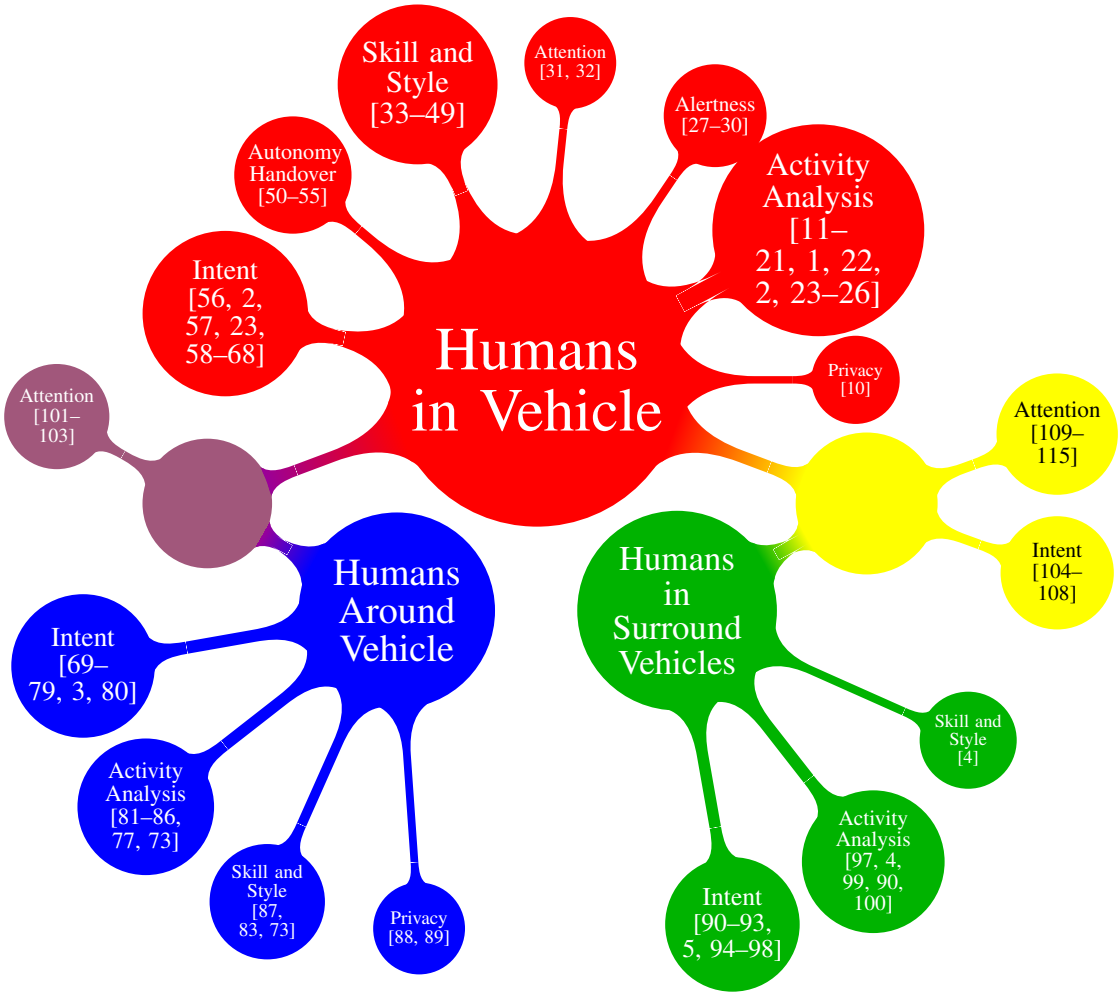


Figure 2.2: Trends in human-centric intelligent vehicle research. The figure visualizes related research studies discussed in this work as they relate to different semantic goals, from maneuver analysis and prediction, to style modeling. Each topic size is proportional the count of studies surveyed it contains.

fail-safe, intelligent vehicles which can earn the trust of humans who would ultimately purchase and use these vehicles.

It is clear that automobile industry has made a firm commitment to support developments towards what can be seen as "disruptive" transformation of automobiles driven by human drivers to intelligent robots who transport humans on the roads. What will then be the role of humans in such a rapidly approaching future? Would they seat as passive occupants, who fully trust their vehicles? Would there be a need for humans to "take over" control in some situations either triggered by the need perceived by the autonomous vehicle or desired by someone in the cabin? How should these autonomous vehicles interact with humans outside the vehicle (either as drivers of non-autonomous vehicles, pedestrians, emergency workers, etc.)? Because the future of intelligent vehicles lies in the collaboration of two intelligent systems, one robot and another human, this study aims to present core research ideas as they relate to

Table 2.1: Overview of human-centric related research studies by research goal and human-centric cues employed. Goal types follow Table 5.3, with [I] - intent and prediction, [Ac] - activity and behavior understanding, [D] - distraction and alertness, [At] - attention, and [S] - skill and style. VD refers to Vehicle Dynamics. PD refers to Pedestrian Dynamics (i.e. position, velocity).

Study	Type	Goal Detail	Cue Type
Jain et al. [116, 58], 2016	I	Lane Change Prediction	Head, Lane, VD, GPS, Map
Tran et al. [2], 2012	I,Ac	Brake	Foot, VD
Lefèvre et al. [56], 2011	I	Intent at Intersections	Map, VD
Molchanov et al. [1, 24], 2015	Ac	Secondary Tasks/Infotainment	Hand, Video
Ohn-Bar et al. [18] [23] 2014	Ac	Secondary Tasks/Infotainment	Head, Hand, Eye, Image
Tawari et al. [22] [32], 2014	Ac,At	Gaze Zone	Head, Eye
Toma et al. [11], 2012	Ac	Secondary Tasks/Phone	Head, Image
Ahlstrom et al. [20], 2012	Ac	Gaze Zone	Head, Eye
Cheng and Trivedi [25], 2010	Ac	Driver/Passenger Classification	Hand, Image
Vicente et al. [31], 2015	At	Gaze Zone	Head, Eye, Image
Liu et al. [30], 2015	D	Distraction Detection	Head, Eye
Jimnez et al. [28], 2012	D	Gaze Zone	Head, Eye
Wllmer et al. [27], 2011	D	Distraction Detection	Head
Lefèvre et al. [37], 2015	S	Style	VD
Schulz et al. [77, 78], 2015	I,Ac	Pedestrian Intent Recognition	PD, Head
Møgelmose et al. [74], 2015	I	Pedestrian Risk Estimation	PD, GPS, Map
Madrigal et al. [72], 2014	I	Intention-Aware Pedestrian Tracking	PD, Social Context
Kooij et al. [3], 2014	I	Pedestrian Path Prediction	PD, Head, Situation Criticality, Scene Layout
Quintero et al. [73], 2014	I,Ac,S	Pedestrian Path Prediction	PD, Body Pose, Subject Style
Goldhammer et al. [70, 83], 2014	I,S	Pedestrian Path and Gait Analysis	PD, Head
Pellegrini et al. [117], 2009	I	Pedestrian Path Prediction	PD, Social Context
Kooij et al. [81], 2016	Ac	Pedestrian Behavior Patterns	PD
Kataoka et al. [85], 2015	Ac	Pedestrian Activity Classification	PD, Video
Choi and Savarese [82], 2014	Ac	Pedestrian Activity Classification	PD, Social Context
Li et al. [90], 2016	I,Ac	Car Fluents	Video, Vehicle Part State
Laugier et al. [96], 2011	I	Behavior and Risk Assessment	VD, Lane, Turn Signal, GPS
Fröhlich et al. [5], 2014	I	Lane Change Intent	Turn Signal
Graf et al. [94], 2014	I	Turn Intent	VD, GPS, Map
Bahram et al. [108], 2016	I	Interaction-Aware Maneuver Prediction	VD, GPS, Map
Ohn-Bar et al. [106], 2015	I	Overtake and Brake Prediction	Head, Hand, Foot, VD, Lane
Jahangiri et al. [91], 2015	I	Intent to Run Redlight	VD, Scene Layout
Gindele et al. [93], 2013	I	Contextual Path Prediction	VD, Map, Lanes
Doshi et al. [105], 2011	I	Lane Change Forecasting	Head, Lane, VD
Aoude et al. [95], 2010	I	Threat Assessment	VD, GPS, Map, Lanes
Tawari et al. [115], 2014	At	Attention and Surround Criticality	Head, VD, Lane
Bar et al. [111], 2013	At	Seen/Missed Objects	Head, Eye, VD, Image
Mori et al. [112], 2012	At	Surround Awareness	Head, Eye, VD
Takagi et al. [114], 2011	At	Gaze Target	Head, Eye, VD
Doshi and Trivedi [109], 2010	At	Attention Focus	Head, Video
Phan et al. [101], 2014	At	Awareness of Pedestrians	VD
Tanishige et al. [102], 2014	At	Pedestrian Detectability	Head, Eye, PD, Video
Tawari et al. [103], 2014	At	Driver and Pedestrian Attention	Head, Eye, PD

Color codes:
Studying humans inside cabin.
Studying humans around vehicles.
Studying humans in surround vehicles.
Studying humans inside cabin and in surround vehicles.
Studying humans inside and around vehicles.

humans in and around vehicles. In this collaboration of human and robot, the need for intelligent vehicles to observe, understand, model, infer and anticipate human behavior is necessary now more than ever.

This thesis follows three main domains where humans and highly automated or self-driving vehicles interact (illustrated in Fig. 2.1):

- **Humans in vehicle cabin:** Whether the humans in the vehicle cabin are active drivers, passengers, or passive drivers, they may still be required to "take over" control in some situations triggered by the perceived need of the autonomous vehicle (for instance, under rare situations such as construction zones or police controlled intersections). In such situations, looking at the humans inside the vehicle cabin is necessary to access readiness to take over. If active drivers, are they distracted, did they pay attention to objects of interest (e.g. traffic signs, pedestrians), are they fatigued? If passengers, are they sitting properly (e.g. for proper airbag deployment in case of emergency), are

they giving directions, are they distracting the driver? If passive drivers, in the case of automated vehicles requiring take over at crucial moments, are they engaged in a secondary task, are their hands free, have they been alert to the changing driving environment?

- **Humans around the vehicle:** In addition to monitoring humans inside the vehicle cabin, observing humans in the vicinity of the intelligent vehicles is also essential for safe and smooth navigation. Because the road is shared with pedestrians, both an automobile driven by humans or intelligent robots who transport humans must be able to sense pedestrian intent and communicate with pedestrians. Where and how are humans around vehicle interacting with the vehicle? These include pedestrians, bike riders, skate boarders, traffic controllers, construction workers, emergency responders, etc. Are they in the path of the vehicle? Are they communicating their intent via body gestures? Are they distracted? Addressing such research issues can result in improved quality of navigation and assistance.

- **Humans in surrounding vehicles:** Intelligent vehicles must take into consideration humans in surrounding vehicles. Activity analysis and observation of intent applies to such humans as well, which operate under specific experience level, aggressiveness, style, age, distraction-level, etc. For instance, imagine two intelligent vehicles arriving at a stop-controlled intersection. In such a situation, both vehicles may be fully autonomous, only one of the vehicles may be fully autonomous, or both may be human-operated. Observing the humans by direct or indirect observation is necessary to acknowledge or give right of way. Are the humans in other vehicles driving in a risky manner? Is their behavior normal or abnormal? What will they do next, and what general and user-specific cues can be leveraged towards this identification? Are they acknowledging right of way at stop-controlled intersection? Are they engaged in secondary tasks, which motivates the ego-vehicle to avoid its vicinity?

We continue by providing an overview of relevant research studies. The studies are categorized in Section 5.6 for providing a highlight of the current research landscape. Section 5.6 studies emerging research topics in vision-based intelligent vehicles for each of the domains where humans and highly automated or self-driving vehicles interact. Section 2.2 follows with an analysis of the publicly available vision tools required for addressing the highlighted research issues. Finally, summary and conclusions are provided in Section 2.2.

2.1 Looking at Humans in and Around the Vehicle: Research Landscape and Accomplishments

The study of human-centric cues for driver assistance is an active research topic in intelligent vehicles, machine learning, and computer vision. Therefore, an extensive amount of work has been done in

Table 2.2: Overview of selected studies discussing different aspects of humans on the road. Methods are categorized according to task and whether humans were observed directly (e.g. body pose cues) or indirectly (e.g. pedal press, GPS/Map, vehicle trajectory).

Goal	Direct	Indirect
Intent and Prediction		
- In Vehicle	[2, 23, 58, 57]	[67, 61, 56, 59, 60, 63, 64, 62, 65, 66, 68]
- Around Vehicle	[69–79, 3, 80]	-
- Surrounding Vehicles	-	[95, 94, 96, 98, 5, 90–93]
- In+Surrounding Vehicles	[104–107]	[108, 97]
Activity		
- In Vehicle	[11–13, 18–20, 118, 21, 1, 55, 51, 22, 2, 23–26]	[14–17, 35]
- Around Vehicle	[81, 82, 84, 83, 85, 86, 77, 73]	-
- In Surrounding Vehicles	-	[90, 97, 100, 99]
Distraction and Alertness		
- In Vehicle	[27–30]	-
Attention		
- In Vehicle	[31, 32]	-
- In+Around Vehicle	[101–103]	-
- In+Surrounding Vehicles	[115, 109–112, 114]	-
Skill and Style		
- In Vehicle	[34]	[33, 35, 36, 119, 49, 37–48]
- Around Vehicle	[87, 83, 73]	-
- In Surrounding Vehicles	-	[4]

the field, from analysis of driver goals and intentions, human-machine interface design and customization, pedestrian activity classification, and up to identification of surrounding aggressive drivers (Fig. 2.1).

As means of identifying research trends, our first step is to give an overview of selected studies employing computer vision and machine learning techniques for intelligent vehicles applications. In order to maintain focus over the a large research landscape, the following approach for clustering research studies is pursued:

- **Domain clustering**: Throughout the chapter we partition the research space based on the three domains in Fig. 2.1, of humans inside the vehicle, around, and in surrounding vehicles. Although all three domains share the human agent, the domain-based clustering is useful because studies tend to focus on one of the three domains. From a vision perspective, methodologies and research goals among papers within the same domain tend to be more similar. Domain clustering also allows comparing and contrasting the domains in terms of what has been done and what has yet to be achieved.

- **Research goal clustering**: Related studies generally attempt to analyze, model, classify, and/or predict activities. This suggests a clustering based on the research task, whether humans inside or outside of a vehicle are concerned. We select seven types of overall research goals found in the surveyed studies. This clustering is employed for gaining a deeper understanding of the research landscape and discussing potential future research directions. Research goals include agent intent analysis and activity prediction (what will happen next?), attention model (where and what is the focus of the agent?), skill and style (what type of agent?), alertness and distraction (what is the state of the agent?), and general activity classification and behavior analysis (how is the agent operating?). Two additional goals not falling into the previous categories are autonomy handover and privacy-related tasks. We emphasize that the chosen research goals are closely related to each

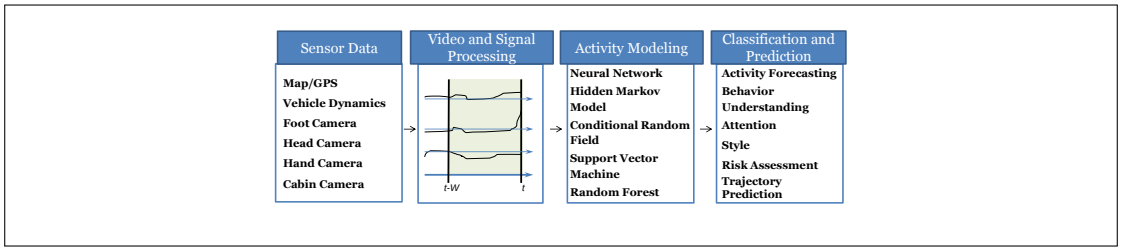


Figure 2.3: Overview of the sensing and learning pipeline commonly used to study humans in the cabin.

other and that there are other potential choices for research goal clustering [120]. Depending on the study, it may fall into one or multiple of the research goals. The research goals are consistent with topics in machine vision and learning-based studies as related to the type of data, methodologies, and metrics employed.

- **Cue type analysis**: A third type of analysis for highlighting trends in related studies can be made based on the type of cues employed in the study. We make a distinction between studies employing direct human-observing cues (e.g. body pose) and indirect cues (e.g. vehicle dynamics, GPS). This is shown in Table 5.3. Furthermore, we detail the specific type of cues employed by selected studies in Table 2.1, which complements the other two clustering techniques described above.

Fig. 2.2 shows a domain-based and research goal-based clustering of the papers listed in the corresponding Table 5.3. An emphasis is put on recent studies (mostly after 2008). In Fig. 2.2, the size of the node is proportional to the number of studies it contains. Fig. 2.2 can be used to draw several conclusions. We first identify trends, and then discuss further detail of the studies in each domain in the following sections (Section 2.1.1, 2.1.2, 2.1.3).

As might be expected, a large number of human-centric studies emphasize humans inside the vehicle. This domain also contains most of the diversity in terms of research goals, but research efforts are not distributed equally. A large number of behavior and activity analysis studies on driver gestures, secondary tasks, distraction, and maneuver classification and prediction have been performed. In-vehicle study of activities allows for a fine sensor resolution of the human agent, from vehicle dynamic sensors and up to eye and gaze analysis. The studies in this cluster still vary drastically in terms of the type of cues and vision techniques employed, as shown in Table 2.1. Certain research tasks, such as skill and style of humans, in-vehicle occupant interaction, and activity analysis of passengers, has seen less attention.

Fig. 2.2 allows for a high-level comparison between the domain of looking at humans inside the vehicle and the other two domains. Although human drivers can analyze fine-grained pose, style, and activity cues for identification of agent intent in all three domains (see Fig. 2.1), fine-grained semantic analysis around and in surrounding vehicles is still in early stages. Looking at humans around the vehicle commonly involves path prediction and to a lesser extent activity classification. Trajectory level path prediction is often done with little notion of skill, style, social cues, or distraction. Future improvement

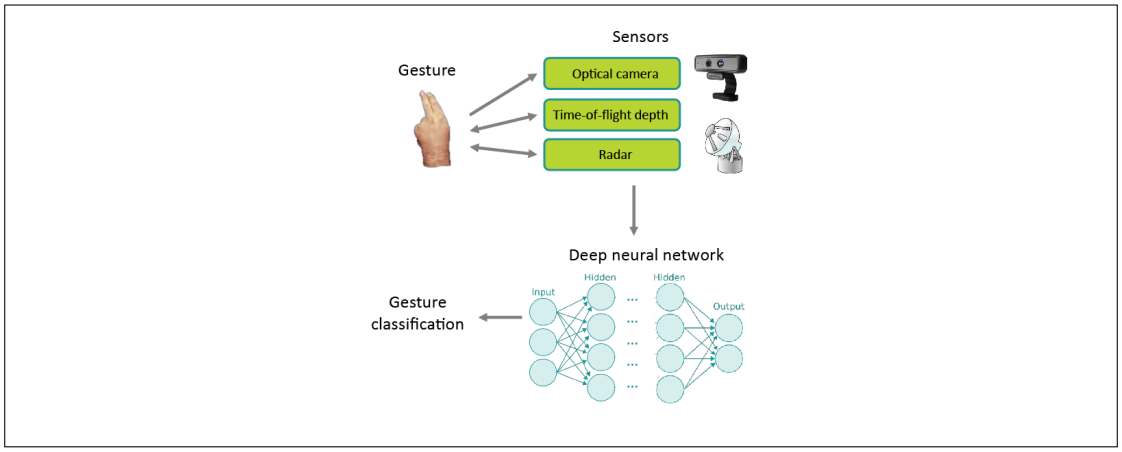


Figure 2.4: A multi-sensor driver gesture recognition system with a deep neural network [1].

in camera and sensing modalities would provide access to better and larger datasets. Consequently, we expect research tasks in the less studied two domains to become more diverse as in the looking inside the vehicle domain. Direct observation of humans in surrounding vehicles has not been done, although humans employ it everyday on the road.

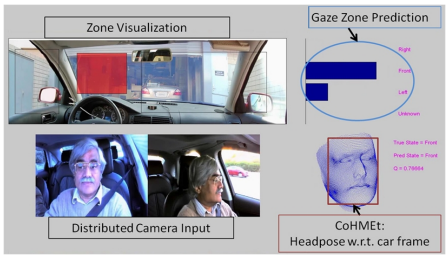
Another main conclusion that can be drawn relates to integrative schemes, which are also shown to be studied to a lesser extent. The studies are limited to attention-related studies as these reason over objects around the vehicle in order to infer surround awareness and gaze target. On the road, holistic understanding of both humans inside, around the ego-vehicle, and in surround vehicles is essential for effective driver assistance and higher vehicle autonomy. Holistic understanding of all three domains is a task performed by everyday human drivers while inferring intents, analyzing potential risk, and smoothly navigating a vehicle [122, 123]. Another relevant research topic is the modeling of social relationships among agents, which are employed by drivers in order to recognize and communicate intents. More specific examples can be found in Section 2.1.4.

Fig. 2.2 and Table 5.3 provide a high-level analysis of trends in related research studies within domains and research goals. Certain research goals are shown to be highly represented in one domain, but almost none existent in another. Nonetheless, even within a certain domain of human study, large variations exist in the types of cues employed for a specific task. Table 2.1 provides a closer look to the type of human-observing cues employed in the surveyed studies.

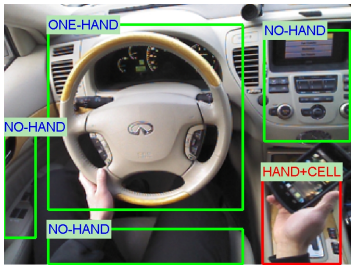
Next, we provide a deeper discussion for each domain as well as integrative frameworks below.

2.1.1 Looking at Humans in the Cabin

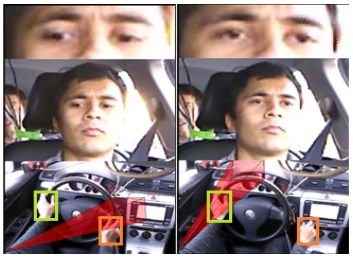
The surveyed papers in Fig. 2.2 show large diversity in terms of the research tasks for studying humans inside the vehicle. Further detail is provided in Table 2.1 in terms of study details and cue analyzed. A highlight of the research tasks is shown in Fig. 2.5, with an example research pipeline in



(a) Gaze zone classification using head cues [32].



(b) Object interaction analysis and secondary task classification with hand cues [121].

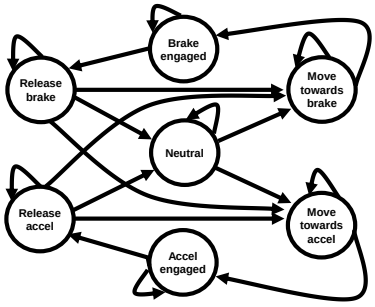


(c) Head, hand, and eye cue integration for secondary task activity analysis [118, 18].

(d) Cabin occupant activity and interaction analysis.

Figure 2.5: Emerging research topics for studying humans inside the vehicle.

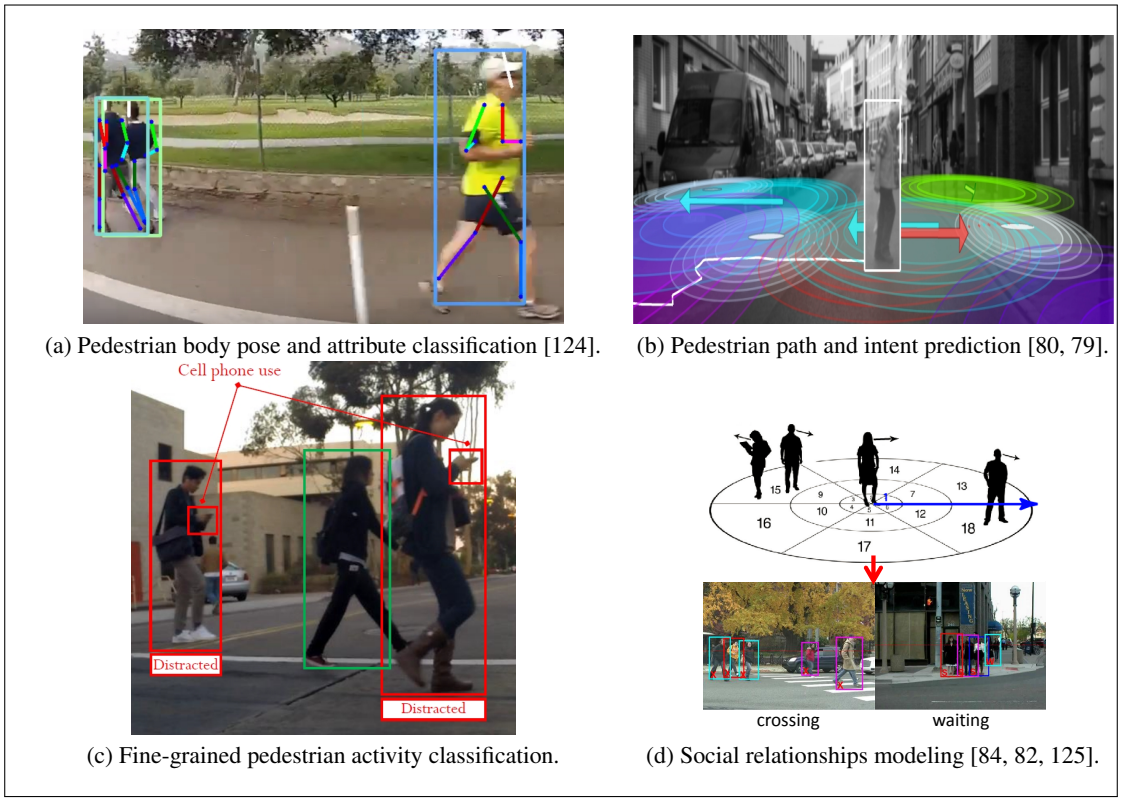
(a) Foot motion tracking.



(b) Activity state model.

Figure 2.6: Foot gesture recognition and prediction using a motion tracker and a temporal state model, such as a Hidden Markov Model [2].

Figs. 2.3 and 2.4. Dynamics of driver body pose, such as head [32], hand [23], eye [28], and foot [2] (Fig. 2.6) can be employed for in-cabin analysis of secondary tasks [11, 18, 31, 22, 20, 126, 127] and



(a) Pedestrian body pose and attribute classification [124].

(b) Pedestrian path and intent prediction [80, 79].

(c) Fine-grained pedestrian activity classification.

(d) Social relationships modeling [84, 82, 125].

Figure 2.7: Emerging research topics for studying people around the vehicle.

intent modeling and maneuver prediction [57, 23, 58, 107, 56]. Certain types of secondary tasks, such as gaze zone estimation and head gesture analysis, are more commonly studied than others, such as driver-object interaction (e.g. infotainment analysis [18] and cell-phone use [11]). Although passenger-related secondary tasks were shown to be critical for driver state monitoring from naturalistic driving studies [128], there are very few vision and learning studies on such tasks. Driver and passenger hand gesture and user identification have been studied in [25, 129, 130], but a large number of research tasks relating to interaction activity analysis has not been pursued. Fig. 2.5 highlights the need for the understanding and integration of multiple cues at different levels of representation. Such holistic modeling is essential for accurate, robust, and natural human-machine interaction. In particular, for studying humans in the cabin under semi-autonomy and control hand off [50, 52–54]. Depth sensors may also be used for improved activity recognition [121, 131–133].

Looking inside the vehicle often involves multiple types of on-board sensors in addition to a camera, such as vehicle dynamics [14–16, 38–40], phone [36, 41–43, 17, 44–48], or GPS [62, 66, 65, 33, 59, 60, 63, 64, 35]. These provide another useful modality for analyzing the behavior of humans inside the vehicle, such as skill and style recognition from inertial sensors [35]. Velocity, yaw-rate, and other vehicle parameters provide a signal useful for intent and maneuver recognition [59, 60, 63, 64]. GPS and map data can provide scene context (e.g. intersection vs. highway), strategic maneuver analysis [134, 135], or

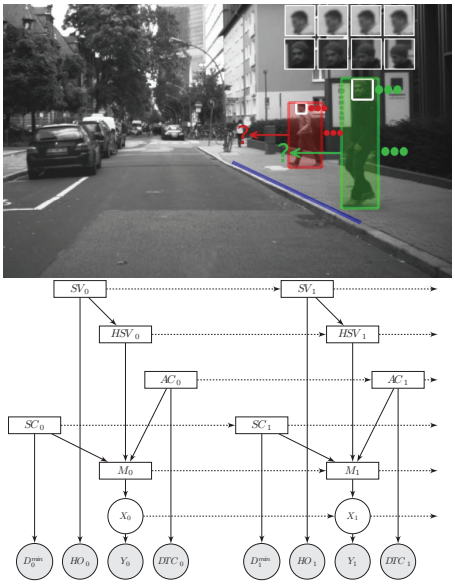


Figure 2.8: Pedestrian path prediction using a Dynamic Bayesian Network for incorporating contextual cues of pedestrian head orientation and situational awareness, situation criticality, and spatial layout cues [3].

be used in tactic and operation prediction models [136, 59]. In Liebner *et al.* [59] turn and stop maneuvers at intersections are predicted using GPS trajectories and a Bayesian Network for modeling driver intent.

2.1.2 Looking at Humans Around the Vehicle

Humans around the vehicle can be sensed with a variety of vision sensors, including color, thermal, and range sensors. Table 2.1 demonstrates a variety of research goals and cues employed to study pedestrians, with a highlight of research tasks shown in Fig. 2.7. The task of analyzing surround pedestrians is related to the heavily-studied visual surveillance tasks of scene and activity modeling [125]. In this work, we emphasize studies performed from movable platforms and leverage the specific geometrical and contextual cues induced by on-road settings. Here, scene information such as lane and road information can be combined with pedestrian detection and tracking for performing intent-aware path prediction and activity classification [81, 84, 82, 77, 78, 74, 72, 80, 3, 73, 70, 83]. Map information and vision-based pedestrian tracking are employed in [74] for risk estimation of pedestrians around a vehicle. Body pose and head pose cues can be used to infer pedestrian intent to cross and predict path [137, 138, 80, 3, 75, 139]. In Kooij *et al.* [3] pedestrian situation awareness (head orientation), distance-based situation criticality, and spatial layout (curb cues) are employed on top of a Switching Linear Dynamical System to anticipate pedestrian crossing (Fig. 2.8). Gait analysis using body pose for walking activity classification has been studied in [83, 85]. Spatio-temporal relationships between people

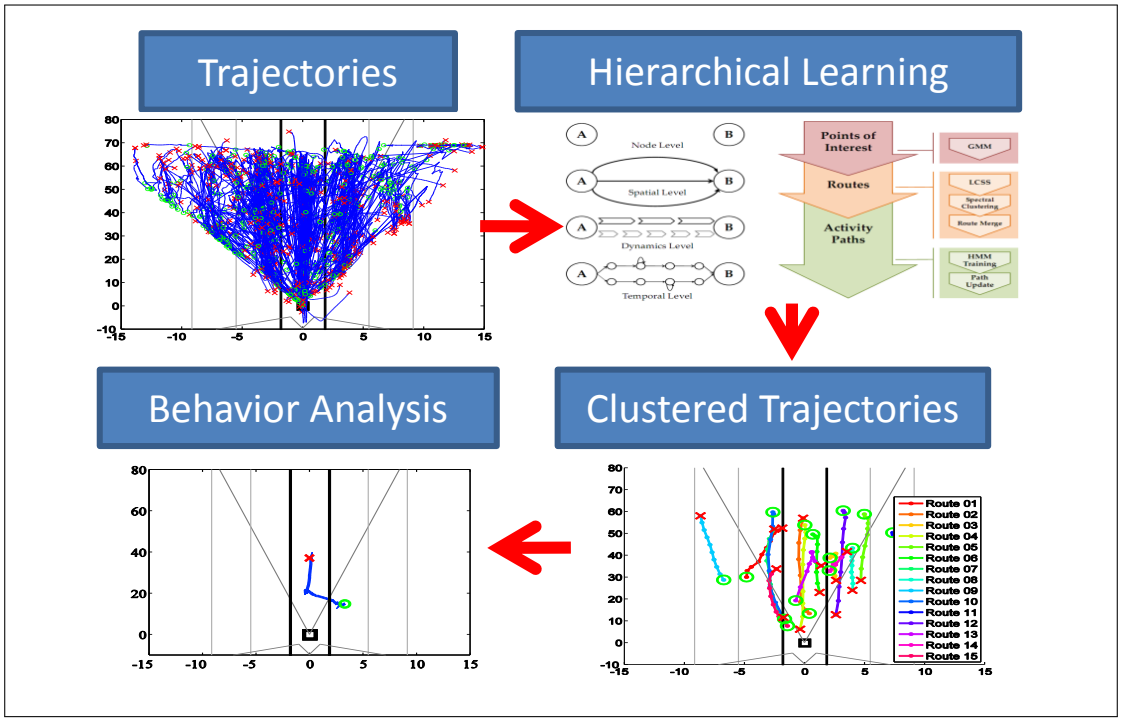


Figure 2.9: Activity analysis of people in surrounding vehicles. In [4], a hierarchical representation of the trajectory dynamics is used to perform behavior analysis of vehicle motion patterns. A Hidden Markov Model is used to perform trajectory classification and detect abnormal trajectory events.

have been incorporated in [84] for activity classification. As shown in Table 5.3, finer-grained semantic analysis of skill, style, attention, distraction, and social interaction inference of people around the vehicle is in its early stages. Several recent naturalistic driving datasets with additional modalities, fine-grained attribute and pose information [140–143] will help to further push the richness of analysis provided by algorithms looking at humans around the vehicle. Increased resolution of the sensing modules will play a key role in advances for intricate analysis of pedestrian state, intent, and social relationship modeling [84, 125]. Because smooth and safe driving often involves navigation around humans (e.g. construction zones) and interaction with pedestrians (Fig. 2.7 depicts some of the relevant research tasks), this domain of human analysis for intelligent vehicles is expected to have high research and commercial activity.

2.1.3 Looking at Humans in Surround Vehicles

Understanding intent of drivers in surround vehicles, a task continuously performed by human drivers, is also useful for machine drivers. The research tasks are therefore shared across the three domains of humans in intelligent vehicles. When looking at humans in surround vehicles, vision-based algorithms can be applied to understand behavior and intent, predict maneuvers, and recognize skill, style, and attention.

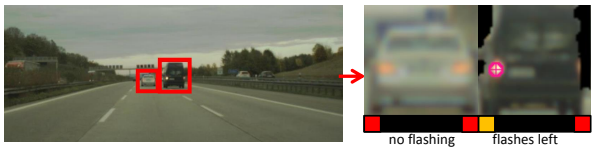
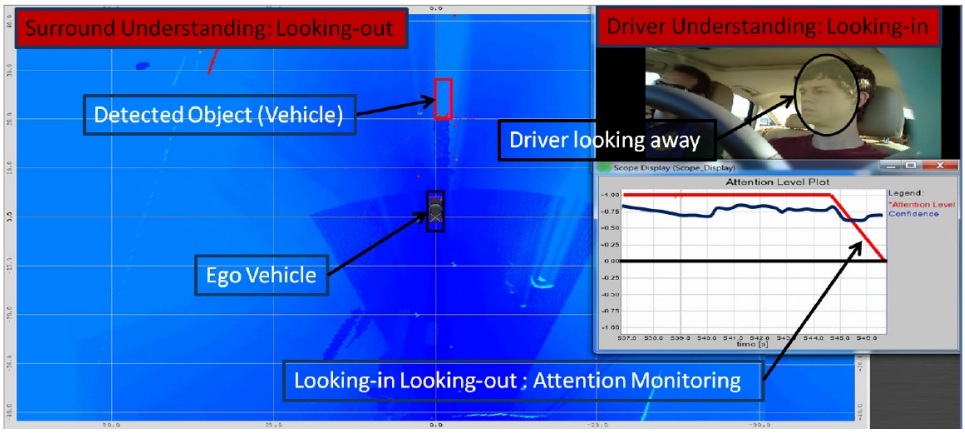
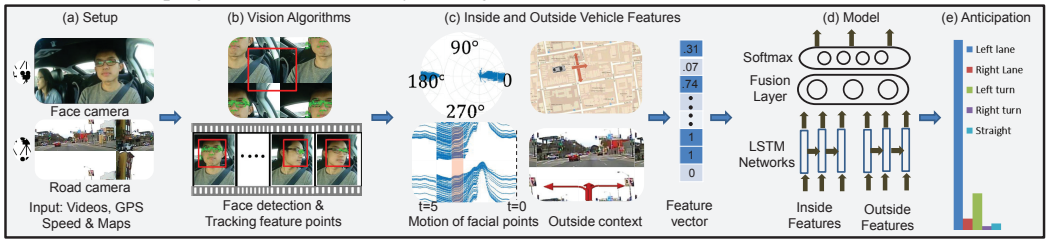


Figure 2.10: Intent detection using turn signal analysis [5]. First, vehicles are detected and tracked using a Mixture-of-Experts model and a Kanade-Lucas-Tomasi tracker. Consequently, light spots are detected, and classification of events is performed with an AdaBoost classifier over frequency-domain features.



(a) Helping a distracted driver by sensing situational need and driver alertness levels [115].



(b) Lane change maneuver prediction using driver and surround cues [116]. A Recurrent Neural Network with Long Short-Term Memory (LSTM) units is employed to fuse cue modalities and capture temporal dependencies.

Figure 2.11: Emerging research topics in integrative frameworks for on-road activity analysis.

Understanding activity and modeling intent of other vehicles is widely researched for path prediction and activity classification [91–93, 144]. Intent modeling is a critical step towards risk assessment [94–97, 62]. Lefèvre *et al.* [61] employs a Dynamic Bayesian model over spatial layout and vehicles state (position, orientation, and speed) cues for detecting conflicting intentions and estimating risk at intersections. In Zhang *et al.* [100], a generative model for modeling traffic patterns at intersections is proposed using vehicle trajectory, orientation, and scene cues. Sivaraman *et al.* [4] proposes learning

trajectory patterns of surround vehicles with a hierarchical representation of trajectory dynamics and a Hidden Markov Model. The trajectory patterns are employed for surround vehicles behavior analysis, including detection of abnormal events. Detection of turn signals [98, 5, 90] is also useful in understanding the intent of humans in surround vehicles (Fig. 2.10). In Fröhlich *et al.* [5], vehicles are detected using a Mixture-of-Experts model and tracked with a Kanade-Lucas-Tomasi tracker. After background segmentation and light spot detection, an AdaBoost classifier is employed over frequency-domain features for performing turn signal analysis. Because predicting intents of other vehicles is crucial to safe driving, a robotic driving system should capture subtle cues of aggressiveness, skill, style, attention, and distraction of humans in surround vehicles. It is known that age, gender, and other properties of the human driver influence driver behavior [91], so that vision-based observation of humans in other vehicles (e.g. body pose cues, preparatory movement of other drivers, age classification, etc.) can be useful when working towards aforementioned research tasks.

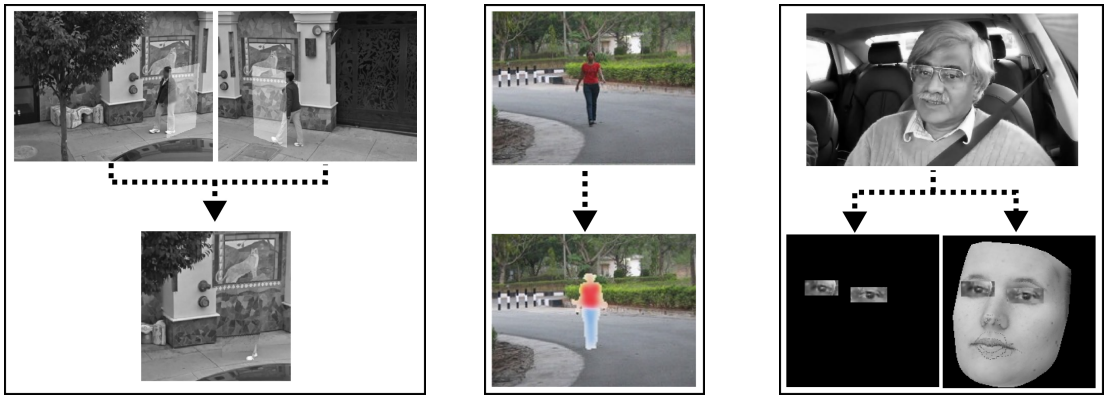
2.1.4 Integrative Frameworks

On the road, humans inside vehicles, around vehicles, and in surround vehicles all interact together. Therefore, intelligent vehicles are vehicles that can integrate information coming from multiple domains for better scene understanding and improved forecasting [145]. Holistic understanding is useful for effective and appropriately engaged driver assistance system, successful human-robot communication, and autonomous driving. Example integrative systems are shown in Fig. 2.9.

As drivers interact with their surrounding continuously, driver activities are often related to surrounding agent cues (e.g. other vehicles and pedestrians). Maneuver prediction [105–107, 146] often requires integrating surround and cabin cues for an improved model of the driver state and consequently better early event detection with lower false positive rates. In Ohn-Bar *et al.* [106], both driver observing cues (head, hand, and foot) and surround agent cues (distance and locations to other vehicles) are integrated with Multiple Kernel Learning to identify intent of the ego-vehicle driver to overtake. Driver attention estimation is another common research theme in integrative frameworks, where driver cues and surround object cues, such as pedestrian detection [103] or salient objects [109], are integrated to estimate attentiveness to surround objects. In Tawari *et al.* [115], situational need assessment and driver alertness levels are employed as cues for an assistive braking system (Fig. 2.11). Jain *et al.* [116] employs multi-modal Long Short-Term Memory networks for maneuver anticipation.

2.2 Naturalistic Datasets and Analysis Tools

The survey of related research studies in Section 5.6 captured the research landscape in terms of what has been done, and what still needs to be done. As in all science and engineering fields, a key component in future research relies on access to naturalistic, high-quality, large datasets which can provide insights into better algorithmic and system designs. Studying user-specific nuances and achieving better



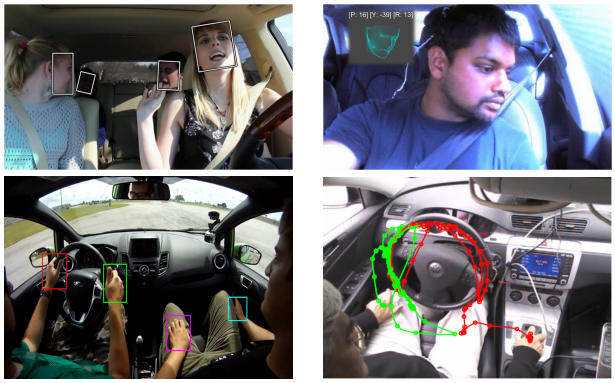
(a) Preserving scene [88] (b) Preserving action [89] (c) Preserving gaze direction [10]

Figure 2.12: Comparison of selected works in de-identification from different applications: (a) Google street view: removing pedestrians and preserving scene using multiple views, (b) Surveillance: Obscuring identity of actor and preserving action and (c) Intelligent vehicles: Protecting driver's identity and preserving driver's gaze.

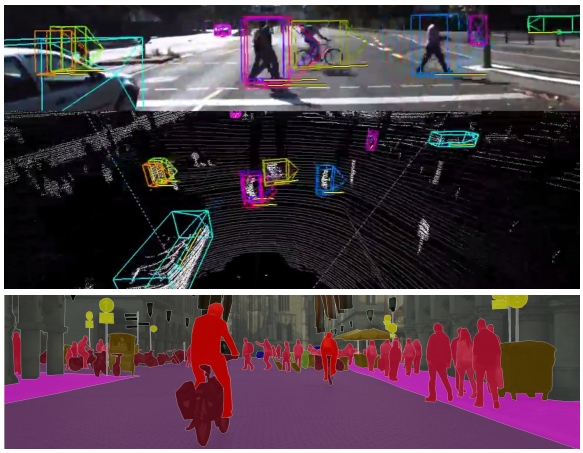
Table 2.3: Overview of selected publicly available naturalistic datasets from a mobile vehicle platform.

Dataset	Description
Studying humans inside cabin	
VIVA-Hands [147, 121] (2014)	Detection, tracking, and gestures of driver and passenger hands in video.
VIVA-Faces [148] (2014)	Detection and pose estimation of in-vehicle occupants' faces.
Studying humans inside cabin and in surround vehicles.	
Brain4Cars [58]	Lane change maneuver prediction with cabin-view camera, scene-view camera, GPS, and vehicle dynamics.
Studying humans around vehicles.	
Caltech [140] (2015)	Body pose and fine-grained classification of pedestrians, including age, gender, and activity.
Studying surround vehicles and humans around vehicles.	
KITTI [141] (2012)	Vehicle and pedestrian 3D tracklets annotated with stereo imagery, GPS, lidar, and vehicle dynamics.
Cityscapes [149] (2015)	On-road object segmentation with stereo video, vehicle dynamics, and GPS.

situational awareness in autonomous systems all require standardized metrics and benchmarks. Furthermore, data accessibility issues are a main reason why integrative frameworks are still little developed and understood on a principled manner. We therefore mention current tools and datasets available to the scientific community for the study of humans in and around vehicles. The discussion further raises issues as to requirements for further progress in the field.



(a) Naturalistic datasets for looking inside the vehicle for face and hand analysis (VIVA dataset).



(b) Naturalistic datasets for looking at humans around a vehicle and in other vehicles (top - KITTI, bottom - Cityscapes).

Figure 2.13: Example images from publicly available datasets (Table 2.3) for analysis of humans inside and outside of the vehicle.

2.2.1 Towards Privacy Protecting Safety Systems

The development of intelligent vehicles requires careful consideration of safety and security of people in and around the vehicle. This article has touched upon the fundamentals needed to deal with safety issues but as naturalistic datasets are developed there are important questions about security and identity.

There is a trade-off between privacy and extracting driver behavior. Many existing state-of-the-art algorithms on driver behavior are able to achieve their purpose due to analysis of raw signal and video input, with possible privacy implications. Privacy preserving considerations may play a role in the construction of publicly available large-scale datasets, especially as current state-of-the-art algorithms for intelligent vehicles require large amounts of data for training and evaluation. Therefore, as a community,

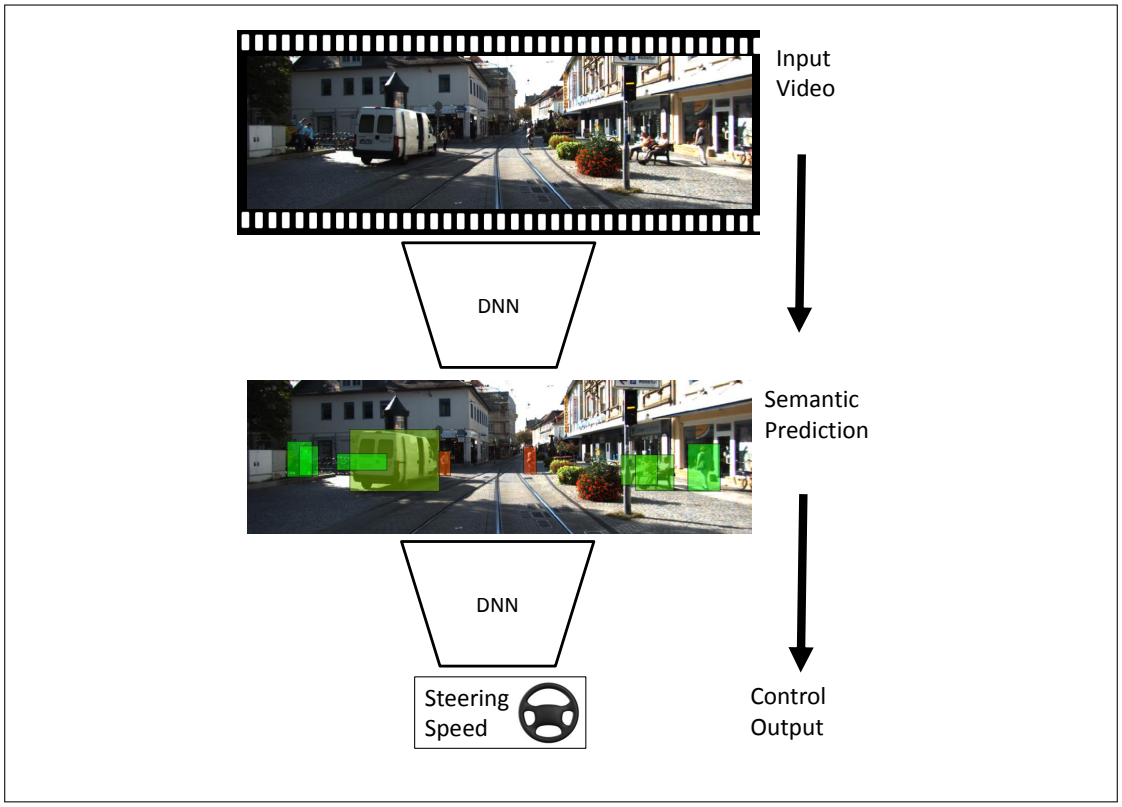


Figure 2.14: Example video-to-control policy pipeline (mediated-semantic perception [6, 7]) with deep networks (DNN), where initial prediction of semantic scene elements is followed by a control policy algorithm.

it is important to raise the standards of both safety and security in the development on intelligent vehicles.

2.2.2 Naturalistic Driving Datasets

Table 2.3 lists recent datasets which are publicly available for the study of humans inside and around the vehicle. As can be seen, only a handful of such standardized datasets currently exist. Because pedestrian detection and tracking is a well-studied problem, such tasks have several publicly available benchmarks, including Caltech pedestrians [150], Daimler [151], KITTI [141], and Cityscapes [149, 152]. The Caltech roadside pedestrians dataset [140] includes body pose and fine-grained pedestrian attribute information. Other datasets are not generally captured in driving settings (e.g. surveillance applications [153], static camera [84], and stroller or hand-held camera [154–156]).

The datasets are visualized in Fig. 2.13, demonstrating the progress that has been made in the field so far. Face and hand detection and analysis can now be measured in harsh occlusion and illumination settings in the vehicle. Similarly, challenging datasets observing surround agents continuously push the field further with comparative evaluations. As can be seen in Fig. 2.13, the majority of the dataset emphasizes basic vision tasks of detection, segmentation, or pose estimation. On exception is the Brain4Cars

dataset [58] which provides annotations for activity anticipation. As methods further progress on such recent benchmarks, additional higher-level semantic tasks such as activity understanding and forecasting could be introduced and evaluated.

2.3 Chapter Concluding Remarks

Intelligent vehicles are at the core of transforming how people and goods are transported. As technology takes a step closer towards self-driving with recent advances in machine sensing, learning, and planning, many issues are still left unresolved. In particular, we highlight research tasks as they relate to the understanding of human agents which interact with the automated vehicle. Self-driving and highly automated vehicles are required to navigate smoothly while avoiding obstacles and understanding high levels of scene semantics. For achieving such goals, further developments in perception (e.g. driveable paths), 3D scene understanding, and policy planning are needed. The current surge of interest in intelligent vehicle technologies is related to recent progress and increased maturity in image recognition techniques [157–160] and, in particular, to the successful application of deep learning to image and signal recognition tasks [161–165]. Deep temporal reasoning approaches [166, 116] have also shown similarly impressive performance, and are useful for a variety of learning tasks (e.g. distraction detection [27]). Furthermore, control policy for self-driving, both mediated-semantic perception approaches [6] and behavior reflex, end-to-end, image to control space approaches [167–175] (e.g. Fig. 2.14) have been making major strides. The exciting and expanding research frontiers raise additional questions regarding the ability of techniques to capture context in a holistic manner, handle many atypical scenarios and objects, perform analysis of fine-grained short-term and long-term activity information regarding observed agents, forecast activity events and make decisions while being surrounded by human agents, and interact with humans.

This chapter is in part a reprint of material that is published in the IEEE Transactions on Intelligent Vehicles (2016), by Eshed Ohn-Bar, and Mohan M. Trivedi. The dissertation author was the primary investigator and author of this paper.

Chapter 3

Modeling Image Context for Object Detection and Localization

This chapter develops a technique for modeling image-level contextual cues for robust object detection and localization. The method is inspired by the multi-level, multi-scale information found in natural or human-made settings, as well as the expertise of humans in leveraging such information when understanding a scene and making decisions. Hence, the spatial context module will repeat throughout the thesis across research tasks. The role of spatial context is initially studied for the task of object detection and localization in images. Robust detection and localization can consequently be used to perform vision-based analysis of agents' behavior, which will be discussed in consequent chapters.

3.1 Modeling Multi-scale Spatial Context

This study aims to analyze the benefits of improved multi-scale reasoning for object detection and localization with deep convolutional neural networks. To that end, an efficient and general object detection framework which operates on scale volumes of a deep feature pyramid is proposed. In contrast to the proposed approach, most current state-of-the-art object detectors operate on a single-scale in training, while testing involves independent evaluation across scales. One benefit of the proposed approach is in better capturing of multi-scale contextual information, resulting in significant gains in both detection performance and localization quality of objects on the PASCAL VOC dataset and a multi-view highway vehicles dataset. The joint detection and localization scale-specific models are shown to especially benefit detection of challenging object categories which exhibit large scale variation as well as detection of small objects.

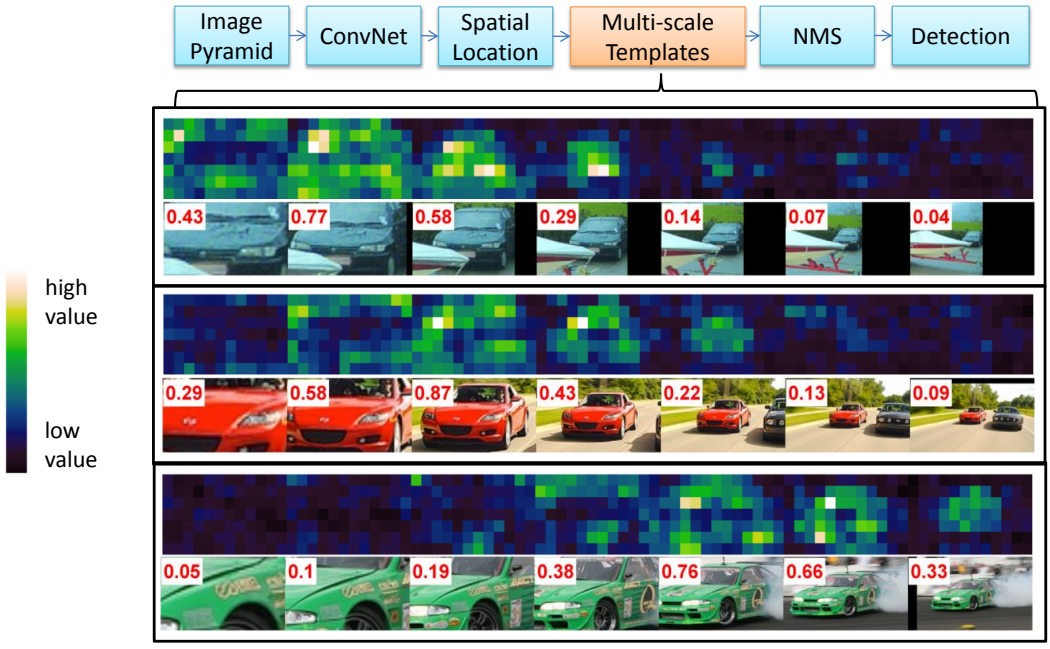


Figure 3.1: Pipeline of the proposed multi-scale structure (**MSS**) approach for studying the role of contextual and multi-scale cues in object detection and localization. Examples of some of the learned MSS models for 'car' over CNN features are shown, with brighter colors implying greater discriminative value. In red text is the overlap of the annotated ground truth object with a fixed model size. Note how each MSS template selects discriminative information across multiple scales, such as road and part information.

3.2 Introduction and Related Research

Visual recognition with computer vision has been rapidly improving due to the modern deep Convolutional Neural Network (CNN). The current success is fueled by large datasets, with pre-training of the network for a supervised object classification task on a large dataset [163], and consequent adaptation for new tasks such as object detection [162, 161] or scene analysis [176, 177]. The success of CNNs is attributed to the rich representation power of the deep network. Therefore, much of the current research is concentrated on better understanding properties captured by CNN representations. When transferring the network from a classification task to a detection and localization task, performance is greatly influenced by the ability to capture contextual and multi-scale information [178]. The main aim of this study is in the evaluation and improvement of this ability for CNNs using better multi-scale feature reasoning.

The biological vision system can recognize and locate objects under wide variability due in part to contextual reasoning. This is of particular importance when different image and object scales are considered. Hence, the tasks of capturing contextual cues and modeling multi-scale information are interleaved. Take for instance a car detection task as depicted in Fig. 3.1. Contextual reasoning appears at different image scales and spatial locations, from fine-grained part information (e.g. bumper, license plate, or tail lights occurring at certain configurations w.r.t. object orientation) and up to contextual scene

cues such as road cues or relationship to other objects. Fig. 3.1 depicts convolutional feature responses computed at twice and half the original image size for a selected feature channel. As can be seen, the responses differ both in magnitude and location depending on the image scale. Responses at different scales contain relevant contextual information for detection and localization. It has been known that CNNs can capture increasingly semantic representations at each layer [179], yet detection performance varies greatly w.r.t. appearance variations (scale, orientation, occlusion, and truncation) [178]. Therefore, contextual multi-scale information can help resolve such challenging cases. This work aims to analyze the benefit of training models that pool features over multiple image scales, both at adjacent and remote scales, on object detection (Fig. 3.1). Furthermore, the inference label space is adjusted to better leverage contextual multi-scale information in the localization of objects.

3.2.1 Contributions

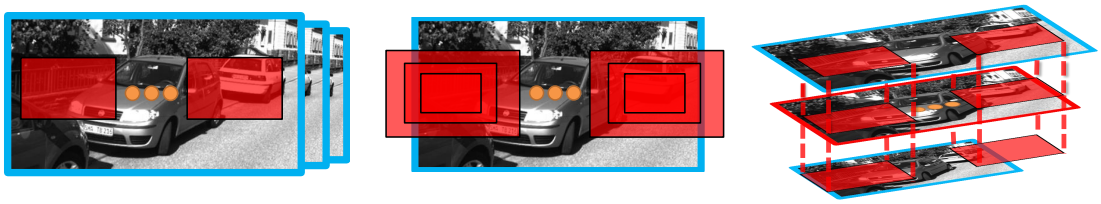
The main contributions presented in this work are as follows:

1. Multi-scale framework: we propose a framework for understanding CNN responses at multiple image scales. By training models that learn to pool features across multiple scales and appropriately designing the inference label space, the proposed framework is used to perform novel analysis useful in obtaining insight into the role of multi-scale and contextual information. In particular, the impact of dataset size and properties, impact of different scales and object properties, types of detection and localization errors, and model visualization are addressed. The framework generalizes current state-of-the-art object detectors which perform single-scale training and independent model testing across scales.

2. Better detection and localization: Replacing the commonly used local region classification pipeline for detection with a proposed set of joint detection and localization, scale-specific, context-aware, multi-scale volume models is shown to improve detection and localization quality. The contextual information is shown to be particularly useful in resolving challenging objects, such as objects at small scale. Experimental results demonstrate generalization of the proposed, **multi-scale structure (MSS)**, approach across feature types (CNN or hand-designed features) and datasets. The approach is light-weight in memory and computation, and is therefore useful for a variety of application domains requiring a balance between robust object detection and computational cost.

3.2.2 Related Research Studies

This study aims to better understand the benefits of improved multi-scale reasoning for object detection and localization. To that end, deep features are extracted at multiple image scales, and models that can perform inference over scale volumes and leverage contextual cues across different scales are



(a) Traditional, single-scale approaches (image pyramid and template pyramid) (b) Our proposed approach

Figure 3.2: Traditional approaches are limited in ability to capture contextual cues due to a single-scale training and testing of a single-scale local region. The proposed Multi-Scale Structure (MSS) approach extends the local regions across scales of an image pyramid to operate on scale volumes. The inference label space is modified as well to predict a localization label. The access to scale volumes across all scales of the image pyramid in training and testing time allows visualizing contextual cues and analyzing their role in detection and localization.

trained. The analysis provided by such models is complementary to existing related research studies discussing schemes for object detection and localization with multi-scale, contextual, and deep architectures, as will be discussed below.

Multi-scale detection: Traditional multi-scale object detection schemes employ a sliding window, which is a local, fixed-sized region in the image. The local region is scored in a classification task for an object presence, a process done exhaustively over different image locations and at different scales. In training, all training samples are re-sized to a fixed template size, thereby removing any scale-specific information and resulting in a single-scale model. In test time, local regions are classified independently across locations and scales. This limits the model's ability to well-localize an object and capture contextual cues. For instance, the example images in Fig. 3.1 would be scored independently, despite the highly structured information across scales. Finally, resolving multiple detections is handled with a heuristic Non-Maxima Suppression (NMS) module, which has no access to the image evidence. Several works have challenged this widely used pipeline. This includes the works of [180, 181, 157, 182, 183], which consider training multiple-resolution models. Such techniques were proposed for better handling appearance variation due to scale. The multi-resolution framework of [184] involves rejecting windows at low resolutions before the rest of the image pyramid is processed, thereby achieving speed gains. As the models trained in the aforementioned studies are still single-scale models, testing involves scoring each image location and scale. The contrast between the aforementioned studies and this work is that we incorporate a scale localization label into the label space of the detector, and consequently train models that operate on all scales of the image pyramid (see Fig. 3.2 for a high-level contrast). The approach explicitly accounts for variation in appearance due to scale and incorporates contextual cues for better localization. We note that the impact on detection and localization quality due to employing features at all scales, both remote and adjacent, has been rarely studied in related studies. As will be shown in Section 3.2.6, the studied framework generalizes the studies of [180, 181, 157, 182] which do not modify the multi-scale sliding window pipeline, and therefore provides complementary analysis.

CNN-based object detection: CNNs are a long-studied class of models [185–188], achieving

impressive performance on a variety of computer vision tasks in recent years [162, 161, 189]. Noteworthy CNN-based detection schemes are the OverFeat [162] and Region-based CNN (R-CNN [161]) architectures. Although both employ a CNN, OverFeat performs sliding-window detection (which is common in traditional object detection), while R-CNN operates on a set of region proposals. We note that both [162, 161] operate in a local-region manner without joint reasoning over multiple scales of an image pyramid. Current improvements over such architectures emphasize 1) The learning and incorporation of deeper networks [190, 191], 2) Resolving different components of the successful R-CNN framework into a single, end-to-end architecture. The original R-CNN framework involves a multi-stage pipeline, from object proposal generation (e.g. Selective Search [192]) to SVM training and bounding box regression. At test-time, a CNN forward pass is performed for each region proposal, which is costly. In contrast, SPPnet [193], Fast R-CNN [194], and OverFeat require only a single forward pass. Fast R-CNN [194] employs a Region of Interest (ROI) pooling layer which operates on region proposals projected to the convolutional feature map. Furthermore, the bounding box regression module is also integrated into the end-to-end training using a sibling output layer. Recently, another boost in performance was introduced in Faster R-CNN [8], which incorporates a Region Proposal Network in order to improve over the Selective Search region-proposal module. Independent testing at multiple scales is shown to improve performance on the PASCAL benchmark in the aforementioned studies, yet no further analysis is shown. Larger gains from multi-scale analysis are generally shown for other domains requiring robustness over large scale variations such as on-road vehicle detection [195] and pedestrian detection on the Caltech benchmark [196, 197]. In general, common CNN and hand-crafted object detectors involve training for and classifying a local region with a single-scale model. The contextual modeling capacity of such models is therefore limited, and detection of objects at multiple scales is done by independent scoring of an image pyramid. Nonetheless, visual information across scales at a given image location is highly correlated. Therefore, pooling features over scales in training and testing may benefit an object detector. Our work leverages a novel multi-scale detection framework in order to study the role of contextual information across image scales in a given spatial location.

Contextual object detection: Our study is relevant to the study of context. Classifying scale volumes directly benefits from contextual cues found at different levels of an image pyramid. Hence, scale and context modeling are interleaved fundamental tasks in computer vision [198, 199, 189, 200, 201]. Careful reasoning over these two tasks has shown great success in a variety of computer vision domains, from image segmentation [202] to edge detection [189]. The Deformable Part Model (DPM) [203, 204] is another example, as it reasons over a lower resolution root and higher resolution parts templates. Commonly, an additional module for capturing spatial and scale contextual interactions is applied over the score pyramid output of a traditional local-region, single-scale detector [161, 205, 203, 206]. In contrast, the studied framework in this work joins the two steps. In Chen *et al.* [207], a Multi-Order Contextual co-Occurrence (MOCO) framework was proposed, extending the Auto-Context idea [208, 209] for context modeling among boxes produced by traditional local region detection schemes. Sadeghi and Farhadi [210] propose visual phrases to reason over the output of object detectors and local context of object

relationships. Desai *et al.* [199] formulate multi-class object recognition as a structured prediction task, rescoring object boxes and replacing NMS for improved modeling of spatial co-occurrence. Li *et al.* [211] propose a hierarchical And-Or model for modeling context, parts, and spatial arrangements, and show large detection performance gains at a car detection task. Unlike the aforementioned, this work aims to study the benefit of incorporation of contextual, multi-scale cues directly into to object detection scheme. This is done both by modifying the detector to operate on scale volumes spanning the entire image pyramid and the inference label space. Analysis regarding the impact of such a framework is lacking in the aforementioned studies.

Multi-scale deep networks for contextual reasoning: Multi-scale deep networks have been previously studied in [200, 202, 212]. Eigen *et al.* [200] predicts depth maps by employing two deep network stacks, one for making coarse global prediction over the entire image and another for local refinement. Similarly to [200], this work aims to analyze the role of capturing information at different image scales. In contrast to [200], we discus the task of object detection and localization, study deep features at more than two image scales, and aim to better capture image appearance variations due to scale. Sermanet *et al.* [212] propose a multi-scale branched CNN for traffic sign recognition. Here, scale refers to different levels of feature abstraction as opposed to image pyramid scales. Although related to our study in capturing context, the method does not employ feature responses or weight learning across image scales for handling scale variation and improved object localization.

A close approach to ours is the work of Farabet *et al.* [202], which proposes a multi-scale CNN for semantic scene labeling of pixels. Consequently, segmentation quality is significantly improved by learning CNN weights which are shared across three image scales. Commonly, multi-scale architectures employ 2-3 image scales at most, while we employ 7-10, and modify the inference label space. The multi-scale CNN is shown in [202] to be better at capturing image evidence at a certain pixel location, yet no insight is given regarding th impact at different object scales (e.g. small objects), contribution of weights at different scales, relationship between object class and context usefulness, or impact on localization quality. Generally, adding responses at multiple image scales is known to benefit a variety of vision tasks, yet analysis on its role for general object detection and localization is lacking. Our study is also motivated by the fact that most current state-of-the-art object detectors do not employ multi-scale features or modeling [162, 161, 194, 8]. Furthermore, the training formulation in this work allows for visualization of the multi-scale, contextual cues. In contrast, most related studies discuss improvement due to multi-scale image features on a performance level only (e.g. features with one image scale vs. two image scales), without providing further insights.

3.2.3 Multi-scale Volumes for Deep Object Detection and Localization

The main approach in which context in object detection will be studied is presented in this section. The method is contrasted with existing schemes which are limited in their contextual reasoning in Fig. 3.2. Instead of training and testing over local image regions (either a sliding window or region

proposals), the approach employs an image pyramid and operates on features at all scales in training and testing. As large scales include fine-grained information, such as part-level information, and small scales include scene-level information, the MSS approach allows a study of the importance of cues at different scales. Furthermore, scale-specific multi-scale models are trained as contextual cues vary greatly w.r.t. the object scale. The MSS approach is also directly comparable to traditional single-scale training/testing baseline as the feature pyramid input to both is kept the unchanged.

3.2.4 Efficient Feature Pyramids

In order to efficiently train and test models which reason and pool over multi-scale features, all experiments are performed in an architecture similar to OverFeat [162, 213] and DeepPyramid DPM [214]. These have shown powerful generalization and flexibility to a variety of tasks, even without fine-tuning [196]. Hence, they are suitable for studying the ability to model context when transferring from the ImageNet classification task to the detection task. Furthermore, they provide *simple and efficient* means for handling multi-scale image pyramid information (order of magnitude faster than the original and widely used R-CNN [161]). By only employing the convolutional layers (discarding the fully connected layers), spatial structure is preserved and image regions can be directly projected to feature responses in an efficient manner without requiring a region proposal mechanism. Although more intricate approaches exist which preserve the fully connected layers (such as faster R-CNN [8]), the used ROI pooling layer in existing approaches still *operates on a single scale* of image features, and so the approach is orthogonal to our study. The network we employ is a truncated version of the winning network of the ILSVRC-2012 ImageNet challenge [163] composed of 8 layers in total. The network is used as a main tool to better understand context in CNNs. Employing deeper networks [8, 191] greatly improves performance by improving *local classification* power, but these are generally evaluated in a single-scale manner (or independent evaluation over multiple scales) and so are also orthogonal to this study. As tasks with large scale variation (e.g. pedestrian detection [196, 215]) require a large image pyramid in order to reach state-of-the-art performance, the approach in this work is also motivated by the need of real-world applications for a trade-off between performance, computational efficiency, and memory requirements. Our study of efficient multi-scale contextual reasoning is directly applicable to such applications.

3.2.5 Multi-scale detection with a single-scale template

First, we introduce notation to clarify and motivate the MSS approach. In traditional object detection, context reasoning is limited as detection is performed in a single scale fashion (tested independently at multiple image scales). First, a feature pyramid is constructed over the entire image at each scale to avoid redundant computation for each striding window. Let $p_s = (x, y, s)$ be a window in the s-th level of a feature pyramid with S scales anchored in the x, y position. Most of the analysis will involve a single aspect ratio model (which is common), and so we do not include that additional parameter in p_s, yet the formulation supports multiple aspect ratio models [216]. Generally, the feature pyramid is at a

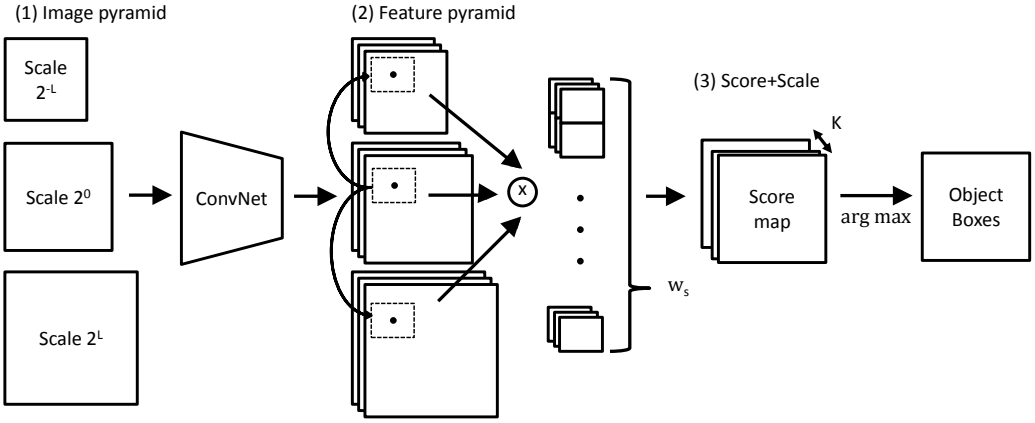


Figure 3.3: Our proposed approach re-samples the original image to obtain an image pyramid. Object-level annotations are converted to multi-scale annotations by obtaining a scale label. The scale label is assigned for each sample based on an overlap of the ground truth in each scale with a fixed model size (Section 3.2.6). Each sample is associated with a feature array that is cropped from the feature pyramid at shifted versions for preserving the same spatial location across scale. Testing involves scoring (represented by the 'X' operation in the figure) using learned multi-scale templates which convert the feature pyramid to an object score map. Note that the feature maps for each scale shown in the figure are at a lower spatial resolution than the original images.

lower spatial resolution than that of the image of the same scale (due to convolution and sub-sampling). Consequently, a zero-based index (x, y) in the feature map can be mapped to a pixel in the original image using a scale factor (cx, cy) based on the resolution of the feature map. Mapping locations over scales can be achieved by a multiplication by the scale factor as well. Each window contains an array of feature values, $\phi(p_s) \in \mathbb{R}^d$, to be scored using a filter w learned by a discriminative classifier, in our case a support vector machine (SVM). The scoring is done using a dot product,

$$f(p_s) = w \cdot \phi(p_s) \tag{3.1}$$

Generally, the template size is defined as the smallest object size to be detected, and further reduction in template size results in degradation of the detection performance. Note that learning and classification only occurs over a local window. A similar pipeline can be described using a template pyramid as studied in [182, 217, 180] and was shown to improve results due to capturing finer features at different scales that would have been discarded by the down-sampling. In this approach, a set of templates are learned, $(w_1, \ldots, w_S)$. In detection, the S templates are evaluated so that each location p in the original image scale is scored using the set of model templates

$$f(p) = \max_{s \in \{1,\ldots,S\}} w_s \cdot \phi(p) \tag{3.2}$$

where we drop s as only one scale of the image is considered. We emphasize that the model filters in this approach are also trained on locally windowed features only, but may capture different cues

for each scale. In principle, this approach is similar to the baseline as it performs the scoring convolution at each scale independently of all other scales (unlike MSS, as shown in Fig. 3.3).

3.2.6 Multi-scale detection with a multi-scale template

The feature pyramid computation and handling is mostly left unchanged in the proposed MSS approach. Spatial locations in the image space can be mapped across scales using a scale factor. As shown in Fig. 3.1, evaluations at the same spatial location occur repeatedly over scales. This mechanism is replaced by considering features from all scales at a given image location, i.e. $\psi(p) = (\phi(p_1), \ldots, \phi(p_S)) \in \mathbb{R}^{d \times S}$ descriptor.

Label space: Next the process of labeling training samples is outlined. Each sample is assigned a label, $y = (y^l, y^b, y^s) \in \mathcal{Y}$ with y^l the object class (in this study only $y^l \in \{-1, 1\}$ is considered), $y^b \in \mathbb{R}^4$ is the object bounding box parameters, and y^s is a scale label. In our experiments, the model dimensions are obtained from the average box size of all positive instances in the dataset (providing a single aspect ratio model). Training instances are sampled directly from the feature pyramid in a simple process where, 1) the multi-scale template is centered on top of each ground truth window spatial location and 2) Overlap with the ground truth is checked in each image scale (as shown in red in Fig. 3.1). Formally, a vector of overlaps F is constructed. If the image at s-th level contains $\hat{y}(s) = \{\hat{y}_1(s), \ldots, \hat{y}_N(s)\}$ ground truth boxes, the template box is centered on a positive sample at the s-th level (denoted as $B(s)$), so that entries of F are computed for each pyramid level,

$$F(s) = \max_{i \in \{1,\ldots,N\}} \text{ov}(B(s), \hat{y}_i^b(s)). \tag{3.3}$$

where ov(a,b) = area($a \cap b$)/area($a \cup b$) for two rectangles, a and b. F is shown for three examples in Fig. 3.1. For instance, for Fig. 3.1 first row, $y^s = (0100000)$. Peaks in $F(s)$ with high overlap imply a positive instance. This process potentially allows for multiple labels over scales to be predicted jointly, i.e. two almost overlapping objects at different scales, but such instances are rare. For simplicity, we only allow a single scale-label association by employing the scale where maximum overlap occurs.

Learning: Two max-margin approaches are studied for learning the multi-scale object templates, leveraging the highly structured multi-scale information, and analyzing importance of contextual information at different scales. Such information would have been ignored if a single-scale template was used.

Parameterization in the image pyramid can be done once over spatial locations at different scales by mapping across region locations with a scaling factor. Although these local regions across scales remain the same both in a traditional single-scale model classification procedure and the MSS approach, this new parameterization implies that we can concatenate features at all scales, as opposed to classifying these separately across scales. Furthermore, the previous section showed how such samples could be labeled, so the problem can now be posed as a multi-class problem.

One-vs-All: There are well developed machine learning tools for dealing with a large-dimensional

multi-class classification problem. A straightforward solution is with a one-vs-all (OVA) SVM, which allows training the multi-class templates quickly and in parallel. Window scoring is done using

$$f(p) = \max_{s \in \{1,\ldots,K\}} w_s \cdot \psi(p) \tag{3.4}$$

The scale of the box is obtained with an $\arg\max$ in Eqn. 3.4. In order to learn the K linear classifiers parameterized by the weight vectors $w_s \in \mathbb{R}^{d \times S}$, the stochastic dual coordinate ascent solver of [218] with a hinge loss is used. The maximum number of iterations is fixed at 5×10^6 and the tolerance for the stopping criterion at 1×10^{-7} for all of the experiments. Training a single multi-scale template on a CPU on average takes less than a minute. For simplicity, this study considers training a model for each scale, so that $K = S$. In general, this may not be the case (e.g. pedestrians occurring at close proximity but at different scales).

Structured SVM: A second approach can be used in order to learn all of the multi-scale templates jointly. A feature map is constructed using the labels of each sample as following,

$$\Phi(p, y) = (\Psi_1(p, y), \ldots, \Psi_K(p, y)). \tag{3.5}$$

$$\Psi_k(p, y) = \begin{cases} \psi(p) & \text{if } y = k \\ 0 & \text{otherwise} \end{cases} \tag{3.6}$$

This approach allows for learning a joint weight vector over all classes $w = (w_1, \ldots, w_K)$, such that

$$f(p) = \max_{y \in \mathcal{Y}} w \cdot \Phi(p, y) \tag{3.7}$$

Where the scale label prediction similar to as in Eqn. 3.4, but the loss function in training is defined differently using other elements of y.

Given a set of image-label pairs of the form $\{p^i, y_i\}$, the model is trained using a cost-sensitive SVM objective function [219–221]

$$\begin{aligned} &\min_{w,\xi} \frac{1}{2}||w||^2 + C \sum_{i=1}^{n} \xi_i \\ &\text{s.t. for } \forall i, \bar{y} \in \mathcal{Y} \setminus y_i \\ &w \cdot (\Phi(p^i, y_i) - \Phi(p^i, \bar{y})) \geq L(y_i, \bar{y}) - \xi_i \end{aligned} \tag{3.8}$$

The loss function, L, is chosen to favor large overlap with the ground truth,

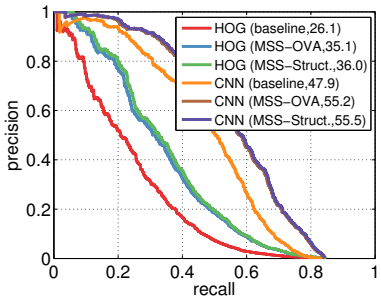


Figure 3.4: Model training comparison on a validation set for 'car' detection using HOG and conv_5 features. Average Precision (AP) is shown in parenthesis. Contextual information captured with MSS is shown to significantly improve detection performance using both one-vs-all (OVA) and structural SVM (Struct.) training.

$$L(y, \hat{y}) = \begin{cases} 0 & \text{if } y^l = \hat{y}^l = -1 \text{ or} \\ & \max_{i \in \{1,\ldots,N\}} \text{ov}(y^b, \hat{y}_i^b) < 0.6 \\ 1 & \text{otherwise} \end{cases} \tag{3.9}$$

Generalization of the single-scale approach

The main aim is to study context. The purpose of introducing the MSS approach is that it generalizes the traditional single-scale approach. Below, we show that in principle, if other scales do not contain additional contextual information, MSS reduces to the traditional single-scale approach. Eqns. 3.4 and 3.7 employ features at all scales for a given spatial location. Such a formulation allows learning the class weights jointly, as in Eqn. 3.7. It can be shown that this is a generalization of the single-scale template baseline. For instance, if no discriminative value is added by adding features at different scales, then the corresponding weights w_s in Eqn. 3.4 will only select features in the single best-fit scale (i.e. a degenerate case). Therefore, for each level s in the pyramid, $w_s \cdot \psi(p)$ becomes identical to $w \cdot \phi(p_s)$ as in Eqn. 3.1. A similar argument demonstrates the same for Eqn. 3.7. Therefore, both of the studied multi-scale template learning approaches can benefit by having access to additional information not accessible to the single-scale template approaches which only employs local window features at one scale. Furthermore, by learning a separate weight for each class, the model can account for appearance variations at different resolutions [182] and learn scale-specific context cues.

3.3 Experimental Evaluation

The experiments aim to quantify the importance of context cues in deeply learned features for a detection and localization task. Initially, the MSS approach is developed on the PASCAL VOC 2007 dataset [222] using its established metrics, followed by analysis on a multi-view highway vehicles dataset

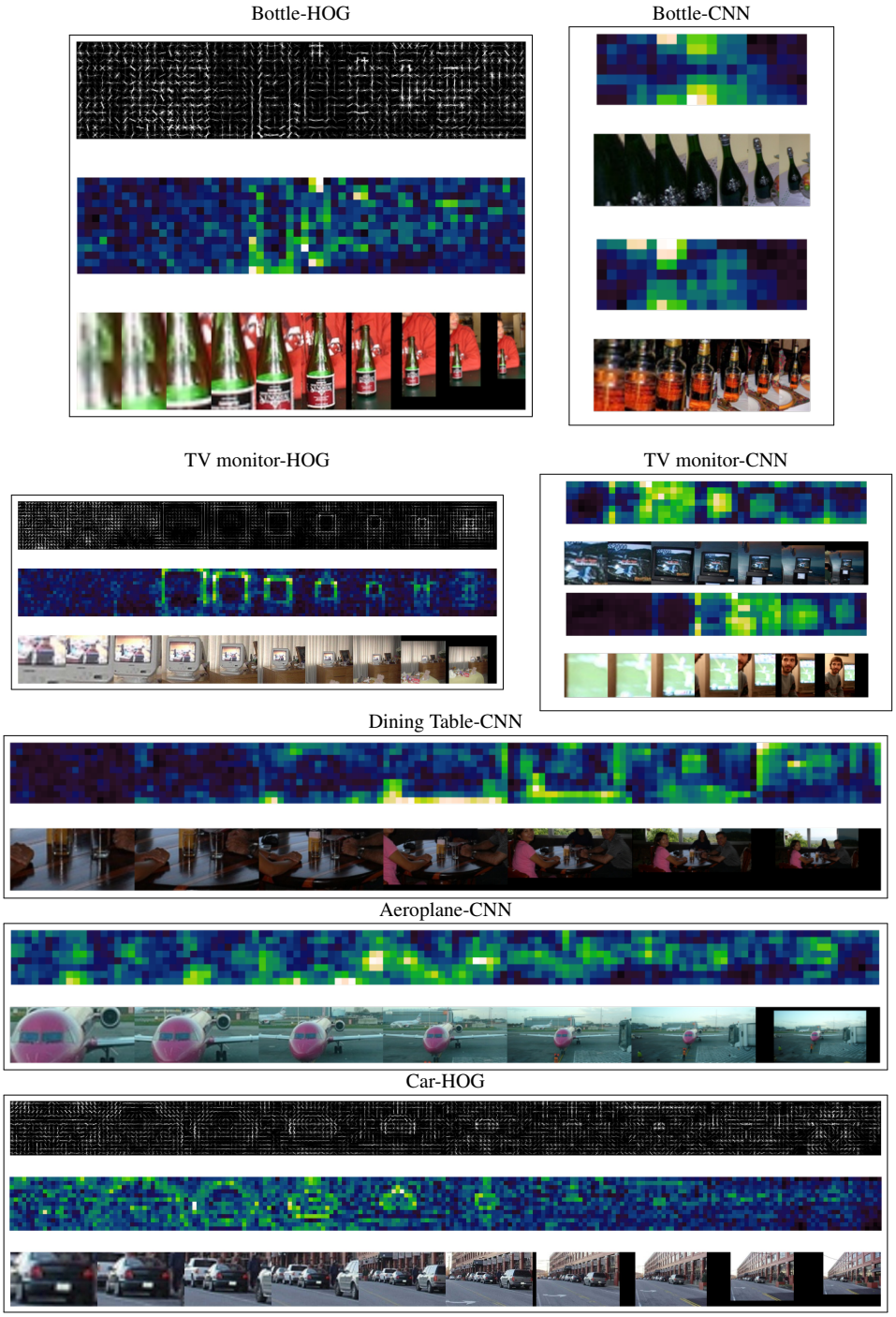


Figure 3.5: Visualization of multi-scale CNN and HOG templates. For each model, the maximum positive SVM weight for each block is shown together with an example instance. Brighter colors imply higher discriminative value. Large amount of discriminative value is placed at nearby and remote scales corresponding to contextual information (e.g. road cues at other scales).

with large variation in object scale.

Features: Two representative visual descriptors are employed in order the study the role of context. Most of the experiments involve the deeply learned features discussed in Sec 3.2.4. The fifth convolution layer output has 256 feature channels. The input to each convolutional or max pooling layer is zero-padded so that the features in a zero-based pixel location (x, y) in the feature space were generated by a receptive field centered at $(16x, 16y)$ in the image space (a stride of 16). As noted by Girshick *et al.* [214], the CNN features already provide part and scale selective cues. This can be enhanced by applying a 3×3 max-pooling layer. For direct comparison with the DeepPyramid approach [214], the same feature extraction and 7-scale pyramid pipeline was implemented in the experiment. The HOG feature implementation of [203] serves as a comparative baseline and studying generalization of experimental analysis across different feature types. HOG is used with a cell size/stride of 8.

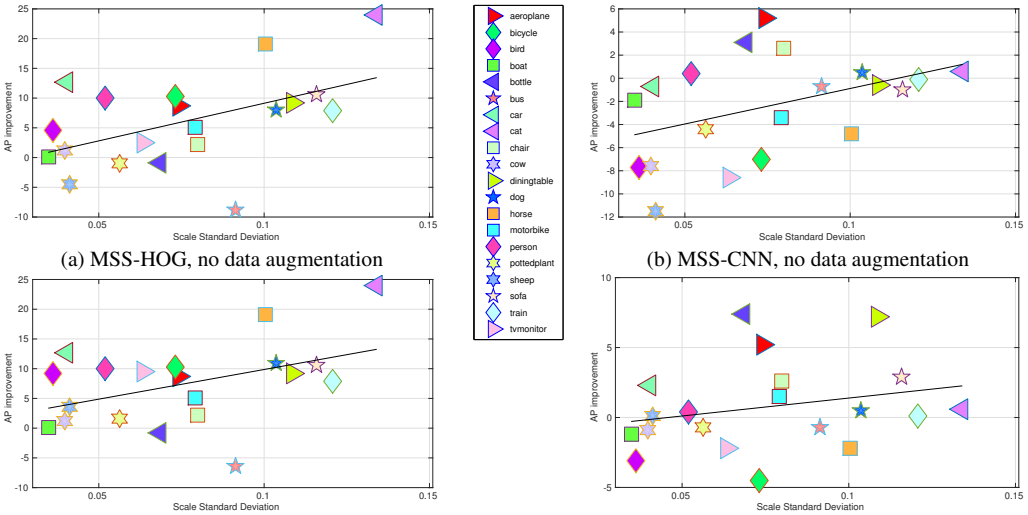
Image pyramid: The scale factor between levels is set to $2^{-1/2}$. The CNN feature pyramid spans three octaves with 7 levels. For HOG features, adding 3 more levels to the image pyramid for a total of 10 was shown to improve performance. In all of the experiments, training instances are extracted directly from the feature pyramid, as opposed to extracting features from cropped image samples. For the CNN feature pyramid, the features used are computed by the fifth convolutional layer which has a large receptive field of size 163×163 pixels.

Data augmentation: Training images are scale-jittered by up to an octave (either down-sampled and zero-padded or up-sampled and center-cropped). In addition to flipping, this data augmentation was essential for obtaining good performance of the MSS approach on all of the object categories.

Hard negative mining: All approaches studied employ an iterative process by which hard negatives are collected for re-training. The process eventually converges, when the number of negative samples generated are below a certain threshold. All of the experiments begin with a random set of 5000 negative samples. For a given object category, the initial negative samples are kept the same across techniques to allow direct comparison. In each iteration, up to 5000 additional negatives are collected. For mining, both images containing positive instances and negative images are used. A threshold of 0.3 overlap is used for mining negative samples from images with object instances.

3.3.1 Analysis on the PASCAL VOC dataset

Learning framework choice: First, we evaluated the choice of learning framework on a validation set of the 'car' category. Fig. 3.4 details the analysis of different learning and features combinations on the car category. Context is shown to benefit both HOG and conv_5 CNN features, as both learned MSS detectors are shown to greatly outperform the baseline in detection Average Precision (AP). Training the templates using the structural SVM allows for joint learning of the MSS templates, yet the improvement is marginal. Because structural SVM training is more costly, one-vs-all models are employed for the remainder of the experiments in this study. The structural SVM formulation may be of interest in the future for bounding box regression [223] or parts integration [203].



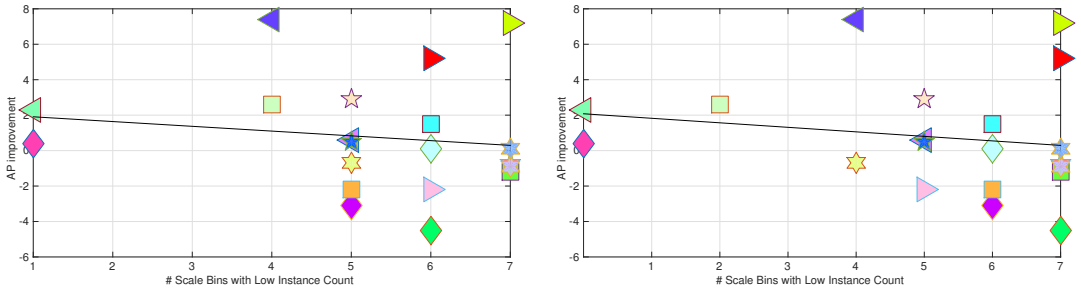
(a) MSS-HOG, no data augmentation

(b) MSS-CNN, no data augmentation

(c) MSS-HOG, with data augmentation

(d) MSS-CNN, with data augmentation

Figure 3.6: Relationship between the scale distribution of class samples in test time and the corresponding improvement in AP with the proposed MSS approach. As shown, our method shines when there is a large spread in the distribution over scales. Although some classes tend to appear in the PASCAL VOC dataset in a narrow scale distribution, this phenomenon is dataset and object specific. Therefore, if more instances at varying scales were to be added, the proposed approach would be better suited for such settings.



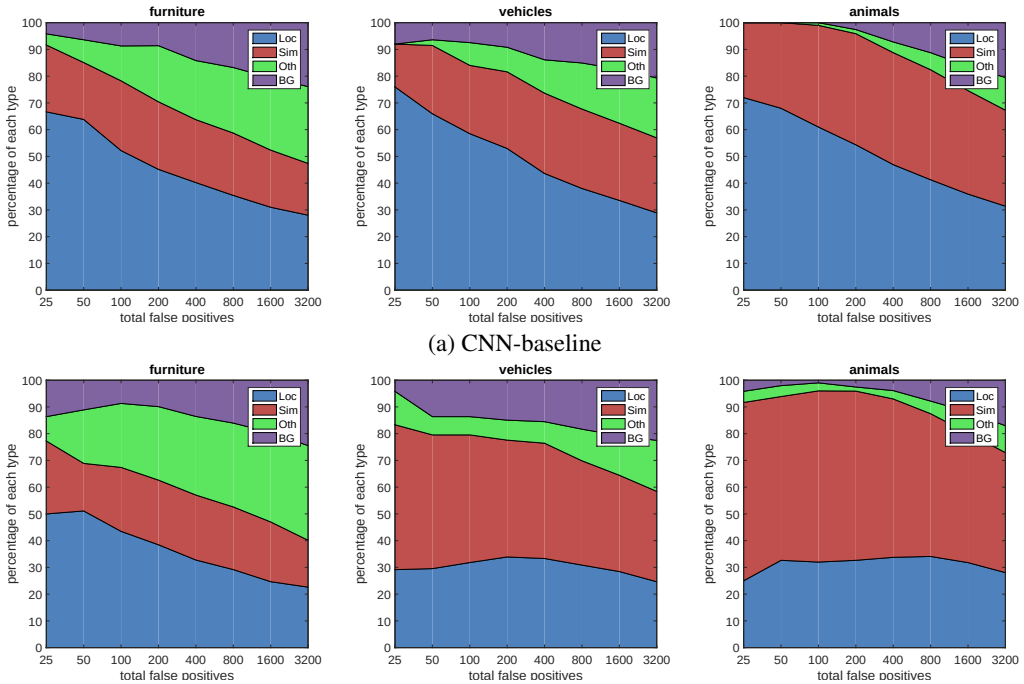
(a) Training set

(b) Testing set

Figure 3.7: Relationship between dataset properties and performance of the CNN-MSS approach. Some of the object classes in the PASCAL VOC benchmark contain a small number of object instances at multiple object scales, which poses a challenge to the scale-specific MSS models.

Visualization of the learned models: Fig. 3.5 depicts some of the learned MSS models for different object categories (positive valued entries in a learned MSS weight model). A single multi-scale template is visualized with a corresponding positive instance for each object category. For a given spatial location in the model, we visualize the learned model weights at each scale. As shown, while the best-fit scale includes large amount of the discriminative value, features from other scales (both adjacent and remote) are also selected. Contextual patterns can be seen, such as selection of road cues for car detection. We also observe the existence of alignment features, where certain appearance cues at one scale may assist in localization at another scale. This is shown by a repetitive shape pattern across the scales.

Relationship between scale-variation, dataset size, and MSS benefit: Our experiments showed



(a) CNN-baseline

(b) CNN-MSS

Figure 3.8: Analysis of the distribution of false positive types [224] for different types of objects on PASCAL VOC 2007. Training and testing is done with a single aspect ratio model. Loc - poor localization, Sim - confusion with a similar category, Oth - confusion with non-similar object category, and BG - confusion with background. The MSS approach is shown to significantly reduce errors due to poor localization.

the MSS method to significantly impact performance on some object classes by up to 7 AP points (e.g. 'bottle' and 'dining table' classes). Overall, 12 out of the 20 object categories benefit from the MSS approach, specifically on challenging object instances (i.e. small objects) and in terms of localization quality. Furthermore, overall mAP is improved with the MSS approach as shown in Table 3.1. Nonetheless, certain object categories do not benefit from incorporation of the multi-scale reasoning. As the reason for this is not immediately clear, we further study it next. A closer inspection of the scale distribution of the different classes reveals some insight, as shown in Fig. 3.6. First, a difference between HOG and CNN features is observed. Because CNN features are more scale-sensitive than HOG, this translates into smaller performance gains due to multi-scale context. Employing HOG on the other hand results in large gains consistently and across all object categories. A second observation is that some classes in the PASCAL VOC dataset exhibit smaller variation in scale. This limits the benefits due to incorporation of multi-scale context, and results in smaller AP improvement. If a certain object class exhibits smaller scale variation in the test set, the contextual cues will be less beneficial, which implies the results are influenced by the object statistics in the test set. Finally, we wish to analyze the role of dataset size on the variation in performance. Because the multi-scale templates require scale-specific instances, a small number of instances in the dataset (even with data augmentation) could lead to sub-optimal learning and

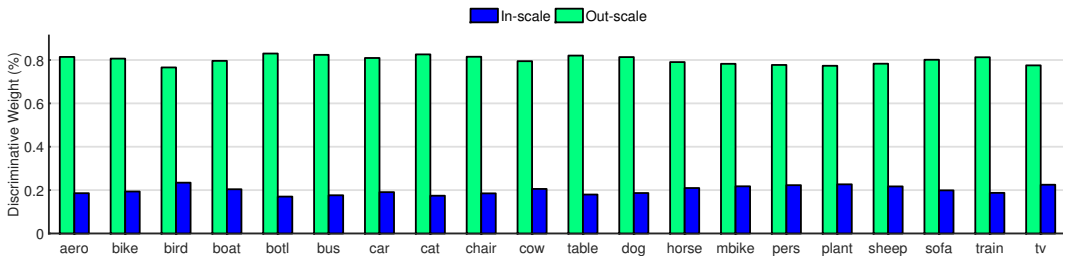


Figure 3.9: For CNN-based detection at a given scale, how important are out-of-scale context features? See Sec. 3.3.1 for details.

consequent reduction in performance gains. The importance of sufficient training instances for training each of the scale-specific MSS template is verified in Fig. 3.7. As shown in Fig. 3.7, classes with low detection AP improvement also contain a small number of objects in multiple image scales. In Fig. 3.7, low instance count is defined as a value under the average number of instances per scale bin across all object categories. Together with the observations in Fig. 3.6 regarding limited scale variability and insufficient training data explain why detection of certain classes, such as 'bottle', 'aeroplane', 'dining-table', and 'sofa', greatly benefit from the multi-scale context framework, and some classes do not (mainly 'boat' and 'bird' which contain small scale variability as shown in Fig. 3.6)). As will be shown next, the MSS approach significantly improves localization quality across all object categories.

Localization quality: Fig. 3.8 demonstrates improved localization due to incorporation of contextual cues across scales. The improvement is consistent over all types of object categories (clustered into three super-classes), including furniture, vehicles, and animals. This type of analysis is encouraging, as CNN-based object detectors are known to suffer from in-accurate localization. Our approach demonstrates the benefit on localization due to explicit incorporation of multi-scale features. This is intuitive, as the existence of certain feature responses at some scales can assist in better localization at another scale.

Context statistics: Training MSS models places discriminative value on each multi-scale cue. Next, we aim to understand how important are such cues in the learning process. For each class, features were divided into two: 1) Features found in the best-fit scale corresponding to the same features that would be employed if a single-scale template (referred to as 'in-scale' features), and 2) 'out-of-scale' features which a placed outside of the best-fit scale. The learned parameters, w, can be decomposed to positive and negative valued entries as $w = w^+ + w^-$. Indices with higher absolute value correspond to locations in the feature space which provide large discriminative value. Single-scale model training involves only 'in-scale' features. Furthermore, if 'out-of-scale' features provided no benefit, we would expect the majority of the discriminative weight to be placed on the best-fit 'in-scale' features only.

By studying the percentage of discriminative weight in w^+ and its distribution across scales for MSS-CNN, Fig. 3.9 demonstrates the clear trend of choosing features that are placed outside of the ground truth scale in training. This is a data-driven affirmation of the proposed approach. Although only positive weights shown in Fig. 3.9, the trends are similar both over positive weights w^+ and negative weights w^-. We can see that context can benefit CNN-detection greatly.

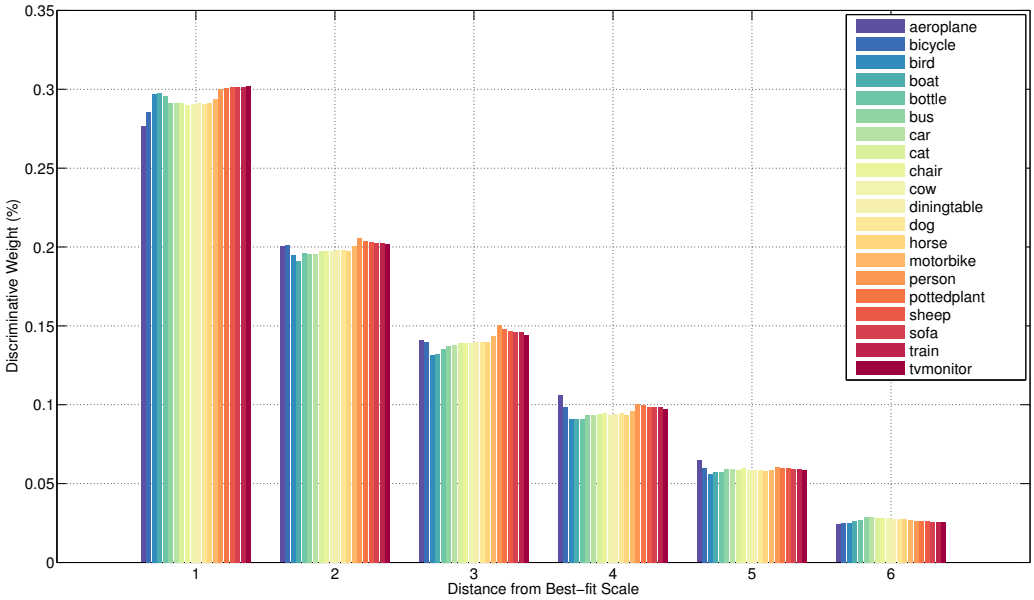


Figure 3.10: Relative to the best-fit scale, how is discriminative value distributed across pyramid levels? Most of the weight is found within adjacent levels (distance of '1' level away), but the contextual cues are shown to span all levels.

Table 3.1: Detection average precision (%) on VOC 2007 test. Column C shows the number of aspect ratio components. Performance improvement due to incorporation of context and multi-scale reasoning (MSS) with HOG and CNN features are shown. For reference, two other baselines, of a three aspect ratio components single-scale model and region proposal-based approach, are included. Note that the results of [214] for one and three aspect ratio components are using the publicly available code.

	C	aero	bike	bird	boat	botl	bus	car	cat	chair	cow	table	dog	horse	mbike	pers	plant	sheep	sofa	train	tv	mAP
HOG	1	13.05	23.54	0.80	1.70	12.85	28.91	27.38	0.68	11.31	8.89	11.04	2.68	13.52	18.49	13.05	5.60	14.58	12.19	16.28	24.48	13.05
HOG-MSS	1	21.72	33.86	10.05	1.81	12.02	22.54	40.04	24.66	13.52	10.08	20.28	13.53	32.57	23.63	23.05	7.24	18.23	22.75	24.20	33.98	20.49
CNN [214]	1	33.54	55.95	24.97	**14.24**	**36.96**	**44.31**	52.33	40.37	**30.07**	44.56	9.09	34.47	51.26	53.39	38.66	**25.22**	40.16	41.36	36.31	**57.97**	38.26
CNN-ours	1	36.68	**60.66**	**33.45**	13.71	17.66	44.02	58.48	49.71	25.12	**46.32**	44.08	41.47	**57.76**	54.18	48.90	22.95	43.84	43.34	42.17	54.96	41.97
CNN-MSS	1	**41.88**	56.17	30.40	12.54	25.05	43.36	**60.75**	**50.27**	27.68	45.41	**51.25**	**41.94**	55.60	**55.71**	**49.30**	22.25	**43.91**	**46.22**	**42.27**	52.78	**42.74**
CNN [214]	3	44.64	64.49	32.43	23.53	35.64	55.92	56.90	39.38	28.07	49.64	42.18	41.38	59.95	55.52	53.92	24.55	46.81	38.89	47.53	59.39	45.04
R-CNN $pool_5$ [161]	-	51.8	60.2	36.4	27.8	23.2	52.8	60.6	49.2	18.3	47.8	44.3	40.8	56.6	58.7	42.4	23.4	46.1	36.7	51.3	55.7	44.2

Table 3.2: The table depicts detection average precision (%) on VOC 2007 test for other methods employing part modeling and CNN features. The results are included for completeness, and meant to be compared with the results in Table 3.1. Our proposed method does not perform any explicit part reasoning.

	C	P	aero	bike	bird	boat	botl	bus	car	cat	chair	cow	table	dog	horse	mbike	pers	plant	sheep	sofa	train	tv	mAP
C-DPM [225]	3	8	39.7	59.5	**35.8**	24.8	35.5	53.7	48.6	**46.0**	**29.2**	**36.8**	**45.5**	**42.0**	57.7	56.0	37.4	**30.1**	31.1	**50.4**	**56.1**	51.6	**43.4**
Conv-DPM [226]	3	9	**48.9**	**67.3**	25.3	**25.1**	**35.7**	**58.3**	**60.1**	35.3	22.7	36.4	37.1	26.9	**64.9**	**62.0**	**47.0**	24.1	**37.5**	40.2	54.1	**57.0**	43.3

A further breakdown of this information is visualized in Fig. 3.10. Here, it is shown that most of the features selected outside of the best-fit scale are located in the adjacent scale (a distance of '1' pyramid level away), which is to be expected. Nonetheless, the MSS models consistently select features at more remote pyramid levels, even up to more than an octave away. This analysis suggests that CNN-based approaches can greatly benefit from careful multi-scale and contextual reasoning, which is not done in

Table 3.3: Results with fine-tuned features on VOC 2007 test. Our approach uses no region proposals (unlike RCNN), a single aspect ratio model, and only conv_5 feature maps.

	C	aero	bike	bird	boat	botl	bus	car	cat	chair	cow	table	dog	horse	mbike	pers	plant	sheep	sofa	train	tv	mAP
CNN	1	41.48	62.58	36.88	16.65	22.23	48.07	61.31	50.78	29.41	49.10	47.54	45.64	62.45	58.13	50.61	25.57	48.58	48.01	44.81	**59.53**	45.47
CNN-MSS	1	46.68	58.09	33.83	15.48	**29.62**	47.41	63.58	51.34	**31.97**	48.19	**54.71**	**46.11**	**60.29**	**59.66**	**51.01**	24.87	48.65	**50.89**	44.91	57.35	46.23
RCNN pool_5 [161]	-	**58.2**	**63.3**	**37.9**	**27.6**	26.1	**54.1**	**66.9**	**51.4**	26.7	**55.5**	43.4	43.1	57.7	59.0	45.8	**28.1**	**50.8**	40.6	**53.1**	56.4	**47.3**
RCNN fc_7 [161]	-	64.2	69.7	50.0	41.9	32.0	62.6	71.0	60.7	32.7	58.5	46.5	56.1	60.6	66.8	54.2	31.5	52.8	48.9	57.9	64.7	54.2
RCNN fc_7 BB [161]	-	68.1	72.8	56.8	43.0	36.8	66.3	74.2	67.6	34.4	63.5	54.5	61.2	69.1	68.6	58.7	33.4	62.9	51.1	62.5	64.8	58.5

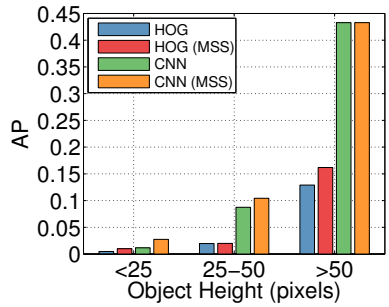


Figure 3.11: Improvement in performance for different object sizes. The largest gains due to incorporating the MSS approach are seen on smaller objects, which include more relevant contextual information throughout the multi-scale features.

most existing approaches for object detection. Simple pooling over both adjacent and remote scales is shown to greatly assist in detection, as shown in Fig. 3.10. Interestingly, a spike at certain remote scales is clearly seen with some categories, such as 'aeroplane', 'bicycle', and 'person'. This observation can be better understood by inspecting the template visualization in Fig. 3.5. For 'aeroplane', many of the scales contain informative contextual information as shown in Fig. 3.5, from wings to other aeroplanes. For 'bicycle', a rider may be found at a further scale. It can also be clearly observed how classes which MSS benefits least ('bird' and 'boat') have the smallest discriminative value placed in other scales out of all object categories. In these classes, contextual information is not selected as much.

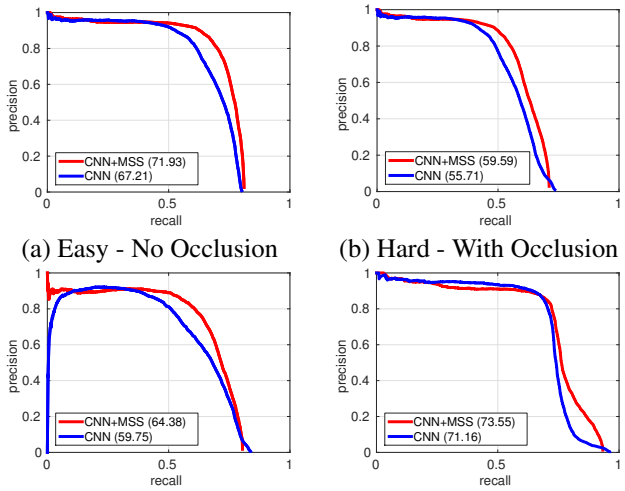
Performance breakdown by scale: As shown in Fig. 3.11, most gains in detection performance with CNN-features come from detection of smaller objects (50 pixels and less in height). This is intuitive, as such objects can benefit from incorporation of contextual cues at other scales.

Comparison with state-of-the-art: The main emphasis in this work is in analysis on modeling multi-scale context and its applications to efficient object detection and localization with deep features. The analysis framework was used to study scale importance, impact of dataset properties, and performance under varying object class and size settings. On PASCAL VOC, certain object classes greatly benefited from the proposed approach in detection, all of the 20 classes benefited in localization quality, and insights were made regarding challenging cases for the MSS approach. By employing only conv_5 feature maps, the method is efficient (requiring a single forward pass for each image scale) and have a low memory impact (no fully connected layers which contain most of the network parameters). As a reference, we provide absolute performance to other related research studies in Tables 3.1, 3.2, and 3.3 with different experimental settings.

For a fair comparison with a baseline, we closely followed Girshick *et al.* [214] in the deep feature pyramid extraction throughout the experiments. Overall, with a single aspect ratio model, our analysis results in a significant improvement of 4.48 mAP over the results of [214], from 38.26 mAP (obtained by the available implementation of [214]) to 42.74. We observed model size to be a crucial parameter, and increasing it results in improvement of the baseline to 41.97 mAP. Large gains in detection performance are shown for HOG, with an mAP increase of over 7 points. As discussed previously, the MSS approach has less impact on objects with little scale variation. Furthermore, as multi-scale templates require scale-specific instances, a small number of instances in the dataset (even with data augmentation), leads to sub-optimal learning and reduced performance gains. On the other hand, certain classes (e.g. 'aeroplane', 'car', 'table', and 'sofa') show large gains in performance. As the method in [214] employs no contextual reasoning, a further gain is obtained by the multi-scale reasoning in overall mAP.

As a reference, although not the main focus of this study, the results of [214] with three aspect ratios are shown, which has an overall 6.78 points improvement up to 45.04 AP, improving over R-CNN in performance with the same convolutional feature maps. The improvement due to multiple aspect ratio components is an orthogonal improvement to MSS as context cues can be incorporated into each of the components. Furthermore, note that unlike R-CNN, [214] and our study does not involve a region proposal mechanism and per-region forward pass through the network (either through the whole network or just through the fully connected layers), which is computationally costly. The CNN-MSS approach (42.74 mAP) performs similarly to other recently proposed approaches of Wan *et al.* [226] and Savalle *et al.* [225] employing multiple aspect ratio components, CNN feature pyramids, and explicit part reasoning. The best relevant results is achieved with R-CNN, fine-tuning, multiple fully connected layers (fc_7), and bounding-box (BB) regression at 58.5 mAP. Compared to R-CNN, the proposed approach is significantly more efficient in memory and computational cost. Furthermore, MSS learns scale-specific appearance and localization models while R-CNN does not. Results are shown both for no fine-tuning and with fine-tuning. R-CNN with the same convolutional features is outperformed on some classes where region proposals are weak. The results post fine-tuning shown in Table 3.3 demonstrate a consistent improvement. This is expected, as fine-tuning is mostly focused on improving local region representation.

Run-time speed: The computational speed is bound by two main factors, the feature pyramid extraction time and the model evaluation (either single-scale or MSS). The feature computation step (a 7 scale deep feature pyramid) is identical for the baseline and the MSS approach, running at ~ 0.4 seconds per image on PASCAL with a Titan X GPU. For the baseline, scoring a window p_s using the features $\phi(p_s) \in \mathbb{R}^d$ involves d operations, which is repeated over S scales ($S \times d$). For a given image location, evaluation with the MSS detector involves S models and an increase of the computational cost by a factor of S, to ($S \times S \times d$). In the current CPU implementation, the run-time of the MSS evaluation takes ~ 0.7 seconds per image. In the future, feature selection could potentially reduce the computational complexity of the detector evaluation for further speed gains.



(a) Easy - No Occlusion (b) Hard - With Occlusion

(c) Height - 20-40 Pixels (d) Height - Above 40 Pixels

Figure 3.12: Results for vehicle detection on highway settings with different evaluation procedures.

3.3.2 Results on Highway Vehicles

The PASCAL VOC 2007 dataset was used for developing the MSS approach and providing analysis in terms of impact of dataset properties, error types and localization quality, generalization to different object types, and sensitivity to object scale. In order to further test the performance of the proposed approach and understand its benefits, we employ a multi-view highway dataset captured using front and rear mounted cameras on a moving vehicle platform [227]. The highway settings are relevant as objects undergo large variation in scale as they enter and leave the scene. Furthermore, because the PASCAL VOC 2007 dataset targets generic object detection, it only contains a handful of images in settings similar to highway settings. The highway vehicles dataset is composed of a total of 1550 images containing 8295 objects. All truncated vehicles are also included in the evaluation. Object occlusion state have also been annotated in order to study performance under occlusion. The settings contain large variation in object height distribution.

The results for vehicle detection are shown in Fig. 3.12. When occluded objects are excluded, the MSS approach results in a significant improvement of 4.72 AP points over the baseline. With the inclusion of occluded objects, the improvement is consistent at 3.88 AP points. On this dataset, a main improvement is in detecting smaller objects and better resolving multiple detection boxes, as shown in Fig. 3.12(c). By observing the curves in Fig. 3.12, we can see how the MSS approach maintains precision at a higher recall over the baseline. This is due to the improved multi-scale reasoning. While the baseline scores objects based on local information and therefore relies on the heuristic NMS alone to resolve responses at nearby locations and multiple scales, the MSS approach can better reason over responses in different scales. This can be clearly seen in the example images in Fig. 3.13, where detection results are shown for both the MSS and the single scale baseline at a fixed recall rate. Fig. 3.13 also shows cases where false positives are reduced due to contextual information available at multiple scales.

(a) CNN-MSS (proposed) (b) CNN (baseline)

Figure 3.13: Results for vehicle detection on highway settings at a fixed recall rate. Observe how the MSS approach better reasons over multi-scale responses, allowing for higher precision at the same recall rate and better localization compared to the single-scale CNN, which employs independent scoring at each scale and relies on NMS alone for resolving multi-scale responses.

3.4 Chapter Concluding Remarks

Modeling image-level spatial context is a first step towards the overall objective of the thesis of developing human-centric and human-inspired algorithms. The role of multi-scale context in object detection with deep features was studied in this chapter of the thesis. An efficient framework for visual analysis of multi-scale contextual reasoning was proposed and studied on the PASCAL object detection benchmark and a highway vehicles dataset. Because the proposed approach operates on scale volumes,

learns scale-specific models, and infers a localization label, it was shown to result in more robust detection and localization of objects. Visualization and feature selection analysis demonstrated how discriminative learning strongly favors multi-scale cues when these are present in training, both in adjacent and remote image scales. Comparative analysis evaluated generalization of the proposed approach for different feature types and dataset settings. As current state-of-the-art object detectors emphasize local region feature pooling in detection, the insights in this study can be used to train better CNN-based object detectors. Robust detection and localization is a key component in a behavior understanding system.

This chapter is in part a reprint of material that has been accepted for publication in the journal of Pattern Recognition (2016), by Eshed Ohn-Bar, and Mohan M. Trivedi. The dissertation author was the primary investigator and author of this paper.

Chapter 4

Visual Analysis of Hand Gestures for Interactivity

Contextual image-level reasoning plays a key role in human-machine interactivity. Throughout my research, another main contribution relevant to human-machine interactivity has been the study of hand gestures. Hands play an integral part in human expression and language. They are among the most important components for a machine to perceive in its environment when interacting with humans. The challenge of providing machines with the skill to recognize the meaning of hand gestures is one of the most potentially useful challenges for modern engineering. Hand pose and movement are used on a daily basis in order to convey thoughts and manipulate objects [228]. Although expression may occur in multiple ways, such as speech, full body pose, or facial movement, this chapter emphasizes the integral role of hands as they coordinate with the aforementioned language tools.

4.1 Real-time, RGB-D based Gesture Recognition for Automotive Interfaces

In this section, we develop a vision-based system that employs a combined RGB and depth descriptor in order to classify hand gestures. The method is studied for a human-machine interface application in the car. Two interconnected modules are employed: one that detects a hand in the region of interaction and performs user classification, and the second performing the gesture recognition. The feasibility of the system is demonstrated using a challenging RGBD hand gesture data set collected under settings of common illumination variation and occlusion.

Recent years have seen a tremendous growth in novel devices and techniques for human-computer interaction (HCI). These draw upon human-to-human communication modalities in order to introduce a certain intuitiveness and ease to the HCI. In particular, interfaces incorporating hand gestures have gained

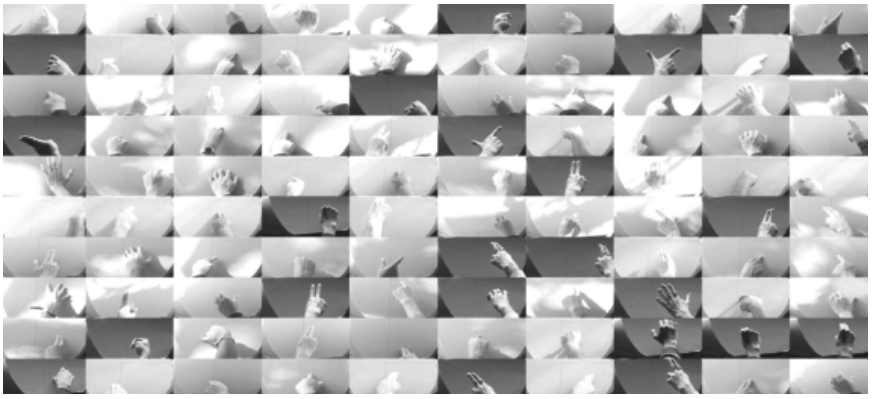
popularity in many fields of application. In this work, we are concerned with the automatic visual interpretation of dynamic hand gestures, and study these in a framework of an in-vehicle interface. A real-time, vision-based system is developed, with the goal of robust recognition of hand gestures performed by driver and passenger users. The techniques and analysis extend to many other application fields requiring hand gesture recognition in visually challenging, real-world settings.

Motivation for in-vehicle gestural interfaces: In this work, we are mainly concerned with developing a vision-based, hand gesture recognition system that can generalize over different users and operating modes, and show robustness under challenging visual settings. In addition to the general study of robust descriptors and fast classification schemes for hand gesture recognition, we are motivated by recent research showing advantages of gestural interfaces over other forms of interaction for certain HCI functionalities.

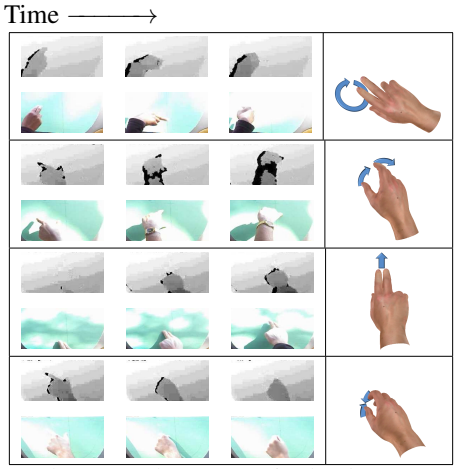
Among tactile, touch, and gestural in-vehicle interfaces, gesture interaction was reported to pose certain advantages over the other two, such as lower visual load, reduced driving errors, and a high level of user acceptability [229–231]. The reduction in visual load and non-intrusive nature led many automotive companies to research such HCI [232] in order to alleviate the growing concern of distraction from interfaces with increasingly complex functionality in today's vehicles [233–235]. Following a trend in other devices where multi-modal interfaces opened ways to new functionality, efficiency, and comfort for certain users (as opposed to interaction approaches based solely on tangible controllers), we propose an alternative or supplementary solution to the in-vehicle interface. As each modality has its strengths and limitations, we believe a multi-modal interface should be pursued for leveraging advantages from each modality and allowing customization to the user.

Advantages for developing a contact-less vision-based interface solution: The system proposed in this work may offer several advantages over a contact interface. First, camera input could possibly serve multiple purposes, in addition to the interface. For instance, it allows for analysis of additional hand activities or salient objects inside the car (as in [236, 127, 237–239]), important for advanced driver assistance systems. Furthermore, it allows for the determination of the user of the system (driver or passenger), which can be used for further customization. Second, it offers flexibility to where the gestures can be performed, such as close to the wheel region. A gestural interface located above the wheel using a heads up display was reported to have high user acceptability in [230]. In addition to allowing for interface location customization and a non-intrusive interface, the system can lead to further novel applications, such as for use from outside of the vehicle. Third, there may be some potential advantages in terms of cost, as placing a camera in the vehicle involves a relatively easy installation. Just as contact gestural interfaces showed certain advantages compared to conventional interfaces, contact-free interfaces and their effect on driver visual and mental load should be similarly studied. For instance, accurate coordination may be less needed when using a contact-free interface as opposed to when using a touch screen, thereby possibly reducing glances at the interface.

Challenges for a vision-based system: The method must generalize over users and variation in the performance of the gestures. Segmentation of continuous temporal gesture events is also difficult.



(a) Large variation in illumination and performance of the gestures.



(b) Example gestures in the dataset.

Figure 4.1: Examples of the challenges for a vision-based in-vehicle gesture interface. Illumination artifacts (saturation, high contrast shadows, etc.) throughout the performance of the gestures in the dataset are shown. Gestures are performed away from the sensor, resulting in frequent self-occlusion. The type of gestures varies from coarse hand motion to fine finger motion. (b) The gestures shown are (top to bottom): clockwise O swipe, rotate clockwise, scroll up, pinch\zoom-in.

In particular, gesture recognition in the volatile environment of the vehicle's interior differs significantly from gesture recognition in the constrained environment of an office. Firstly, the algorithm must be robust to varying global illumination changes and shadow artifacts. Secondly, since the camera is mounted behind the front-row seat occupants in our study and gestures are performed away from the sensor, the hand commonly self-occludes itself throughout the performance of the gestures. Precise pose estimation (as in [240, 241]) is difficult and was little studied before in settings of harsh illumination changes and large self-occlusion, yet many approaches rely on such pose information for producing the discriminatory features for gesture classification. Finally, fast computation (ideally real-time) is desirable.

In order to study these challenges, we collected a RGB-Depth (RGBD) dataset of 19 gestures, performed 2-3 times by 8 subjects (each subject preformed the set as both driver and passenger) for a total of 886 instances. Examples of gesture samples and the challenging settings are shown in Fig. 4.1.

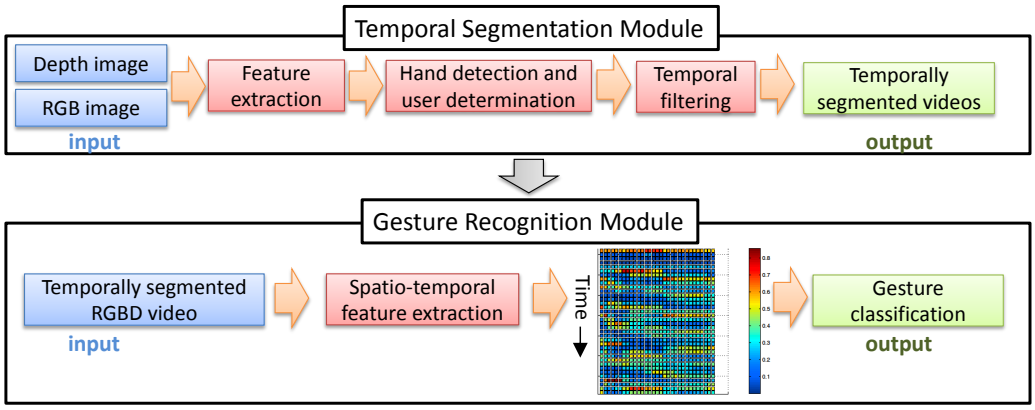


Figure 4.2: Outline of the main components of the system studied in this work for in-vehicle gesture recognition. First, the hand detection module provides segmentation of gestures and determines the user, which is either the passenger or the driver. This is followed by spatio-temporal feature analysis for performing gesture classification.

The dataset collected allows for studying user and orientation invariance, the effects of occlusion, and illumination variability due to the position of the interface in the top part of the center console. Different common spatio-temporal feature extraction methods were tested on the dataset, showing its difficulty (Table 4.4).

In this study, we pursue a no-pose approach for recognition of gestures. A set of common spatio-temporal descriptors [242–244] are evaluated in terms of speed and recognition accuracy. Each of the descriptors is compared over the different modalities (RGB and depth) with different classification schemes (kernel choices for a Support Vector Machine classifier [245]) for finding the optimal combination and gaining insights into the strengths and limitations of the different approaches. Finally, the gesture dataset is used to study effects of different training techniques, such as user-specific training and testing, on recognition performance. The results of this study demonstrate the feasibility of an in-vehicle gestural interface using a real-time system based on RGBD cues. The gesture recognition system studied is shown to be suitable for a wide range of functionalities in the car.

4.2 Related Research Studies

As the quality of RGB and depth output from cameras improve and hardware prices decline, a wide array of applications spurred an interest in gesture recognition in the research community. Relevant literature related to gesture recognition and user interfaces is summarized below.

Video descriptors for spatio-temporal gesture analysis: Recent techniques for extracting spatio-temporal features from video and depth input for the purpose of gesture and activity recognition are surveyed in [246, 247]. Generally, hand gesture recognition methods may extract shape and motion features that represent temporal changes corresponding to the gesture performance, as in [244, 248].

These can be extracted locally using spatio-temporal interest points (as in [249, 250]) or sampled densely. Such features may be hand crafted, as done in this work, or learned using a convolutional network [251]. Information of pose, although difficult to obtain in our application, is also highly useful for recognition, as demonstrated in [252–257].

Hand gesture recognition with RGBD cues: The introduction of high-quality depth sensors at a lower cost, such as the Microsoft Kinect, facilitated the development of many gesture recognition systems. In particular, hand gesture recognition systems were developed with applications in fields of sign language recognition [258–261], driver assistance [18, 121], smart environments [262, 263, 248], video games [264], medical instrumentation [265, 266], and other human-computer interfaces [267–269]. Hand gesture recognition systems commonly use depth information for background removal purposes [270–273]. [270] proposed using a Finger-Earth Mover's Distance (FEMD) for recognizing static poses. Hand detection is commonly performed using skin analysis [272, 260]. In [260], depth information is used to segment the hand and estimate its orientation using PCA with a refinement step. The classification of static gestures is performed using an average neighborhood margin maximization classifier combined with depth and hand rotation cues. In [261], a nearest neighbor classifier with a dynamic time warping (DTW) measure was used to classify dynamic hand gestures of digits from zero to nine. A Hidden Markov Model (HMM) may also be used [274] for gesture modeling. Minnen *et al.* [275] used features of global image statistics or grid coverage, and a randomized decision forest for depth-based static hand pose recognition. There has been some work in adapting color descriptors to be more effective when applied to depth data. As noted by [276], common RGB based techniques (e.g. spatio-temporal interest points as in Dollár *et al.* [277]) may not work well on the output of some depth sensors, and need to be adjusted as in [278].

In this work we focus on approaches that do not involve tracking of hand pose. Each descriptor is applied to the RGB and depth modality separately, and finally these are early-fused together by concatenation. Common spatio-temporal feature extraction methods such as a histogram of 3D oriented gradients (HOG3D) [243], motion boundary descriptors and dense trajectories [244], and other forms of gradient-based spatio-temporal feature extraction techniques [242] will be evaluated on the challenging dataset. For classification, an SVM classifier is employed [245].

Hand gesture interfaces in the car: Finally, we briefly review works with affinity to the vehicle domain. A similar effort to ours was reported in Zobl *et al.* [279], where a CCD camera and NIR LEDs illumination in a simulator were used to perform gesture recognition out of an elaborate gesture inventory of 15 gestures. The gestures used were both static and dynamic. Static gestures may be used to activate the dynamic gesture recognizer. A HMM is employed to perform the dynamic gesture recognition. The inventory is not explicitly mentioned, as well as the speed of the algorithm, and only one subject was used. There also has been some work towards standardization of the in-vehicle gestural interaction space [280]. Althoff *et al.* [281] studied 17 hand gestures and six head gestures using an infrared camera, and a HMM and rule-based classifier. Endres *et al.* [282] used a Theremin device, a contact-less device consisting of two metal antennas. Moving the hand alters the capacity of an oscillating current, generating a signal which is fed to a DTW classifier.

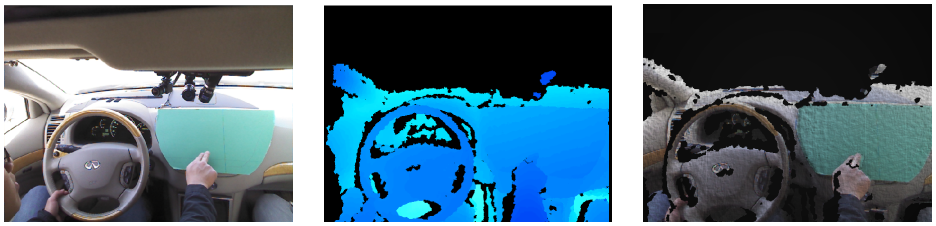


Figure 4.3: Camera setup (color, depth, and point cloud) for the in-vehicle vision-based gesture recognition system studied in this work.

4.3 Hand Gesture Recognition in the Car

4.3.1 Experimental Setup and Dataset

The proposed system uses RGB and depth images in a region of interest (ROI). In our experiments, this ROI was chosen to be the instrument panel (shown in Fig. 4.1 and Fig. 4.3). In order to demonstrate the feasibility of the system, we collected a dataset containing 19 hand gestures. The dataset is publicly available at `http://cvrr.ucsd.edu/LISA/hand.html`. Each gesture was performed about three times by eight subjects. Each subject performed the set two times, once as the driver and once as the passenger. The gestures are all dynamic, as these are common in human-to-human communication and existing gestural interfaces. The size of the RGB and depth maps are both 640×480, and the ROI is 115×250. Altogether, the dataset contains 886 gesture samples. The main focus of this work is recognition of gestures under illumination artifacts, and not the effects of the interface on driving. Therefore, subjects were requested to drive slowly in a parking lot while performing the gestures, as the gestures were verbally instructed. Subjects 1 and 4 performed the gestures in a stationary vehicle. It was observed that following the initial learning of the gesture set, both passenger and driver carried the gestures more naturally. At times this resulted in the hand partially leaving the pre-defined infotainment ROI, as strokes became large and more flowing. These large and inaccurate movements provided natural variations which were incorporated into the training and testing set.

Fig. 4.4 shows the illumination variation among videos and subjects. A temporal sum was performed over the number of pixel intensities above a threshold in each gesture video to produce an average intensity score for the video,

$$\text{Intensity Score} = \frac{1}{m \times n \times T} \sum_{t=1:T} |\{(x, y) : I_t(x, y) > 0.95\}| \tag{4.1}$$

That is, the average number of high intensity pixels over the $m \times n$ images I_t in a video of length T. A large variation in the dataset is observed in Fig. 4.4, both within the same subject and among subjects.

Interface location: Among previously proposed gestural interfaces, the location of the interface

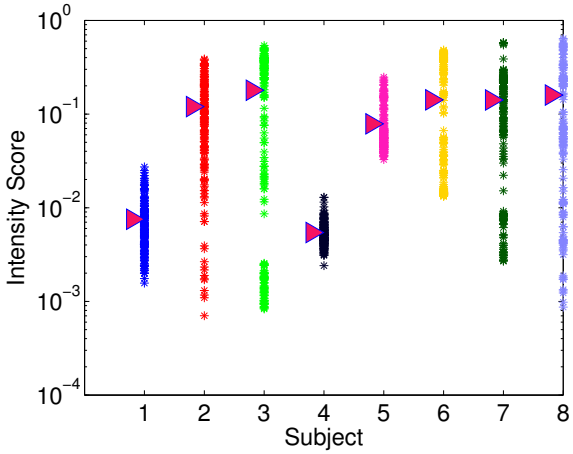


Figure 4.4: Illumination variation among different videos and subjects as the average percent of high pixel intensities (see Eqn. 1). Each point corresponds to one gesture sample video. The triangles plot the overall mean for each subject. Videos with little to no illumination variation were taken using subjects 1 and 4.

varies significantly. In our study, the gestures were performed by the center console, as shown in Fig. 4.3. We chose a position that would be difficult for a vision-based system due to illumination artifacts and self-occlusion. In future design, the location of the interface should depend on whether the system aims to replace or supplement existing secondary controls, and the type of feedback that will be used.

Gesture inventory: The inventory is as follows. Two-finger swipe gestures: *swipe left*, *swipe right*, *swipe down*, *swipe up*, *swipe V*, *swipe X*, *swipe + (plus)*. The motion in these is mostly performed with the fingers, and not with the hand, as opposed to the scroll where the fingers move with the entire hand in the direction of the scrolling: *scroll left*, *scroll right*, *scroll down*, and *scroll up*. One tap gestures can be done with one or three fingers, *one tap-1* and *one tap-3*. Next we have the *open* and *close*, a fist following a spread open palm or vice-versa. Finally, we use a two finger *pinch* as shown in Fig. 4.1-bottom, and the *expand* (opposite motion), as well as *rotate counter-clockwise* and *rotate clockwise* (Fig. 4.1-second row). We note that there were small variations in the performance of some of the gestures; for instance the *swipe X* and *swipe +* can be performed in multiple ways, depending on the starting position of the hand.

Gesture functionality: The 19 gestures are grouped into three subsets with increasing complexity for different in-vehicle applications as shown in Table 4.2. A set of functionalities is proposed for each gesture.

For GS1 (phone), the *open* and *close* gestures are used to answer or end a call. Scrolls provide volume control, and the *swipe +* provides the 'info/settings/bring up menu' button. GS2 involves additional gestures for music control. Swipes provide the 'next' and 'previous' controls. A tap with one finger pauses, and with three fingers allows for a voice search of a song. Finally, the X and V swipes provide feedback and ranking of a song; so that the user can 'like' or 'dis-like' songs. This gesture set contains

Table 4.1: Attribute summary of the eight recording sequences of video data used for training and testing. Weather conditions are indicated as overcast (O) and sunny (S). Time of capture was done in afternoon and mid-afternoon. Skin-color varies from light (C1) to intermediate (C2) and dark brown/black (C3).

Subject	Gender	Weather	Skin Color
1	M	O	C2
2	M	S	C2
3	M	S	C2
4	M	O	C3
5	M	S	C1
6	M	S	C3
7	F	S	C3
8	M	S	C1
Total Samples: {# Driver, # Passenger} = {450, 436}			

Table 4.2: Three subsets of gestures chosen for evaluation of application-specific gesture sets.

Gesture Set 1 (GS1) Phone	Gesture Set 2 (GS2) Music\Menu Control	Gesture Set 3 (GS3) Picture\Navigation
SwipePlus	SwipeX	SwipeUp
SwipeV	SwipeV	SwipeDown
Close	OneTap3	OneTap1
ScrollUp2	ScrollUp2	ScrollUp2
ScrollDown2	ScrollDown2	ScrollDown2
Open	OneTap1	ScrollRight2
	SwipeRight	ScrollLeft2
	SwipeLeft	RotateCntrClwse
		RotateClwse
		Expand\ZoomOut
		Pinch\ZoomIn

gestures that can be used for general navigation through other menus if needed. Finally, the more complex GS3 contains refined gestures purposed for picture or navigation control. A one finger tap is used for 'select', the scrolls for moving throughout a map, two finger rotation gestures rotate the view, and *expand* and *pinch* allows for zoom control. *Swipe up* and *swipe down* are used for transition between bird-eye view to street view.

4.3.2 Hand Detection and User Determination

Both recognition and temporal segmentation must be addressed. Since recognition was found to be a challenging task on its own, it is the main focus of this study. In particular, spatio-temporal features are evaluated in terms of speed, performance, and varying generalization. Although temporal segmentation is a difficult problem as well, in this work we employ a simple segmentation of temporal gestures using a hand presence detector, so that the hand must leave the ROI between different gestures.

The first module in the system performs hand detection in a chosen ROI. The classification may be binary, detecting whether a hand or not is present in the ROI, or multiclass for user determination, as in [283]. In the latter case, a three class classification performs recognition of the user: 1) no one; 2) driver; or 3) passenger. This is done with a simplified version of the histogram of oriented (HOG) algorithm

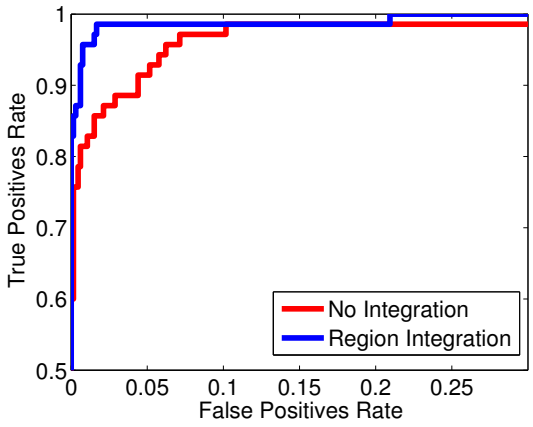


Figure 4.5: Driver hand presence detection in the instrument panel region. As the instrument panel region is large with common illumination artifacts, cues from other regions in the scene (such as the wheel region) can increase the robustness of the hand detection in the instrument panel region.

[284] which will be described below and an SVM classifier. For clarity and reproducibility, we detail the implementation of the visual features extraction used in this work.

HOG spatial feature extraction: Let $I(x,y)$ be an $m \times n$ signal. The discrete derivatives G_x and G_y are approximated using a $1D$ centered first difference $[-1,0,1]$ to obtain the magnitude, G, and quantized orientation angles into B bins, Θ. The image is split into $M \times N$ blocks. We found that overlapping the blocks produces improved results, and throughout the experiments a 50% overlap between the cells is used. Let G^s, Θ^s denote a cell for $s \in \{1,\ldots,M \cdot N\}$, so that the q^{th} bin for $q \in \{1,\ldots,B\}$ in the histogram descriptor for the cell is

$$h^s(q) = \sum_{x,y} G^s_{x,y} \cdot \mathbf{1}[\Theta^s(x,y) = \theta] \tag{4.2}$$

where $\theta \in \{-\pi + \frac{2\pi}{B} : \frac{2\pi}{B} : \pi\}$ and $\mathbf{1}$ is the indicator function. The local histogram is normalized using an L2-normalization: $h^s \rightarrow h^s / \sqrt{\|(h^s)\|_2 + \epsilon}$. Finally, the descriptor at frame t is the concatenation of the histograms from the cells

$$h_t = [h^1,\ldots,h^{M \cdot N}]. \tag{4.3}$$

For additional details and analysis on this part of the algorithm we refer the reader to [283].

Region integration for improved hand detection: The specific setup of location and size of the ROI can have a significant impact on the illumination variation and background noise in the ROI. Because the location of the ROI in our setup produces common illumination artifacts, we found that using visual information from other ROIs in the scene improves hand detection performance under ambiguous and challenging settings [133]. For instance, features extracted from the wheel, gear shift, and side hand-rest regions were shown to increase detection accuracy for the driver's hand in the ROI (Fig. 4.5).

4.3.3 Spatio-Temporal Descriptors from RGB and Depth Video

The first module described in the previous section produces a video sequence, which then requires spatio-temporal feature extraction for the classification of the gesture instance. We consider four approaches, each is applied to the RGB and depth video independently. These are compared in Table 4.3 in terms of extraction time and dimensionality. In calculation of extraction time, we time feature extraction for each video, divide by the number of frames, and average over the videos in the dataset. Given a set of video frames, we choose a descriptor function, $\phi : \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^T \rightarrow \mathbb{R}^d$, for producing the d dimensional feature vector for gesture classification.

HOG: A straightforward temporal descriptor is produced by choosing a vectorization operator on the spatial descriptors in each frame, h_t, $t \in \{1, \ldots, T\}$. In this case, the video is first resized to $T = 20$ frames by linear interpolation so that the descriptor is fixed in size.

$$\phi(I_1, \ldots, I_T) = \Big[h_1, \ldots, h_T\Big] \tag{4.4}$$

The pipeline for this algorithm contains three parameters, namely M, N, and B. We use $B = 8$ orientation bins in all of the experiments, and fix $M = N$, so that only one parameter can be varied, as shown in Fig. 4.6.

HOG^2: Another choice of ϕ is motivated by [242, 285]. In this case, the spatial descriptors are collected over time to form a 2D array (visualized in Fig. 4.2) of size $T \times (M \cdot N \cdot B)$. Changes in the feature vector correspond to changes in the shape and location of the hand. Consequently, the spatial HOG algorithm described in Section 4.3.2 is applied again using a $M^1 \times N^1$ grid of cells and B^1 angle quantization bins to extract a compact temporal descriptor of size $M^1 \cdot N^1 \cdot B^1$. The approach is termed HOG^2, since it involves applying the same algorithm twice (once in the spatial domain, and then again on those histograms over time). In this case, $\phi : \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^T \rightarrow \mathbb{R}^{M^1 \cdot N^1 \cdot B^1}$

$$\phi(I_1, \ldots, I_T) = \text{ HOG } \left(\begin{bmatrix} h_1 \\ \vdots \\ h_T \end{bmatrix} \right) \tag{4.5}$$

As in [242], we also use the mean of each of the spatial HOG features over time in the feature set. Generally, the dimensionality of HOG^2 is much lower than the corresponding temporal HOG concatenation. There are three additional parameters for the second operation of HOG, but we fix those to be the same as in the spatial HOG feature extraction so that $M^1 = M$, $N^1 = N$, $B^1 = B$.

HOG-PCA: Alternatively, we can reduce the dimensionality of the concatenated histograms descriptor (HOG) using Principal Component Analysis (PCA). In this case, we pre-compute the eigenspace using the training samples, and at test time project the temporal HOG concatenation feature using the eigenspace to derive a compact feature vector. Studying this operation is useful mainly for comparison with HOG^2.

HOG3D (Kläser *et al.* [243]): A spatio-temporal extension of HOG, where 3D gradient orienta-

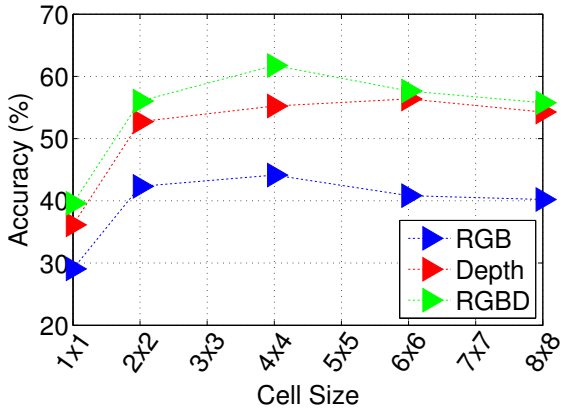


Figure 4.6: Varying the cell size parameters in the HOG-based gesture recognition algorithm with a linear SVM for a RGB, depth, and RGB+Depth descriptors. A fixed 8 bin orientation histogram is used. Results are shown on the entire 19 gestures dataset using leave-one-subject-out cross-validation (cross-subject test settings).

tions are binned using convex regular polyhedrons in order to produce the final histogram descriptor. The operation is performed on a dense grid, and a codebook is produced using k-means. In our experiments, we optimize k over $k \in \{500, 1000, 2000, 3000, 4000\}$. k-means is run five times and the best results are reported.

DTM (Heng *et al.* [244]): The dense trajectories and motion boundary descriptor uses optical flow to extract dense trajectories, around which shape (HOG) and motion (histograms of optical flow) descriptors are extracted. Trajectory shape descriptors encode local motion patterns, and motion boundary histograms (MBH) are extracted along the x and y directions. Similarly to HOG3D, we follow the author original implementation with a dense sampling grid and a codebook produced by k-means.

We emphasize that in our implementation, only HOG3D and DTM require codebook construction with k-means. In these, a video sequence is represented as a bag of local spatio-temporal features. k-means is used to produce the codebook by which to quantize features, and each video is represented as a frequency histogram of the visual words (assignment to visual words is performed using the Euclidean distance). The other techniques involve a global descriptor computed over the entire image patch. Furthermore, we experimented with a range of descriptors, such as the Cuboids [277] and HON4D [276], but even after parameter optimization these did not show improvement over the aforementioned baselines.

4.3.4 Classifier Choice

SVM [245] is used in the experiments due to its popularity in the action recognition literature with varying types of descriptors [244, 243]. In SVM classification, a Mercer similarity or kernel function needs to be defined. We study the following three kernel choices. Given two data points, $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^d$, the linear kernel is given as,

$$K_{LIN}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j \tag{4.6}$$

Table 4.3: Comparison of average extraction time per frame in milliseconds for each descriptor and for *one modality* - RGB or depth. Note that extracting RGBD cues from both modalities will require about twice the time. Experiments were done in C++ on a Linux 64-bit system with 8GB RAM and Intel Core i7 950 @ 3.07 GHz x 8. Asterisk * prefix - requires codebook construction.

Descriptor	Extraction Time (in ms)	Dimensionality
HOG	2.8	2560
HOG^2	2.83	256
HOG-PCA	3.25	256
DTM[244]	54	2000*
HOG3D[243]	372	1000*

an RBF-χ^2 kernel,

$$K_{\chi^2}(\mathbf{x}_i, \mathbf{x}_j) = exp(-\frac{1}{2C}\sum_{k=1:d}\frac{(x_{ik}-x_{jk})^2}{x_{ik}+x_{jk}}) \tag{4.7}$$

where C is the mean value of the χ^2 distances over the training samples, and a histogram intersection kernel (HIK),

$$K_{HI}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1:d}\min(x_{ik}, x_{jk}) \tag{4.8}$$

4.4 Experimental Evaluation and Discussion

Spatio-temporal descriptor analysis: The descriptors mentioned in Section 4.3 were compared with the three kernels in Table 4.4. The results are shown for the entire 19 gesture dataset with leave-one-subject-out cross validation (cross-subject test settings). As shown in Fig. 4.6, a 4×4 cell size in computing the HOG-based descriptors was shown to work well. The temporal HOG descriptor shows best results across modalities and kernels. Although lower performing descriptors benefit significantly from the non-linear kernels, the benefits for HOG are small. Overall, the DTM and HOG3D baselines are outperformed by the rest, possibly since these are densely sampled over the ROI yet background information does not contain useful information for the recognition (unlike other action recognition scenarios).

Inspecting the different HOG descriptors studied in this work, we observe that although the HOG^2 shows comparable results to DTM and HOG3D, it is outperformed by the HOG scheme. Interestingly, it appears to contain complementary information to the HOG scheme when combined, more so than when using the HOG-PCA scheme (although the two descriptors have the same dimensionality). This is the main reason for which HOG-PCA was studied in this work, and not for improving the results over HOG. Because HIK SVM with the HOG+HOG^2 descriptor showed good results, it is used in the remaining experiments.

Evaluation on gesture subsets: As mentioned in Section 5.4.1, a 19 gesture dataset may not be suitable for the application of an automotive interface. A set of three subsets was chosen and experiments were done using three testing methods, with results shown in Table 4.5. The three test settings are as

Table 4.4: Comparison of gesture classification results using the different spatio-temporal feature extraction methods on the entire 19 gesture dataset in a leave-one-subject-out cross validation (cross-subject test settings). Average and standard deviation over the 8 folds are shown. In bold are the best results for each modality and for each kernel for the SVM classifier; Linear, RBF-χ^2, and histogram intersection kernel. The best result overall is prefixed by an asterisk.

	K_{LIN}	K_{χ^2}	K_{HI}
Descriptor \ Modality	RGB (%)		
DTM	41.3 ± 14.1	47.0 ± 12.3	47.7 ± 12.0
HOG3D	35.8 ± 9.5	39.1 ± 8.5	37.8 ± 6.4
HOG	44.1 ± 11.8	46.5 ± 15.9	47.3 ± 14.1
HOG-PCA	44.3 ± 8.8	38.0 ± 9.2	42.1 ± 11.6
HOG^2	33.1 ± 8.9	35.4 ± 9.1	34.9 ± 8.6
HOG+HOG-PCA	45.4 ± 12.7	47.2 ± 14.3	49.0 ± 14.1
HOG+HOG^2	$\mathbf{47.9 \pm 13.8}$	$\mathbf{50.8 \pm 17.2}$	$\mathbf{52.3 \pm 16.2}$
Descriptor \ Modality	Depth (%)		
DTM	37.1 ± 9.5	40.8 ± 9.9	43.2 ± 11.8
HOG3D	40.6 ± 7.8	43.0 ± 11.4	44.2 ± 8.6
HOG	55.2 ± 13.9	57.0 ± 17.0	57.4 ± 15.6
HOG-PCA	49.1 ± 11.9	48.7 ± 13.7	48.8 ± 13.4
HOG^2	46.9 ± 12.8	49.6 ± 14.4	49.0 ± 14.7
HOG+HOG-PCA	55.9 ± 13.6	57.1 ± 16.7	57.8 ± 15.7
HOG+HOG^2	$\mathbf{57.5 \pm 14.6}$	$\mathbf{57.6 \pm 17.9}$	$\mathbf{58.6 \pm 15.8}$
Descriptor \ Modality	RGB+Depth (%)		
DTM	47.8 ± 13.2	51.5 ± 15.3	54.0 ± 14.8
HOG3D	36.7 ± 8.5	41.3 ± 9.0	44.6 ± 9.7
HOG	61.8 ± 15.7	62.1 ± 15.5	62.2 ± 16.8
HOG-PCA	56.2 ± 12.4	56.5 ± 13.7	57.3 ± 13.2
HOG^2	49.6 ± 14.6	52.3 ± 13.5	52.3 ± 14.5
HOG+HOG-PCA	62.2 ± 15.8	62.5 ± 16.0	62.1 ± 16.1
HOG+HOG^2	$\mathbf{63.3 \pm 15.3}$	$*\mathbf{64.5 \pm 16.9}$	$\mathbf{63.1 \pm 16.7}$

follows: *1/3-Subject*: a 3-fold cross validation where each time a third of the samples from each subject are reserved for training and the rest for testing. *2/3-Subject*: Similarly to 1/3-Subject, but two thirds of the samples are reserved for training from each subject and the remaining third for testing. *Cross-subject*: leave-one-subject-out cross validation. Results are done over 8 subjects and averaged.

The purpose of such a study is mostly in evaluating the generalization of the proposed algorithm, as well as the effect of user-specific training. The confusion matrix for each gesture subset using 2/3-Subject test settings are shown in Fig. 4.8. Table 4.5 reveals a lower accuracy on the challenging cross-subject testing, as expected. The reason is that within the 8 subjects there were large variations in the execution of each gesture.

Basic interface with a mode switch: Equipped with insight on the least ambiguous gestures from Fig. 4.8, we study a final gesture subset (Fig. 4.7) that provides a basic gesture interaction at high recognition accuracy (shown in Table 4.6). One possibility is to use one of the gestures, such as a one tap with three fingers (OneTap3) in order to navigate among functionality modes.

Table 4.5: Recognition accuracy and standard deviation over cross-validation using different evaluation methods discussed in Section 5.4. Increasing the number of user-specific samples results in improved recognition. RGB+Depth is the two descriptors concatenated and a HIK SVM. The overall category is the mean over the column for each modality, for showing the best modality settings and the effects of the test settings.

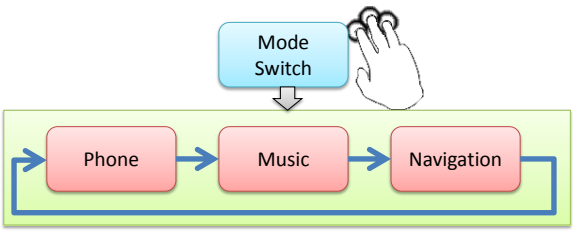
	1/3-Subject	2/3-Subject	Cross-Subject
	RGB (%)		
GS1	95.0 ± 1.1	96.5 ± 2.7	75.5 ± 16.7
GS2	91.0 ± 1.7	94.7 ± 1.3	63.8 ± 16.6
GS3	91.4 ± 2.0	94.6 ± 1.7	56.2 ± 14.7
Overall	92.5 ± 1.6	95.3 ± 1.9	65.2 ± 16.0

	Depth (%)		
GS1	92.7 ± 0.3	94.1 ± 1.6	80.9 ± 12.4
GS2	90.5 ± 1.5	93.6 ± 1.9	72.6 ± 19.4
GS3	87.0 ± 2.1	90.3 ± 2.1	67.3 ± 16.0
Overall	90.1 ± 1.3	92.3 ± 1.9	73.6 ± 15.9

	RGB+Depth (%)		
GS1	95.6 ± 1.1	96.5 ± 1.6	82.4 ± 15.1
GS2	92.9 ± 1.8	96.1 ± 1.2	73.8 ± 13.7
GS3	93.2 ± 1.9	96.0 ± 2.2	72.0 ± 15.6
Overall	$\mathbf{93.9 \pm 1.6}$	$\mathbf{96.2 \pm 1.7}$	$\mathbf{76.1 \pm 14.8}$

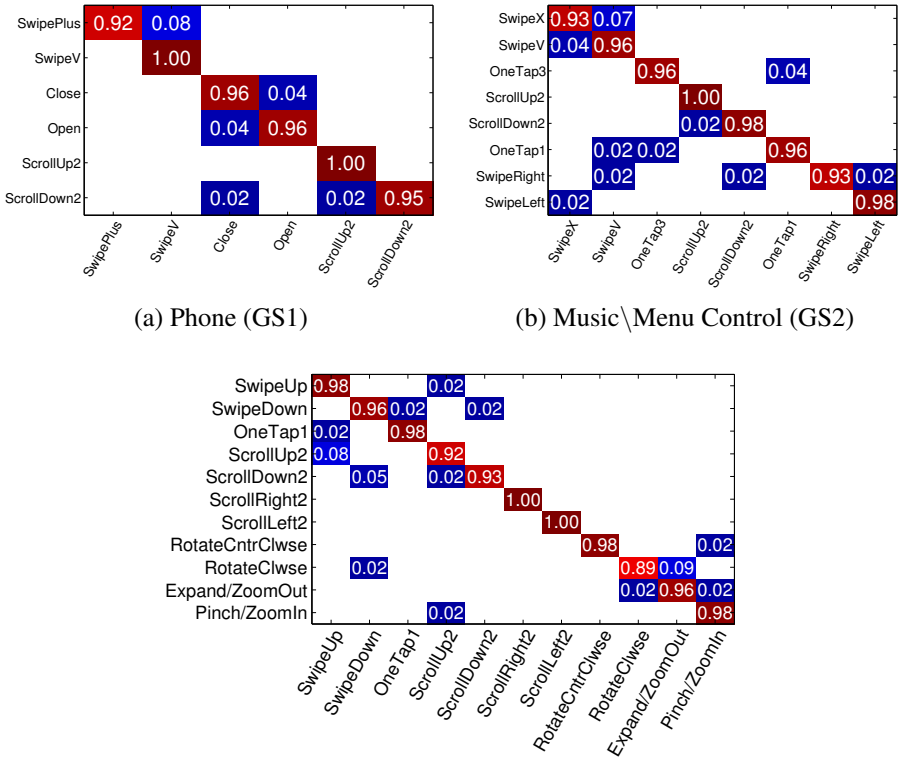
Table 4.6: Recognition accuracy using RGB+Depth and a HIK SVM on Gesture Set 4.

	1/3-Subject	2/3-Subject	Cross-Subject
	RGB+Depth (%)		
GS4	98.4 ± 0.5	99.7 ± 0.6	92.8 ± 8.8



Gesture Set 4 (GS4)
OneTap3
SwipeV
ScrollUp2
ScrollDown2
ScrollRight2
ScrollLeft2

Figure 4.7: Equipped with the analysis of the previously proposed gesture subsets, a final gesture set composed of less ambiguous gestures is defined and studied. The subset is designed for basic interaction, with one of the gestures used to switch between different functionality modes.



(a) Phone (GS1)

(b) Music\Menu Control (GS2)

(c) Picture\Navigation (GS3)

Figure 4.8: Results for the three gesture subsets for different in-vehicle applications using $2/3$-Subject test settings, where $2/3$ of the samples are used for training and the rest for testing in a 3-fold cross validation. A RGB+Depth combined descriptor was used. Average correct classification rates are shown in Table 4.5.

4.5 Analyzing Driver Hand Motion Patterns

Observing hand activity in the car provides a rich set of patterns relating to vehicle maneuvering, secondary tasks, driver distraction, and driver intent inference. This work strives to develop a vision-based framework for analyzing such patterns in real-time. First, hands are detected and tracked from a monocular camera. This provides position information of the left and right hands with no intrusion over long, naturalistic drives. Second, the motion trajectories are studied in settings of activity recognition, prediction, and higher-level semantic categorization.

4.6 Hand Detection and Tracking Dataset

Hands are used by drivers to perform primary and secondary tasks in the car. Hence, the study of driver hands has several potential applications, from studying driver behavior and alertness analysis to infotainment and human-machine interaction features. The problem is also relevant to other domains of robotics and engineering which involve cooperation with humans. In order to study this challenging computer vision and machine learning task, this work introduces an extensive, public, naturalistic video-based hand detection dataset in the automotive environment. The dataset highlights the challenges that may be observed in naturalistic driving settings, from different background complexities, illumination settings, users, and viewpoints. In each frame, hand bounding boxes are provided, as well as left/right, driver/passenger, and number of hands on the wheel annotations. Comparison with an existing hand detection datasets highlights the novel characteristics of the proposed dataset.

The detection and tracking of human hands has been studied extensively in the vision and learning community. In more recent years, the field has seen growing interest with the introduction of cheaper range sensors [132, 1], ego-centric applications [286–289], and driver study [24, 290, 291, 130, 231]. Until recently, the majority of studies have emphasized human-machine interaction (HMI) applications and gesture analysis in relatively constrained settings, as opposed to more naturalistic, out of the lab, social, and "in the wild" settings. Higher level semantic analysis of hand gestures would benefit from better detection and tracking of hands, which is challenging due to the tendency of the hand to deform and occlude itself. The dataset proposed in this work follows the more recent trends of leaving the constrained, in-front-of-the-sensor lab settings, and provides the full challenge of occlusion, hand-hand and hand-object interaction, illumination variability, and more. Specifically, we strive to create a hand detection dataset that incorporates the conditions encountered in a naturalistic driving setting.

In the domain of driving, several key motivations exist for the vision-based study of human hands. First, in the interest of the safety of a vehicle's occupants and their surroundings, our motivation to pursue the challenge of detecting vehicle occupants' hands is that successful detection will provide a major indication of the driver's level of attentiveness to the road. Drivers who regularly engage in distracting secondary tasks involving hands during vehicle operation, such as text messaging or eating, are reportedly common [292]. Second, driver hands provide a unique modality of understanding driver behavior [106].

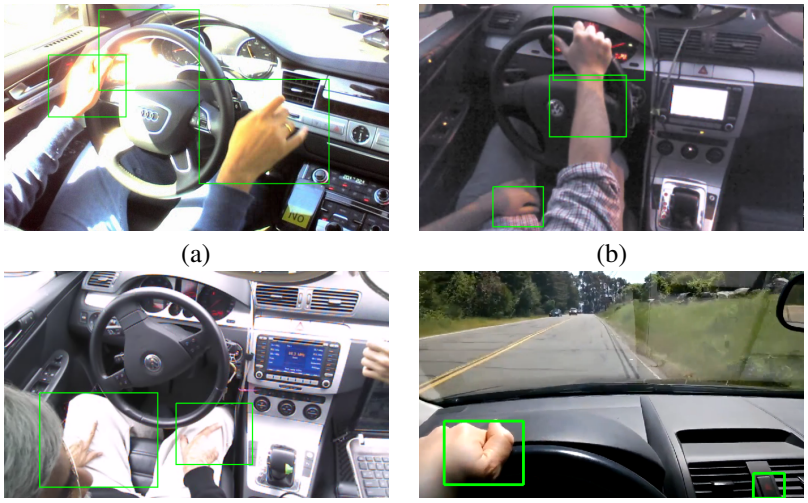


Figure 4.9: Challenges in the dataset. (a) Varying illumination conditions may cause false positives and missed detections. Sunlight causes the detector to consider the bright spot on the steering wheel as a hand. Realistic driving scenarios are prone to volatile illumination, and thus the inclusion of severe illumination settings in the hand detection dataset is vital. (b) Skin-colored non-hand objects, such as faces, forearms, and car interiors, may cause false positives in hand detection when the detector relies heavily on color information. Utilizing additional cues, such as context or motion cues, may make the detector more robust against false positives due to skin-colored objects. (c) A detector may miss a hand if it is occluded by another object, self-occluded, or otherwise not completely visible within the frame of the image. In this example, the passenger's left hand is not detected due to being partially out of the frame. (d) Introduction of different viewpoints may cause errors in detection because the perceived size and shape of the hand as well as background variability. This is useful for evaluating detector generalization capacity.

When maneuvering on a freeway or turning in an intersection, driver hands provide information of the driver's style and experience level. Third, large scale naturalistic driving studies could immensely benefit from automatic or semi-automatic analysis of driver hands and secondary tasks. Recently, the SHRP 2 Naturalistic Driving Study has been collecting raw data from 3,100 drivers throughout their everyday driving routines, which contain data looking into and out of the vehicle using camera sensors [293]. The purpose of the study is to understand the role of driver behavior in vehicular safety. The study is advantageous because pre-crash conditions and patterns in a driver's behaviors may be examined in detail, which may shine light on the role that driver behavior plays in a crash, demonstrate how drivers use hands to regain control of a vehicle, and provide valuable insight in the design of autonomous driving systems. The SHRP 2 study provides a dashboard view looking into the vehicle, which shows the driver's hands [127], thus demonstrating the direct applicability of the SHRP 2 data to the task of automatic analysis of hand positions and motion patterns in long-term video.

This work presents the following contributions:

Dataset: As a benefit to the research community, we assembled an annotated a video-based dataset for the task of hand detection under challenging naturalistic driving settings. We make this dataset accessible to the community as part of the Vision for Intelligent Vehicles and Applications (VIVA) chal-

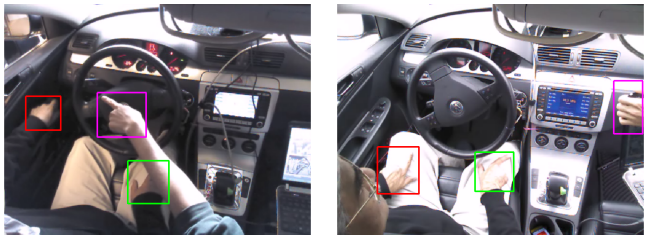


Figure 4.10: Visualization of annotations for a given video. Passenger hands are also annotated in the VIVA dataset as they may influence the behavior of the driver or may provide a further challenge in hand detection.

lenge[1]. We provide a method for participating research groups to publicly compare detection algorithms and results on a readily available online framework.

Analysis: A benchmark algorithm based on boosting decision trees over color and shape descriptors [215] is tuned for the settings of hand detection and is used for experimental analysis. This work demonstrates how a hand detector can greatly benefit from employing deeper decision trees.

Metrics: The work establishes suitable metrics and evaluation procedures on the dataset. The metrics emphasize overall precision-recall curve as well as performance at low false positives rates.

4.7 Challenges of A Naturalistic Driving Setting

A vision-based hand detection dataset must include the challenges encountered in a naturalistic driving setting in order to be fully representative of hands in a vehicle. Existing hand analysis research often circumvents the issues that are prevalent in realistic driving situations by constraining the hand detection problem such as by limiting the search space [294] or by fixing the hand and background colors [295]. A general hand detection dataset currently exists [296], which occasionally incorporates challenges that overlap with those found in a naturalistic driving setting. However, the occurrences of these challenges are uncommon in the general hand detection dataset as the imagery of said dataset are hand-picked photographs obtained via crowd-sourcing, while imagery found in a naturalistic driving setting and in-vehicle camera system will typically come from videos in a non-selective manner. Thus, when analyzing hands within vehicles, we do not have the ability to control the environment, to enforce an allowable range of clothing colors upon the driver, or to select which images are clear enough to analyze. Instead, the challenges that are often avoided in the field of hand analysis must now be considered in the context of a naturalistic driving setting.

This section outlines some of the challenges that exist in a naturalistic driving setting that we strive to represent within the VIVA hand detection dataset.

Illumination conditions: Varying illumination conditions (Figure 4.9(a)) and overexposure often cause false positives during detection [127].

[1]Dataset publicly available at http://cvrr.ucsd.edu/vivachallenge/

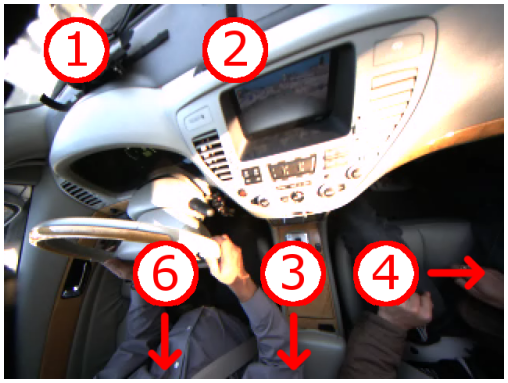


Figure 4.11: Camera positions indexed as in the dataset: 0 - handheld (not shown), 1 - front left, 2 - front right, 3 - back, 4 - side, 5 - top (current view), 6 - first-person.

Non-hand objects of similar color: Detectors that rely heavily on color features [297] may result in many false positives due to skin-colored non-hand objects, including faces, forearms, clothing, and car interiors (Figure 4.9(b)). While relying on color for detection may be beneficial in locating potential hand locations, further techniques must be employed to reject non-hand detections, such as a context detector [296].

Occlusion and truncation: Occlusion of hands by other objects and self-occlusion are challenges in the hand detection problem [23]. Figure 4.9(c) shows a passenger hand on the right that is missed by a detector because the hand is only partially visible. An improved detector must be able to locate hands even when the hands are partly occluded or out of frame. The necessity to detect occluded hands is important because driver hands that are not clearly visible may actually be involved in other activity, which identifies the driver's distracted state.

Camera viewpoints: Varying camera viewpoints may contribute to both false positives and false negatives due to representations of the hand that are rarely seen from other viewpoints. Changing the viewpoint may drastically change the perceived size of the hand, the orientation of the hand, and the level of occlusion of the hand. Figure 4.9(d) demonstrates both a false negative and a false positive that occurs in the first-person viewpoint. The hand is incompletely detected, and is thus considered a miss, while the hazard light button is falsely detected as a hand. An improved, generalized hand detector should be able to detect hands regardless of the viewpoint. While the camera viewpoint would typically be known if a hand detection system were built into a vehicle, we create a dataset with varied viewpoints with the intent to encourage the generalizability of detector submissions.

4.8 Description of the Dataset

In this section, we describe the VIVA hand detection dataset in detail, including the annotation format, sources of imagery, and categorized counts of images.

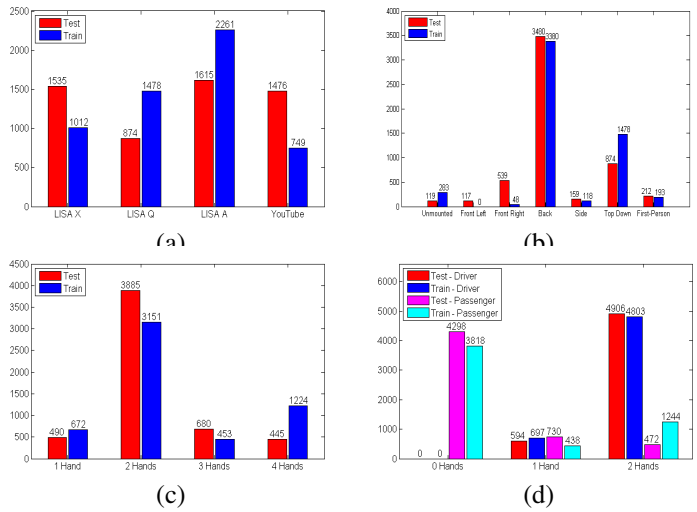


Figure 4.12: Dataset statistics. (a) Counts of images by vehicle type. Our three testbed vehicles are marked separately, but all vehicles from YouTube videos are grouped together. (b) Counts of images by viewpoint. The number of images from the back viewpoint largely dominates over the other viewpoints, and thus we consider the imagery from the back viewpoint as the easier of two levels of difficulty in our dataset. (c) Counts of images by the total number of visible hands. The maximum number of visible hands is 4, and there is always at least 1 hand visible in each image. (d) Counts of images by the number of visible driver and passenger hands. There is always at least 1 driver hand, and there is usually no visible passenger hands.

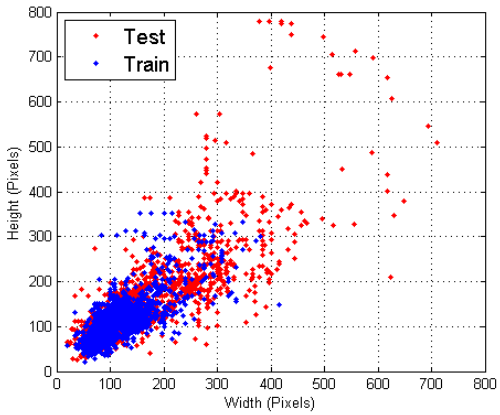


Figure 4.13: Annotation bounding box sizes for both the training and test set. The sizes of the hands are largely similar between the training and test set. The test set includes imagery in which the hands appear much larger than the hands in the training set.

4.8.1 Annotations

Placement of bounding boxes: Each hand present in a given image is annotated with an axis-aligned bounding box. Partially occluded hands have a bounding box that encompasses the entire hand including the occluded portions of the hand. When a hand is partially out of frame, a bounding box is drawn only around the portion of the hand within the frame. Completely occluded hands and hands completely out of frame have no bounding box. Each image in the training and test sets has at least one

annotated hand belonging to the driver and at most four annotated hands belonging to the driver and a single passenger. Figure 4.10 exemplifies typical annotated images from the dataset.

Format of ground truth: The format of the annotations follows the format supported by Piotr's Computer Vision MATLAB Toolbox (PMT) [298]. Each bounding box is specified by its top-left point, width, and height $[x,\ y,\ w,\ h]$. Additionally, each bounding box is assigned to one of four classes depending on whether the hand belongs to the driver or to a passenger and whether the hand is the owner's left or right hand. We note that left-right hand information is useful for many potential in-vehicle applications [299].

4.8.2 Sources of Imagery and Camera Positions

We collect and annotate data from various sources and viewpoints with the intent to create a diverse and challenging detection task.

The VIVA detection dataset is comprised of images gathered primarily from videos recorded from our lab. Three lab test-beds were used, labeled as LISA X, LISA Q, and LISA A. The viewpoints in these are either from behind the driver or top down from the rear view mirror. We also include images from YouTube videos of drives to further diversify the VIVA hand detection dataset. The majority of the selected YouTube videos have similar viewpoints as those observed in our testbeds imagery. The remaining YouTube imagery uses unfixed cameras, such as head-mounted or handheld cameras.

Figure 4.11 shows the possible camera positions from viewpoint 5 (top view). Handheld camera imagery in our dataset is viewed in a position similar to viewpoint 3 or 4, but are not classified as such because the camera position is not fixed in these cases.

4.8.3 Temporally Preceding Frames

For each image in the VIVA detection dataset, we make available up to three temporally preceding frames as is provided with the KITTI detection dataset [141, 195, 300]. The set of temporally previous frames do not have bounding box annotations and serves only to augment the detection data. The temporally previous frames will be useful to detection algorithms that utilize motion cues.

4.8.4 Annotation Statistics

In this section we present the counts of each image type and each image source.

Figure 4.12(a) shows we have over 2000 annotated images from each of our three testbed vehicles. To further diversify the dataset, we also include over 2000 images from YouTube which use imagery in unknown vehicles.

Figure 4.12(b) presents the number of images provided for each viewpoint. The distribution of imagery by viewpoint was selected based on the availability of imagery and our endeavor to create various levels of difficulty within the dataset. Imagery from the back view is most common in our dataset, and we

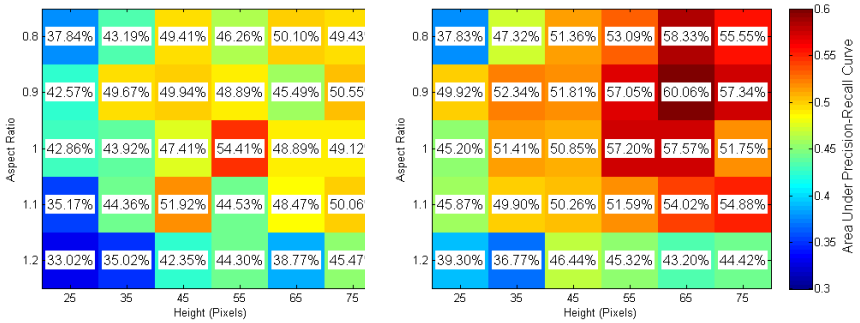


Figure 4.14: AP values for a grid search over model heights and aspect ratios with tree depth 2 (top) and tree depth 4 (bottom).

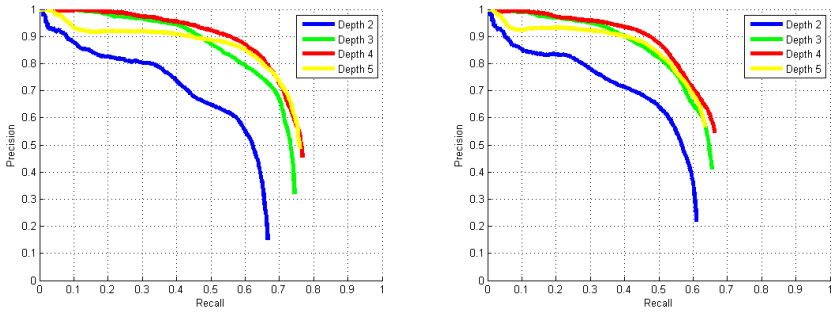


Figure 4.15: PR curves using boosted trees of depth 2, 3, 4, and 5 for both the L1 (left) and L2 (right) difficulty levels. The model height is held constant at 65 pixels and aspect ratio 0.9. Increasing the tree depth improves performance in terms of AP until a depth of 4. Further increases to the tree depth decrease performance due to overfitting.

intend for this viewpoint to be the easier portion of the dataset. A subset of the test data consisting of only back view imagery and larger instances (above 70 pixels in height) constitutes the easier difficulty level in the hand detection challenge which we denote as level-1 (**L1**) evaluation setting. Imagery from other viewpoints and instances greater than 25 pixels in height serve as the more difficult portion of our dataset, and correct hand detection for these viewpoints is reserved for detection algorithms that are capable of hand detection regardless of the camera viewpoint. The level-2 (**L2**) setting includes imagery from all viewpoints (including the images from the L1 setting) and serves as the more difficult evaluation setting.

The majority of the dataset uses imagery in which both of the driver's hands are visible and neither of a passenger's hands are visible. We provide counts of images with a specified total number of visible hands and a specified number of visible driver and passenger hands in Figures 4.12(c)–(d).

The annotated bounding box dimensions for both the training and test sets are plotted in Figure 4.13. The majority of the hand sizes are similar between the training and test set, though the amount of overlap decreases as the size of hands increases. The test set uses some YouTube videos that are of higher resolution than other imagery in our dataset, which causes hands to appear larger in terms of pixels.

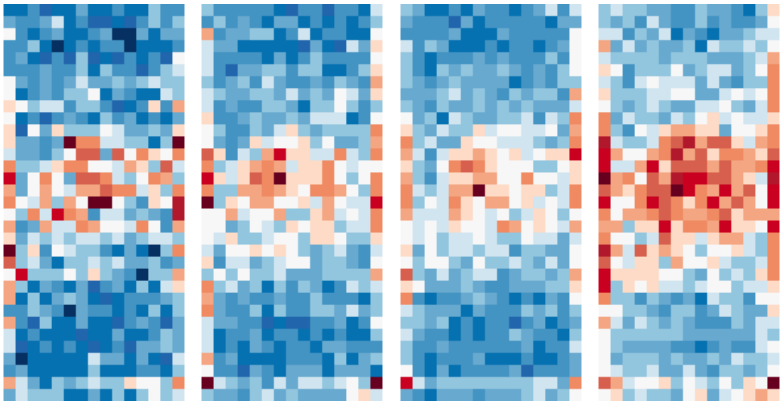


Figure 4.16: Model visualizations for detectors with tree depths of 2, 3, 4, and 5 (left to right). Model height and aspect ratio are held constant at 65 pixels and 0.9, respectively. Warmer colors represent the larger weights assigned to the corresponding locations within each considered window. The deeper colors in the visualization for the detector with a tree depth of 5 suggests that this detector may have overfit to the training data.

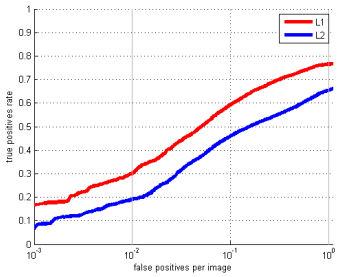


Figure 4.17: ROC curves for the detector with height 65 pixels, aspect ratio 0.9, and tree depth 4 on both the L1 and L2 difficulty levels. The incorporation of all viewpoints (L2) provides more challenging settings.

4.9 Experimental Evaluation

We use the Aggregate Channel Features (ACF) object detector [215] from the PMT [298] to test the viability of the VIVA hand detection dataset. This section describes evaluation metrics, the ACF detector, and the results of the detector on the hand detection set when we sweep through basic model parameters. We use the precision-recall (PR) curve and the area under the PR curve (AP) to evaluate how a parameter affects performance. We also publicize the average recall (AR) metric for each detection submission, computed from the ROC curve over 9 evenly sampled points in log space between 10^{-2} and 10^{0} false positives per image. The AR metric is suitable for summarizing detection performance at lower false positive rates. A detection is considered correct when it satisfies the PASCAL criterion. That is, a detection is correct when the proportion of overlap between the predicted bounding box and the ground truth bounding box is greater than 0.5 [222].

4.9.1 Detector Overview

The ACF detector utilizes 10 feature channels, a normalized gradient magnitude channel, 6 gradient orientation channels, and LUV color channels. Features are formed by aggregating and smoothing the channels, and AdaBoost is used to train decision trees based on these features. Object detection is performed using a sliding-window approach. An advantage of the ACF detector is that fast multiscale detection is achieved using feature pyramids which are quickly derived by computing features of octave-spaced scaled images and using approximations for scales between octaves [215]. The output of the ACF detector is a set of axis-aligned bounding boxes along with a score proportional to the confidence of detection for each box [215].

The ACF detector is highly successful in pedestrian detection [215], we thus treat the ACF detector as an effective multiscale object detector to test the viability of the hand detection dataset.

To maintain simplicity in training an ACF detector to evaluate the VIVA dataset, we only sweep through parameters that govern the size of the model and the complexity of the weak learners used in AdaBoost. We first select to use boosted trees of depth 2, and we perform a grid search over 6 model heights ranging from 25 to 75 pixels and 5 model aspect ratios from 0.8 to 1.2. The ACF parameters we keep constant are the number of classifiers in each of the four AdaBoost stages ([32, 128, 512, 2048]) and the non-maximal suppression threshold at which lower-scoring bounding boxes are suppressed if they overlap with other bounding boxes (0.2). All other ACF model parameters are left as their default values. We retain the AP obtained by each detector with depth 2 trees. We then repeat this process using detectors with depth 4 trees. Figure 4.14 shows the AP values obtained in both grid searches.

Using the model dimensions with the highest AP in the depth 4 model size grid search (height of 65 pixels and aspect ratio of 0.9), we sweep the tree depths to ensure that a tree depth of 4 best suits this dataset. Figure 4.15 shows the PR curves using detectors with tree depths of 2, 3, 4, and 5 on both the L1 and L2 evaluation settings. AP increases as tree depth increases until a depth of 4. The detector with a tree depth of 5 performs worse than the detector with a tree depth of 4, suggesting that a detector with a tree depth higher than 4 suffers from overfitting. Visualizing the models in Figure 4.16 provides further evidence that the detector with a tree depth of 5 overfits to the training data. In this visualization, warmer colors represent the larger weights assigned to the corresponding locations within each considered window, and the deeper colors in the depth 5 case (far right) suggest that this detector may have overfit to the training data.

Using the detector with model height 65 pixels, aspect ratio 0.9, and tree depth 4, we compute the AP for both the L1 and L2 settings: 70.09% for L1 and 60.06% for L2. We also generate an ROC curve (Figure 4.17) to better visualize the performance of the detector in terms of its true positive rate and number false positives per image. We also calculate AR for the L1 and L2 settings: 53.84% for L1 and 40.42% for L2.

Our initial results are promising, but suffer from false and missed detections. Typical high-scoring false positives are shown in Figure 4.18. The top row contains hands, but the poor fit of the

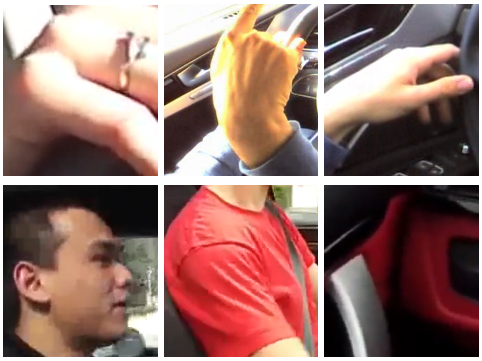


Figure 4.18: Typical high-scoring false positives from our trained ACF detector. The bounding boxes for the hands in the top row are poorly fit, thus causing such instances to be marked as false positives. The false positives in the bottom row suggest that our detector is heavily color-based, as skin-colored objects such as faces or objects with a red hue are detected as a hand.

bounding box prevents these detections from being true positive detections. The false positives in the bottom row suggest that our detector is heavily color-based because faces and red objects are mistakenly detected as hands. Further improvements to our detection system must be able to reject these types of false positives and must form better-fitting bounding boxes for each detection.

4.9.2 Cross Dataset Comparison

We performed a cross dataset comparison to assess whether the images provided in the VIVA hand detection dataset may be superseded by images provided in a general hand detection dataset. We selected the diverse hand detection dataset created by Mittal *et al*. [296] which includes annotated photographs in indoor and outdoor settings. Cross dataset training and testing resulted in AP of less than 10% in both cases, showing the difficulty of the hand detection problem and the domain differences among the datasets.

4.10 Trajectory-based Hand Activity Analysis

This study is concerned with construction of robust, vision-based tools for studying hand motion patterns under naturalistic, real-world settings. Since the study of human hands is an active field in the computer vision, machine learning, and human-machine interaction communities, the methods developed in this work are relevant to a wide array of applications. Inferring hand activity is especially important in the operated vehicle, as hands are a common medium for expressing and conveying information. For instance, it may provide vital information about the state of attentiveness of the driver. In order to clearly motivate the study, we list potential applications below.

Motivating applications: First, hand tracking allows the study of *preparatory movements* for maneuvers [301, 236]. Such information may be useful when providing alerts and support to the driver

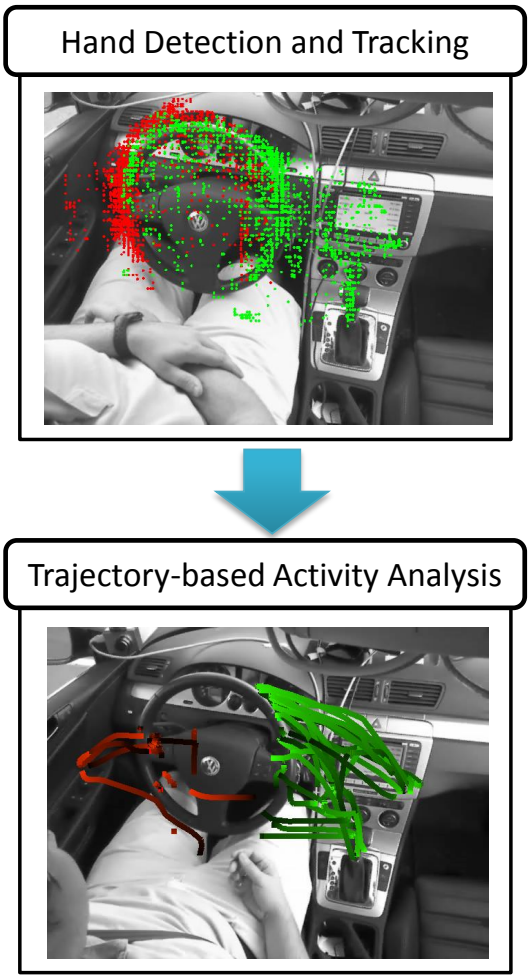


Figure 4.19: Motion patterns are studied in terms of activity classification, prediction, and high-level semantics by observing hand movement in naturalistic driving settings. First, driver hands are detected and tracked in real-time in order to produce trajectories in real-time processing. The figure depicts left and right hand positions (in red and green respectively) for an entire drive. Trajectories are formed and used for several proposed driver assistance applications.

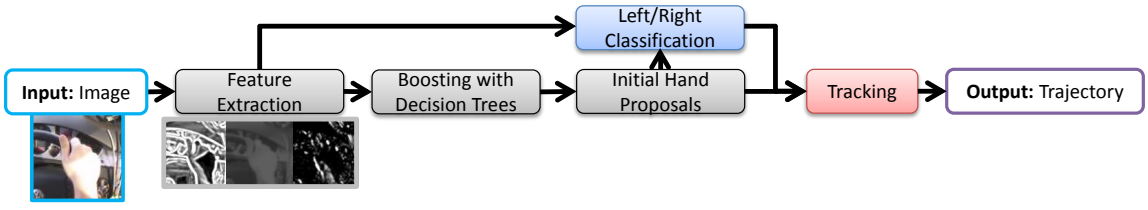


Figure 4.20: The hand detection module. Hand location proposals are outputted by AdaBoost with color (LUV colorspace pixels) and gradient (normalized gradient and histogram of oriented gradients). These are classified as left or right hands, and tracking provides the hand trajectories.

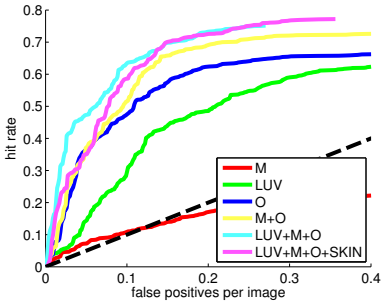


Figure 4.21: The impact of each of the studied features on detection performance is shown (M-gradient magnitude, O-gradient orientation, SKIN-learned skin-likelihood classifier, and LUV colorspace pixels).

[302]. For instance, while performing a sharp turn a driver may shift the hand position while the turn is ongoing in order to further turn the wheel, an action which may lead to an accident. Another example is in preparing for an overtaking maneuver, where a driver may shift the hand position together with a sequence of head and body pose dynamics to prepare for the overtake [236]. A second potential application is in *monitoring distraction levels*, as hand-vehicle and hand-object interactions (such as text messaging, handling navigation, etc.) can potentially increase visual, manual, and cognitive load [235]. This important application is presseing as drivers today are increasingly engaged in secondary tasks behind the wheel (23.5% of the time according to [235]). A third possible application lies in providing a framework for hand gesture recognition for *interactivity*, as in [130]. Finally, *long term analysis* of hand motion can provide useful insight into crash and near-crash events. For instance, in studying gestures performed by the driver for re-gaining control following an unexpected event. The framework proposed in this work can be immediately applied to other applications of hand gesture recognition [303], such as tutoring applications as in [248].

4.11 Hand Detection Module

In this section we specify the image pre- and post-processing, feature extraction, and training and testing routines for the hand detector.

Hand detection is a challenging task, studied extensively in the vision community. In our dataset, some main challenges are common occlusion by objects and self-occlusion of the hand, deformation, and rotation (see Fig. 4.22). Color, edge, and texture cues are commonly used for hand detection [304, 133]). The detection scheme of aggregate channel features from [215] is employed due to the fast detection (30 frames per second on a 640×480 image) and state-of-the-art detection performance.

For evaluating the hand detection module, 922 hand instances are used for training and 1516 for testing. Color features, in particular LUV colorspace pixel values, were shown to work significantly better compared to RGB or HSV in detection. For gradient orientation features, 6 orientation bins are used and gradient magnitude. An AdaBoost classifier is trained in four stages, with number of trees starting

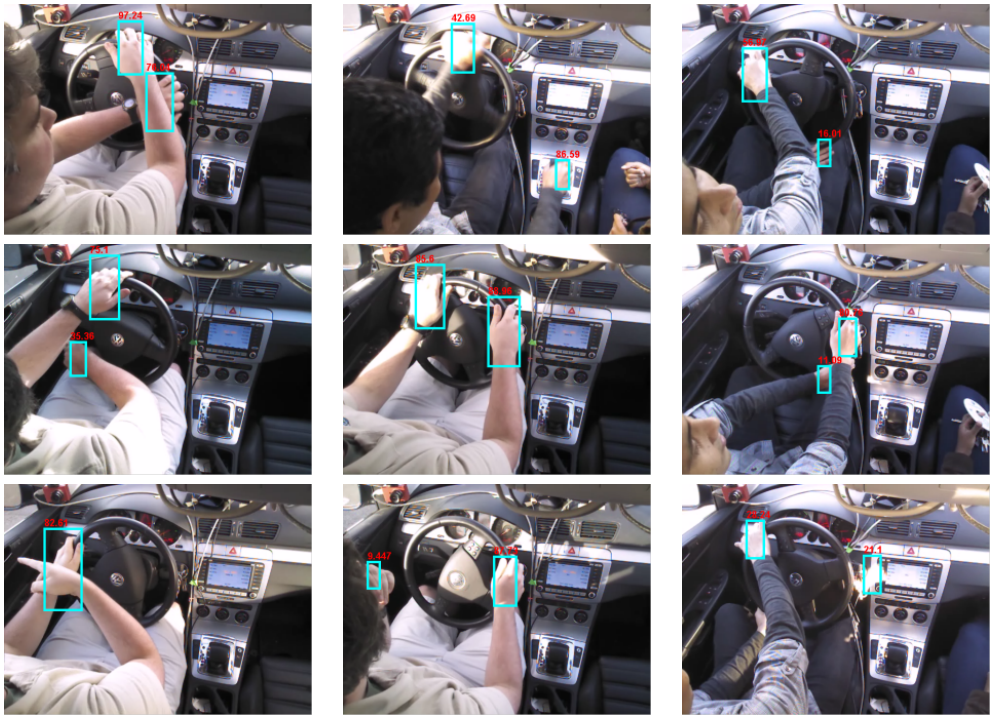


Figure 4.22: Depiction of successful detection results (top two rows) and challenging settings (bottom row). The method is shown to be robust to moderate occlusion by objects in the car, self-occlusion, variation in pose and rotation. Nonetheless, false positives still occur under heavy illumination variability. These are handled by tracking.

at 32 and increasing by a factor of 4 in each stage. Bootstrapping is performed at each stage, with hard negatives collected and used for re-training. We experimented with additional feature channels, such as different transformations for extracting skin colored pixels using a learned skin-likelihood classifier. We found no benefit over using the simple LUV color features (Fig. 4.21).

As mentioned, the hand detector runs at 30 fps on a CPU, which we found crucial for analyzing hours of captured video quickly. We noticed many of the false detections occurring in the proximity of the actual hand (the arm, or multiple detections around the hand). Therefore, window size and padding had a significant effect on false positive rates (see Fig. 4.21). Neighboring responses were removed using non-maximum suppression with a threshold of 0.2.

Left and right hand classification: Hand proposals provided by the hand detector are given to a binary linear Support Vector Machine (SVM) [245] for left and right hand classification. The already computed gradient features are used. Color cues were not shown to be beneficial for the left/right classification. Finally, detections are tracked using a standard Kalman filter.

Figure 4.23: Visualizing hand locations of entire drives. In red are left hand positions and in green are right hand positions. The scatter plots above show several hours of collected video.

4.12 Trajectory Learning

The output of the hand detector is used as part of an activity modeling framework. Common applications with trajectory studies (e.g. surveillance) involve a set of assumptions which may not hold in our data, such as a pre-defined number of points of 'entry' and 'exit' of the moving agents for defining the activities [305]. Furthermore, trajectories that are similar semantically may contain large performance variability. For example, turning maneuvers may begin or end anywhere on the wheel, with varying velocity profiles, or with one or two hands. At times, turning may produce a very slight change in hand positions, yet we would like to recognize such events. Temporal events of no motion, which usually provide temporal segmentation information, are also difficult to interpret. In our domain, such 'stop' states can occur during turns, lane changes, or regular driving, and are therefore not trivially defined. In addition to distinguishing among subtle and intricate movements, gesture performance is also effected by the *preferred neutral hand* position of the driver. Because of the uniqueness of the trajectories, we turn to a careful study of both the appropriate choice of trajectory features and the temporal modeling technique.

4.12.1 Trajectory Features

The following trajectory features are studied.

Position features: A signal of the position of the hands in each frame,

$$F_t^j = (f_{t-L+1}^j, \ldots, f_t^j) \tag{4.9}$$

Figure 4.24: A dataset of transition reaching and retracting gestures is used for the experiments. Left hand trajectories are shown in red and right hand trajectories are shown in green. Trajectory color encodes time, with brighter being more recent in the trajectory. Shown are reaching gestures to left side rest, gear, and instrument cluster.

with $j \in \{1, 2, 3, 4\}$ so that for each dimension of position and each hand we obtain a windowed time series (for a total of $L \times 4$ sized descriptor. That is, $f_t^j \in \{x_t^{left}, y_t^{left}, x_t^{right}, y_t^{right}\}$ which are the image plane positions provided by the hand detector. L is the trajectory length.

In addition to these, trajectory shape and dynamic information can be captured in the following features.

Displacement features: Given the component displacements at time t, $\Delta f_t = f_t - f_{t-1}$, the displacement features for the trajectory are

$$V_t^j = \Delta F_t^j = (\Delta f_{t-L+1}^j, \ldots, \Delta f_t^j) \tag{4.10}$$

Normalized displacement features: Inspired by [244], the displacement feature vector is normalized by the sum of the magnitudes of the displacement vector

$$\bar{V}_t^j = \frac{\Delta F_t^j}{\sum_{i=t-L+1}^{t} ||\Delta f_i^j||} \tag{4.11}$$

Transition histogram of displacements: Proposed in [306], this histogram descriptor utilizes quantization of the displacements in V into three levels of magnitude after normalization by the maximum displacement magnitude in the trajectory. Orientation is binned into 8 sectors of the unit circle, producing a total of 24 quantization bins. Finally, a zero displacement bin is added. A transition matrix counts the frequency of occurrence from the consecutive entries in V. The final histogram descriptor is therefore of size $25 \times 25 = 625$.

Temporal pyramid of Fourier coefficients: For each dimension of F, the short Fourier transform [255] is applied and the low frequency coefficients are used. The trajectory F is recursively partitioned into levels to further capture temporal structure of the trajectory. In our experiments, we use two levels of partitioning the original trajectory, as no gains were made by further partitioning.

4.12.2 Temporal Modeling

Characterization of trajectory paths involves learning of the temporal dynamics of the hand movement. Four supervised modeling techniques are compared. An SVM classifier is studied with a linear kernel and a non-linear RBF kernel [245]. Both the regularization parameter C and the spread parameter γ are grid optimized. As a classical benchmark for temporal modeling, a Hidden Markov Model (HMM) learned using the Baum-Welch method and expectation maximization (EM) [307] is also evaluated. The available implementation of [308] is used, and the number of states is optimized over $\{1, 3, 5, 7\}$. A more recent development over the HMM was demonstrated with Conditional Random Fields (CRF). We employ the Latent-Dynamic CRF (LDCRF) [309], which provides an advantage over HMM due to discriminative training.

4.13 Experimental Settings

The model and features will be evaluated in terms of three performance measures.

Activity classification: Each motion pattern is manually annotated with a starting frame and an end frame, interpolated to be the same size (a 20-dimensional vector), and classified into a pre-defined set of activities. The purpose of these experiments is to compare the performance of different features and classifiers. Cross-subject cross-validation is employed, where training and testing are done on disjoint subjects. Such cross-validation is employed in all of the tests below as well. Furthermore, we use **normalized accuracy** as the performance metric, where true positives in each class are normalized by the number of instances in the class before the final averaging. This takes care of unbalanced classes in evaluation.

Activity prediction: Assume an event annotation ending at a certain time, t_e. In prediction, we query the model δ seconds before t_e (i.e. at $t_e - \delta$) for a label given the sequence of observations $F_{t_e-\delta}$. There are two possible training procedures. In one, referred to as the **fixed model** procedure, only one model is trained over the annotated events once. That model is used for prediction at different δ values in testing. In the second procedure, referred to as the **shifted model**, a model is trained on samples shifted by δ (i.e. shifting δ involves re-training) and tested on the δ-shifted test samples. Both procedures allow for activity prediction, but the shifted model case requires the evaluation of multiple models corresponding to trajectory patterns specific to each choice of δ.

Abnormal event detection: Measuring the quality of the modeling can also be done on a semantic level. Can the models be used in order to distinguish critical events specific to our application domains? The important notion of 'abnormality' is a useful measure for evaluating the framework. It also allows for direct comparison with data-driven learning of models using unsupervised techniques. Traditionally, novelty detection is achieved by inspecting the scores provided by the temporal models. This is expressed in low log-likelihood scores for a CRF or HMM model. For the SVM models, we employ the point to hyper-plane distance as a confidence measure. SVM scores are normalized using a coupling

approach [245]. In all cases, a cursor for the maximum posterior probability is thresholded in order to detect an abnormal event,

$$\max_{c \in \{1,\ldots,C\}} P(c|F) < \epsilon_{abnormal} \tag{4.12}$$

in a C class problem. Due to the highly complex nature of the hand trajectories, unsupervised approaches for obtaining the motion path labels may also be of interest. We also evaluate a data-driven, unsupervised trajectory analysis framework using fuzzy C-means clustering [305] and a outlier-aware K-means algorithm. In the latter case, the standard K-means iteration is performed, but at every step we use the Euclidean distance in order to discard samples that are distant from the centroid of the clusters before updating of the new centroids. The number of samples to discard is chosen according to a parameter which is fixed in each iteration. Both of the clustering algorithms contain a notion of outliers, which is essential for learning models for abnormality detection.

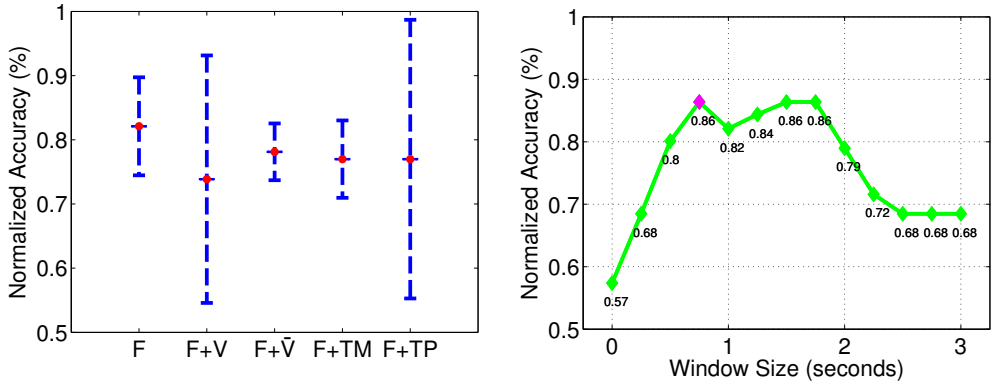
4.14 Experimental Evaluation

In order to evaluate the framework a video dataset composed of over an hour of driving was used. The analysis is focused on hand motion patterns which are clearly defined and are important for the study of attentiveness-reaching and retracting trajectories. Reaching motions involve hand-object interaction associated with secondary in-vehicle tasks.

Dataset: A total of 60 trajectory instances were annotated in terms of start and end, focusing on transition motions. Six classes were defined among the four regions of wheel, instrument cluster, gear shift, side rest. All trajectories must initiate or terminate on the wheel. Visualization of some of the samples is shown in Fig. 4.24. Increasingly intricate motion patterns can be defined in the future. All experiments employ cross-subject cross validation, where training and testing is performed on disjoint subjects.

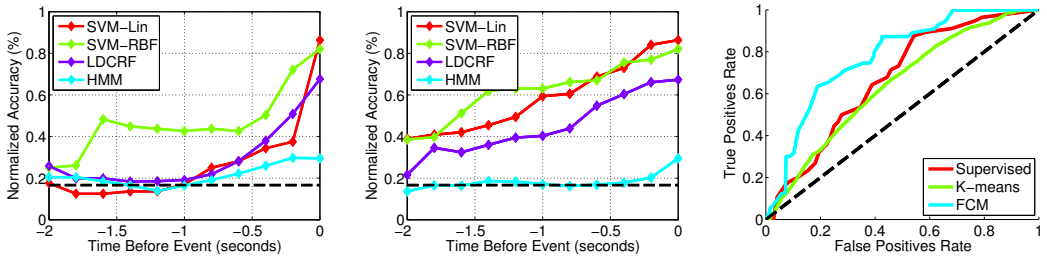
For abnormal event detection, 36 events of abnormal activity were annotated. These are events that are semantically abnormal when compared to the previous six classes of gestures which are commonly performed while driving. These include rear-mirror adjustment, driver touching the face, and driver reaching back over the shoulder to inspect and perform a reverse maneuver. Most of these involve a hand motion that is not only when abnormal compared to the six defined gesture classes, but might also be considered abnormal in certain driving scenarios (e.g. on a highway). User-specific event definition and study is left for future work.

Feature analysis: On the transition motions dataset, position features alone were shown to work well out of the five types of trajectory features studied, with no clear benefit by the explicit addition of dynamic features. The analysis is shown in Fig. 4.25(a) in terms of the average normalized accuracy and standard deviation over the cross validation. Furthermore, given an event annotation, the optimization for the window size L to include in computation of the trajectory features is shown in Fig. 4.25(b). Both the



(a) (b)

Figure 4.25: Evaluation of the trajectory features studied in activity classification. (a) Position features (**F**) are shown to work well. The abbreviations are: $\boldsymbol{V}$-displacement features, $\bar{\boldsymbol{V}}$-normalized displacement features, **TM**-transition histogram of displacements, and **TP**-temporal pyramid of Fourier coefficients. (b) Given the annotated end of a gesture, we optimize for the temporal window size L of the time series. A 0.75 seconds window is shown to work best, and is used in the activity prediction experiments.



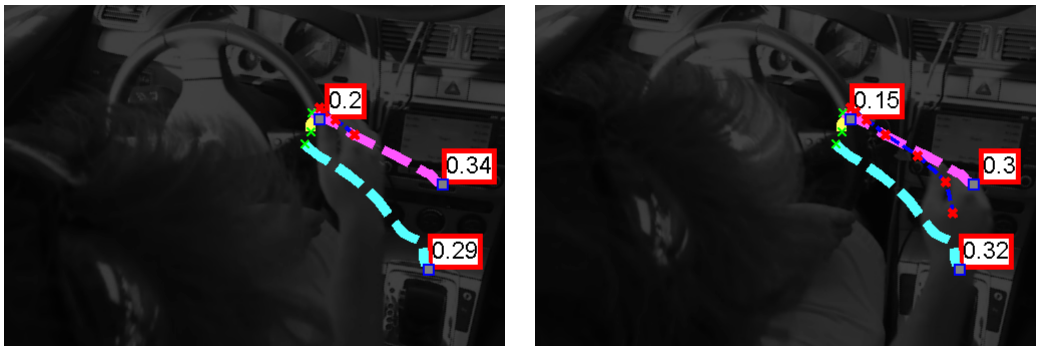
(a) Fixed Model (b) Shifted Model (c) Abnormal Event Detection

Figure 4.26: Evaluation of the four modeling techniques in terms of predictive power is shown in (a) and (b). The black line depicts the random guess case. In fixed model, one model is obtained by training once using the annotated events. In shifted model, a model is learned for each δ time before an event. (c) Detecting abnormal hand activities using supervised clusters, or data-driven unsupervised clusters using K-means with outlier removal or fuzzy C-means (FCM).

position features and a window size of 0.75 seconds are employed for the remainder of the experiments. The results were produced with a linear SVM.

Temporal modeling: The four classification techniques are evaluated in terms of predictive power. As mentioned in Section 4.13, prediction can occur using two procedures. Overall trends are similar in both of the procedures, as shown in Fig. 4.26. In fixed model, where one model is trained on the annotated event end ($\delta = 0$) and evaluated at different δ values to produce predictions, an SVM with an RBF kernel is shown to work best, while a linear SVM tops for classification of the gestures at $\delta = 0$. The trend is similar for the shifted model procedure (Fig. 4.26(b)), yet prediction rates improve overall due to the training on the shifted time series. Common ambiguous trajectories occur in reaching gestures, where a hand may reach towards the lower part of the instrument cluster or the gear shift. An example is shown in Fig. 4.27(b).

(a) Prediction of wheel to instrument panel reaching.



(b) Prediction of wheel to gear shift reaching.

Figure 4.27: Early classification of hand motion patterns. In blue is the current and previous hand trajectory (with the actual corresponding frame shown for each instance). Red crosses depict the previous hand locations in the trajectory. We plot the top three trajectories (centroids by averaging) matching to the current trajectory with the SVM probability score. Only right hand information is shown for clarity. In (a), notice how a large horizontal trajectory from the left part of the wheel is classified correctly as towards instrument cluster. In (b) note how a more difficult sample is first classified incorrectly as towards instrument cluster, but as more information becomes available the gear reaching label is correctly predicted.

Abnormal event detection: The preliminary results in Fig. 4.26(c) shows the data-driven approach with a membership threshold using fuzzy C-means works best. In the future, unsupervised discovery of events would be essential for representing user-specific motion patterns, such as a driver's neutral hand position.

4.15 Chapter Concluding Remarks

Observing and understanding hands is crucial for human-machine interactivity. This chapter of the thesis summarized algorithmic and experimental considerations when looking at hands, specifically for understanding in-cabin hand gestures. The dataset has been publicly released as part of an on-going challenge at the IEEE Conference on Computer Vision and Pattern Recognition and IEEE Intelligent Vehicles Symposium, which is how the VIVA (Vision for Intelligent Vehicles and Applications) challenge came to be. A main part of my research has been concerned with understanding human hands, from

contextual detection of infotainment activity and up to temporal modeling of hand gestures. Therefore, this chapter is strategically placed and discussed to set the stage for a deeper discussion on behavior sensing and modeling. The next chapter will include hand gestures, but extend the work to additional cues and situations.

A summary of this chapter is detailed below,

- We studied the feasibility of an in-vehicle, vision-based gesture recognition system. First a hand detection and user determination step was used, followed by a real-time spatio-temporal descriptor and gesture classification scheme. A careful evaluation of different temporal descriptors showed the challenging nature of the dataset, with RGBD fusion proving to be beneficial for recognition. Future extensions should further analyze the role of each of the spatio-temporal descriptors in increasing illumination-, occlusion-, and subject-invariance of the system. Temporal segmentation of gestures without requiring the hand to leave the ROI may result in a more comfortable interface to use. The studied RGBD feature set might be useful for studying other activities in the vehicle or general action recognition applications [236, 127].

- This work studied vision-based hand activity analysis in naturalistic settings. In order to tackle the intricate nature of the trajectory problem, multiple temporal trajectory features and classification schemes were studied in supervised settings. The transition gestures studied and other visual-manual tasks may be correlated with head cues [18], and their integration will be studied next.

This chapter is in part a reprint of material that is published in the IEEE Transactions on Intelligent Transportation Systems (2014), by Eshed Ohn-Bar, and Mohan M. Trivedi. The dissertation author was the primary investigator and author of this paper.

This chapter is in part a reprint of material that is published in the IEEE Intelligent Transportation Systems Conference (2014), by Eshed Ohn-Bar, and Mohan M. Trivedi. The dissertation author was the primary investigator and author of this paper.

Chapter 5

Multi-Cue Behavior Modeling, with Applications to Driver Assistance

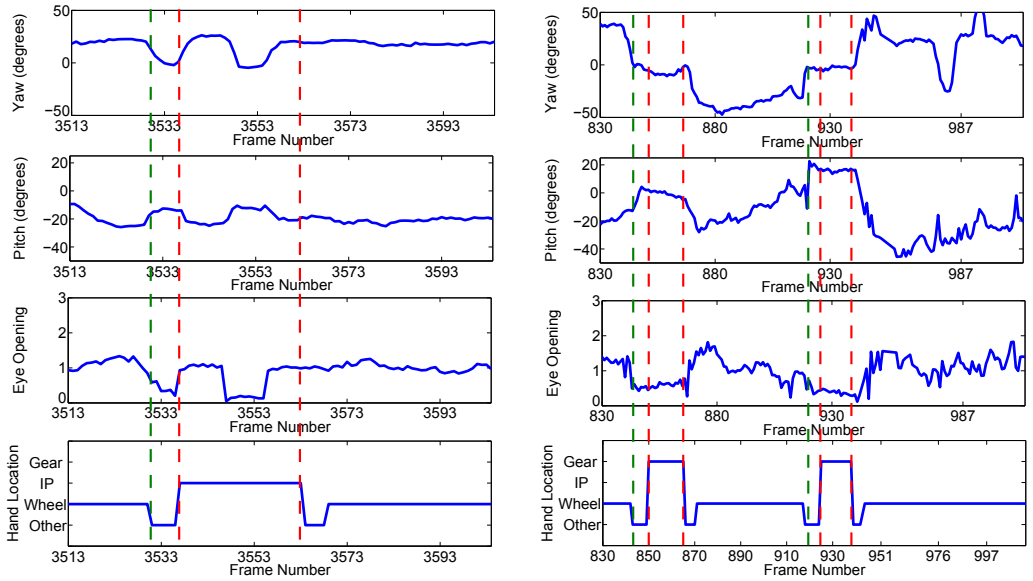
This chapter utilizes the contextual detection and hand activity recognition modules described in the previous chapters in order to perform a deeper study of behavior and its prediction for safety-critical applications. Effective behavior modeling is holistic, spanning multiple cues and their coordination over time. First, we study coordination of hand, head, and eye activity for infotainment and interactivity applications. Second, we extend to a general study of the complex interplay between driver (hand, head, and foot), vehicle (speed, yaw-rate, etc.), and surround spatio-temporal context (agents, scene information) cues for understanding and predicting activity.

5.1 Hand, Head, and Eye Coordination Model

In this section, we discuss a fusion framework for modeling spatio-temporal context and dependencies among hand, head, and eye cues for representing activity. We propose a multiview, multimodal vision framework in order to characterize driver activity based on head, eye, and hand cues. Leveraging the three types of cues allows for a richer description of the driver's state and for improved activity detection performance. First, regions of interest are extracted from two videos, one observing the driver's hands and one the driver's head. Next, hand location hypotheses are generated and integrated with a head pose and facial landmark module in order to classify driver activity into three states: wheel region interaction with two hands on the wheel, gear region activity, or instrument cluster region activity. The method is evaluated on a video dataset captured in on-road settings.

Secondary tasks performed in the vehicle have been shown to increase inattentiveness [235], which, in 2012 was a contributing factor in at least 3092 fatalities and 416,000 injuries [313]. According to a recent survey, 37% of the drivers admit to having sent or received text messages, with 18% doing so

Figure 5.1: Hand, head, and eye cues can be used in order to analyze driver activity. Notice the guiding head movements performed in order to gather visual information before and while the hand interaction occurs.



(a) Instrument Panel Interaction (b) Gear Shift Interaction

Figure 5.2: Hand, head, and eye cue visualization for (a) an instrument cluster activity sequence and (b) gear shift activity sequence. Green line: indication of start of head and eye cues (yaw, pitch, and opening) before the hand activity. Red lines: start and end of the hand activity. See Section 5.2.2 for further detail on the cues.

regularly while operating a vehicle [314]. Furthermore, 86% of drivers report eating or drinking (57% report doing it sometimes or often), and many reported common GPS system interaction, surfing the internet, watching a video, reading a map, or grooming.

Because of the above issues, on-road analysis of driver activities is becoming an essential component for advanced driver assistance systems. Towards this end, we focus on analyzing where and what

hands do in the vehicle. Hand positions can provide the level of control drivers exhibit during a maneuver or can even give some information about mental workload [315]. Furthermore, in-vehicle activities involving hand movements often demand coordination with head and eye movements. In fact, human gaze behavior studies involving various natural dynamic activities including driving [316, 317], typing [318], walking ([319]), throwing in basketball [320], batting in cricket ([321]) etc., suggest a common finding that gaze shifts and fixations are controlled pro actively to gather visual information for guiding movements.While specific properties of the spatial and temporal coordination of the eye, head and hand movements are influenced by the particular tasks, there is strong evidence to suggest that the hand usually waits for the eyes either for the target selection or for the visual guidance for the reach, or both [322]. For this, a distributed camera setup is installed to simultaneously observe hand and head movements.

The approach is purely vision-based, with no markers or intrusive devices. There are several challenges that such a system must overcome, both for the robust extraction of head [323] and hand cues [127]. For the head, there are challenges of self-occlusion due to large head motion and volatile illumination conditions during driving. Furthermore, the hand detection is challenging as the human hand is highly deformable and tends to occlude itself in images. The problem is further complicated by the vehicular requirement for algorithms to be robust to changing illumination. Most existing works involve hand detection under indoor or naive settings. Under such constraints, the hand may be the main salient object in the scene or exhibiting the most motion [324], skin-color techniques may be used [325], or a depth-based threshold could provide the main cue [326]. As single cues, such techniques were shown to perform poorly on our dataset [133]. Therefore, a main emphasis in this work is towards the robust localization of hands in the scene. In particular, we are interested to know whether the hand is engaged in a specific region of the vehicle or not, and how many hands are on the wheel.

The framework in this work leverages two views for driver activity analysis, a camera looking at the driver's hand and another looking at the head. The multiple views framework provides a more complete semantic description of the driver's activity state [327]. As shown in Fig. 5.3, these are integrated in order to produce the final activity classification. First, the hand detection technique is discussed, then a detailed description of relevant head and eye cues is given, followed by a description of head, eye and hand cueintegration scheme. Lastly, experimental evaluations is presented on naturalistic driving.

5.2 Feature Extraction Modules

The proposed framework leverages two views for activity analysis, a camera looking at the driver's hand and another looking at the head. As shown in Fig. 5.3, each of these provide a zone of activity. First, we detail the gaze zone cues used for activity analysis. These are head and eye features extracted from the head view. Three zones are defined: the wheel, the gear, and the instrument cluster. Next we discuss the hand zone cues. Hand activity is detected in each zone using a visual descriptor and a classifier, which provides a probability output for each zone. The zones are fused to produce a hand-only three class activity classification. Finally, the hand-only probability cues and the gaze zone cues are

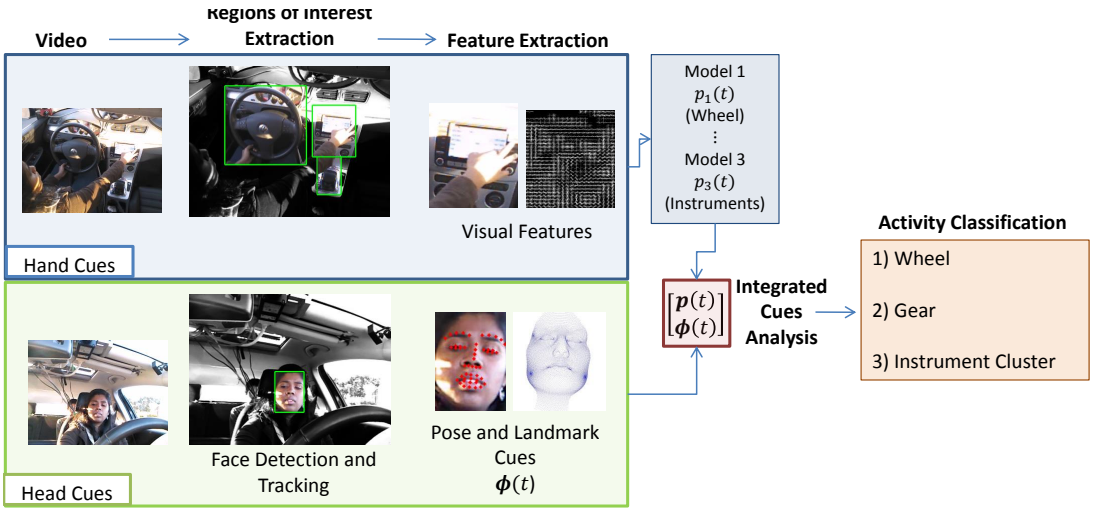


Figure 5.3: The proposed approach for driver activity recognition. Head and hand cues are extracted from video in regions of interest. These are fused using a hierarchical Support Vector Machine (SVM) classifier to produce activity classification.

Table 5.1: Driver activity recognition dataset collected. Training and testing is done using cross-subject cross-validation.

Subject	Video Time (min)	# Samples Annotated	Head	Environment	Time	Vehicle	# Activity Classes
1	13:11	3195		Sunny	12pm	LISA-Q	5
2	18:00	2574		Sunny	12pm	LISA-Q	5
3	9:08	10115	✓	Sunny	4pm	LISA-X	4
4	10:05	4491	✓	Sunny	5pm	LISA-X	4

Table 5.2: Types of activities in the dataset collected.

Loction	Activity Types
Radio	On/Off Radio
	Change Preset
	Navigate to Radio Channel
	Increase/Decrease Volume
	Seek/Scan for Preferred Channel
	Insert/Eject CD
	On/Off Hazard Lights
Climate Control	On/Off AC
	Adjust AC
	Change Fan Direction
Side Rest	Adjust Mirrors
Gear	Park/Exit Parking

integrated in order to produce the final activity classification.

5.2.1 Hand Cues

In the vehicle, hand activities may be characterized by zones or regions of interest. These zones (see Fig. 5.3) are important for understanding driver activities and secondary tasks. This motivates scene representation in terms of these salient regions. Additionally, structure in the scene can be captured by leveraging information from the multiple salient regions. For instance, during interaction with the instrument cluster, visual information from the gear region can increase the confidence in the current activity recognition, as no hand is found on the gear shift. Such reasoning is particularly useful under occlusion, noise due to illumination variation, and other visually challenging settings [133]. In [127, 121], edge, color, texture, and motion features were studied for the purpose of hand activity recognition. Since we found that edge features were particularly successful, in this work we employ a pyramidal representation for each region using Histogram of Oriented Gradients (HOG) [284], with cell sizes 1 (over the entire region), 4, and 8 for a $8 + 128 + 512 = 648$ dimensional feature vector.

5.2.2 Head and Eye Cues

As motivated earlier, knowing where the driver is looking can provide important cues about any on-going driver activities. While precise gaze information is ideally preferred, its estimation is very challenging, especially when using remote eye tracking systems in a real-world environment such as driving. However, a coarse gaze direction, i.e. gaze zone, is often sufficient in a number of applications, and can be relatively robustly extracted in driving environments [32]. In the case of driver activity recognition, the temporal dynamics of gaze zone can provide important cues.

To infer a driver's gaze zone, we use head-pose and eye-state. With recent advancements in facial feature tracking methods [328, 329], in our implementation, we have used facial features-based geometric approach for head pose estimation. It has shown robust performance in the driving environment with good accuracy. For implementation details, we encourage the reader to refer to [330] by Tawari *et al.* An additional benefit of using facial features for estimating head pose is that it allows for facial landmark analysis, such as level of eye opening. Head pose alone provides a good approximation of gaze zone, but neighboring zones (e.g. instrument cluster region and gear region) are often confused [32]. In such cases, eye-state such as eye-opening can help to disambiguate between confusing zones. In our implementation, the eye state at time t is estimated using two variables: area of the eye and area of the face. Area of the eye is the area of a polygon whose vertices are the detected facial landmarks around the left or right eye. Similarly, the area of the face is the area of the smallest polygon that encompass all the detected facial landmarks. To compute the level of eye opening, we divide area of the eye by the area of the face at every time t. This normalization will allow the computation of eye opening to be invariable to driver's physical distance to the camera, where closer distances makes the face appear larger in the image plane. Finally, a normalization constant learned for each driver representing his or her normal eye-opening state is used such that after normalization values < 1 represent downward glances and values > 1 represent upward glances (visualized in Fig. 5.4.

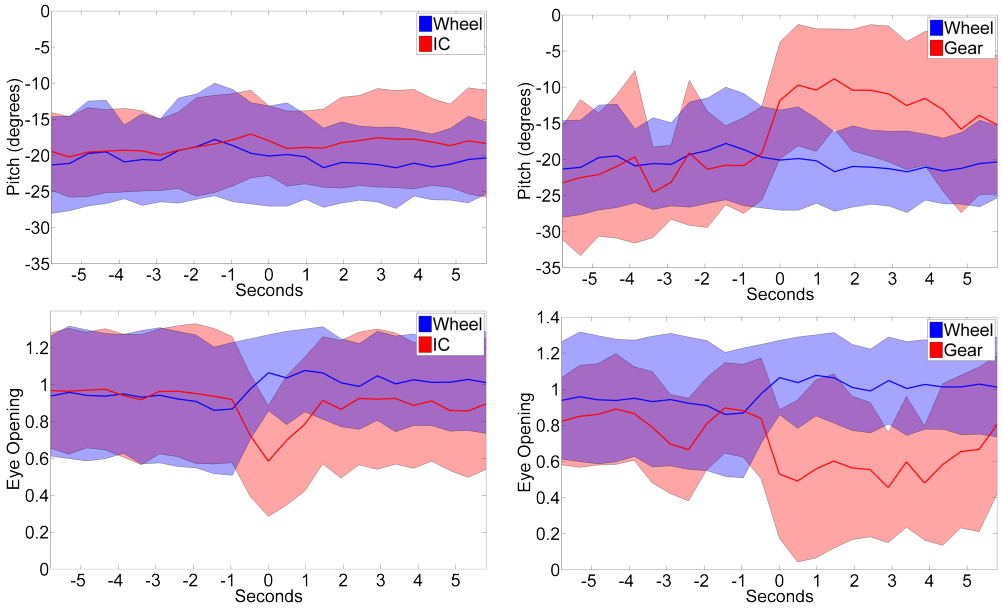
The eye-opening cue in addition to head pose, has potential in differentiating between glances towards the instrument cluster and glances towards the gear, as shown in Fig. 5.4. Figure 5.4 shows the mean (solid line) and standard deviation (semi-transparent shades) of two features (i.e. head pose in pitch and eye opening) for three different driver activities, using the collected naturalistic driving dataset. The feature statistics are plotted 6 seconds before and after the start of the driver hand activity, where time of 0 seconds represents the start of the activity. Using the eye opening cues alone, we can observe that when the driver is interacting with the instrument cluster he or she glances towards the IC at the start of the interaction. However, when the driver is interacting with the gear, while there is some indication of a small glance before the start of the activity, there is significant glance engagement with the gear region after the start of the event. Formally,the eye state at time t is given by,

$$e(t) = \frac{A_{eye}(t)}{A_{face}(t) \times e_o}, \tag{5.1}$$

where, $A_{eye}(t)$ is the area of the convex hull of the fudicial points around the eyes, $A_{face}(t)$ is the area of the convex hull of all the facial landmarks (as illustrated in 5.3), and E_o is a normalization constant learned for each driver to represent his or her normal eye-opening state. The final feature vector $\boldsymbol{\phi}(t)$ at time t consists of head pose (yaw and pitch) and eye-state.

The eye-opening cue in addition to head pose, has potential in differentiating between glances towards the instrument cluster and glances towards the gear, as shown in Fig. 5.4. Figure 5.4 shows the mean (solid line) and standard deviation (semi-transparent shades) of two features (i.e. head pose in pitch and eye opening) for three different driver activities, using the collected naturalistic driving dataset. The feature statistics are plotted 6 seconds before and after the start of the driver hand activity, where time of 0 seconds represents the start of the activity. Using the eye opening cues alone, we can observe that when the driver is interacting with the instrument cluster he or she glances towards the IC at the start of the interaction. However, when the driver is interacting with the gear, while there is some indication of a small glance before the start of the activity, there is significant glance engagement with the gear region after the start of the event.

Driver interactions with the infotainment system and the gear show unique pattern combination with head pose, eye opening and hand locations. Figure 5.4 shows time synchronized plots of head pose, eye opening, hand activity for two typical events: interacting with instrument cluster and interacting with gear. In Fig. 5.4 head pose in yaw and pitch are measured in degrees, where a decreasing value in yaw represents the driver looking rightward and an increasing value in pitch represents the driver looking downward. In the plot for eye opening, a value of 1 represents the normal size of eyes, values greater than one could represent looking upward, and values less than one could represent looking downward. Hand locations in the image plane are also plotted in a time-synchronized manner. The green dotted line indicates the start of head and eye cues before the hand movement. The dotted red lines indicates the start and end of the hand movement. These plots show the presence of hand, head and eye movements while the driver interacts with the infotainment system (Fig. 5.4(a)) and with the gear (Fig. 5.4(b)). While



(a) (b)

Figure 5.4: Head and eye cue statistics visualization for (a) instrument cluster (IC) activity sequences against normal wheel interaction sequences and (b) gear shift activity sequences against normal wheel interaction sequences. Time $t = 0$ represents the start of the respective driver activity. The blue and red line represent the mean statistics of respective cues (i.e. head pose in pitch, eye opening) for 6 seconds before and after the start of the driver hand activity. The lighter shades around the solid line indicate the standard deviation from the respective mean statistics.

the latency of each cue is circumstantial, we experimentally validate the presence of head and eye cues strengthen activity recognition.

As the above cues may occur before or after an associated hand cue (i.e. looking and then reaching to the instrument cluster), the head and eye features are computed over a temporal window. Let $\boldsymbol{h}(t)$ represent the features containing the head pose (in pitch, yaw and roll in degrees) and the level of eye opening (for both left and right eye) at time t and δ be the size of the time window to be used for temporal concatenation. Then, the time series $\boldsymbol{\phi}(t) = [\boldsymbol{h}(t-\delta), \ldots, \boldsymbol{h}(t)]$ is the feature set extracted from the head view at time t to be further used in the integration with hand cues.

5.3 Activity Recognition Framework

In this section, we detail the learning framework for fusion of the two views and performing activity classification. The classifier used is a linear kernel SVM [331], and fusion is done using a hierarchical SVM which produces the final activity classification.

Because the hand and head cues are different in nature, first a multiclass Support Vector Machine (SVM) [332] is trained to produce activity classification based on the hand view region features only. A weight, $\mathbf{w}_i$ is learned for each class $i \in \{1, \ldots, n\}$ where n is the number of activity classes. In this

work, we focus on three activity classes: 1) Wheel region interaction with two hands on the wheel; 2) Gear region interaction; 3) Instrument cluster interaction. The weights for all of the classes are learned jointly, and classification can be performed using

$$i^{\star} = \underset{i \in \{1,\ldots,n\}}{\arg\max}\ \mathbf{w}_i^T \mathbf{x} \tag{5.2}$$

where **x** is the feature vector from all the regions in the hand view.

In order to measure the effectiveness and complementarity of the hand and head cues, activity recognition will be studied using hand-only cues and integrated hand and head cues. Hand cues can be summarized using normalized scores,

$$p(i|\mathbf{x}) = \frac{\exp\left(\mathbf{w}_i^T \mathbf{x}\right)}{\sum_j \exp\left(\mathbf{w}_j^T \mathbf{x}\right)} \tag{5.3}$$

These posterior probabilities can be calculated at every frame and are abbreviated in Fig. 5.3 as p_i. For the fusion of the hand and head views, the hand cues are concatenated with the windowed signal of head features to produce the feature set at time t,

$$\mathbf{x}(t) = \begin{pmatrix} p_1(t) \\ \vdots \\ p_n(t) \\ \boldsymbol{\phi}(t) \end{pmatrix} \tag{5.4}$$

The fused feature vector is given to a hierarchical second-stage multiclass SVM to produce the activity classification.

The classes in our dataset are unbalanced. For instance, one activity class such as wheel region two-hands on the wheel may occur in the majority of the samples. Nonetheless preserving all of the samples for the wheel region in training could be beneficial in producing a robust classifier which can generalize over the large occlusion and illumination challenges occurring in the wheel region. Therefore, we also incorporate a biased-penalties SVM [333], which adjusts the regularization parameter in the classical SVM to be proportional to the class size in training.

5.4 Experimental Evaluation and Discussion

The proposed driver hand activity recognition framework is evaluated on naturalistic driving data from multiple drivers. Using hand annotated ground truth data of driver hand activity, we show promising results of integrating head and hand cues.

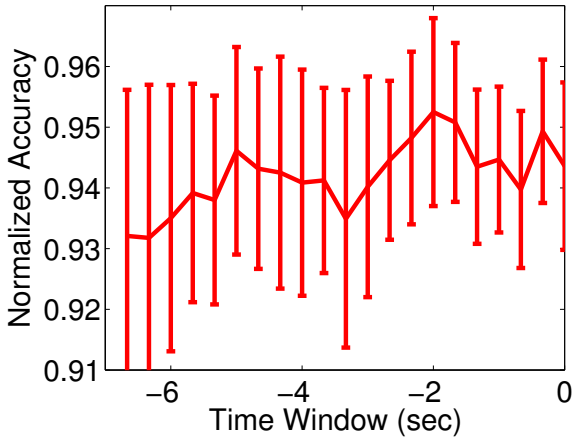
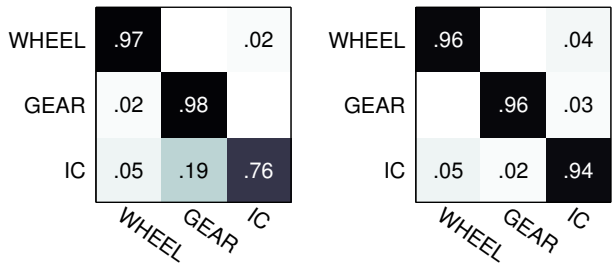


Figure 5.5: Effect of varying the time window before an event definition for the head cues. Normalized accuracy (average of the diagonal of the confusion matrix) and standard deviation for activity classification is reported after integration with hand cues.



(a) Hand Only (**90%**) (b) Hand+Head (**94%**)

Figure 5.6: Activity recognition based on hand only cues and hand+head cue integration for three region activity classification. **IC** stands for instrument cluster.

5.4.1 Experimental Setup and Dataset Description

The naturalistic driving dataset is collected using two cameras, one observing the driver's hands and another observing the driver's head. Multiple drivers (three male and one female) of varying ethnicity and varying age from 20 to 30, as well as varying driving experience participated in this study. Before driving, each driver was instructed to perform, at his or her convenience, the following secondary tasks any number of times and in any order of preference:

- *Instrument cluster (IC) region activities:* On/off radio, change preset, navigate to radio channel, increase/decrease volume, seek/scan for preferred channel, insert/eject a CD, on/off hazard lights, on/off/adjust climate control.
- *Gear region activities:* Observed while parking and exiting parking.
- *Wheel region activities:* Observed under normal driving conditions.

The drivers practiced the aforementioned activities before driving in order to get accustomed to

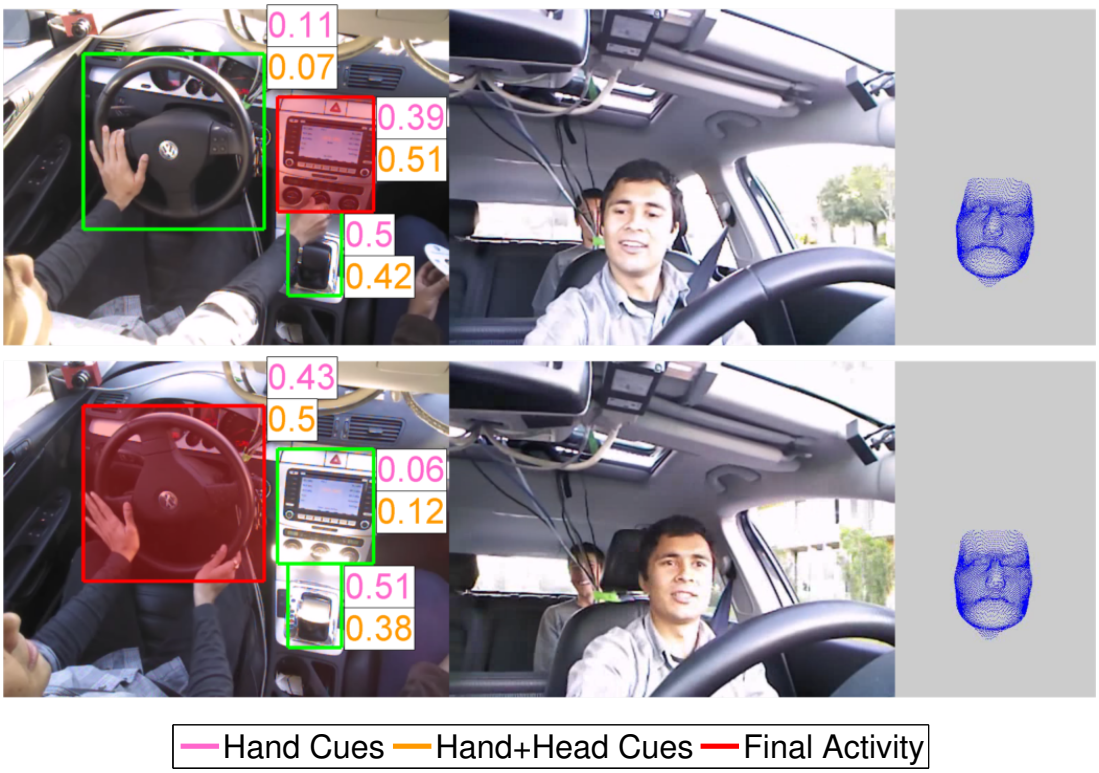


Figure 5.7: Visualization of the advantage in integrating head, eye, and hand cues for driver activity recognition. We show the hand view, head view, and the fitted head model. In purple are the probabilities of the activity based on hand cues alone. In orange are the rescored values using a hierarchical SVM and head and eye cues. Note how in the above scenarios, the incorrect hand-based predictions were corrected by the rescoring based on head and eye cues.

the vehicle. In addition, instructors also prompted the drivers to instigate these activities randomly but cautiously. Driving was performed in urban, high-traffic settings.

Ground truth for evaluation of our framework is obtained from manual annotation of the location of driver's hands. A total of $11,147$ frames from many number of driver activities during the drives were annotated: 7429 frames of two hands in the wheel region for wheel region activity, 679 frames of hands on the gear, and 3039 frames of interaction in the instrument cluster region. As the videos were collected in sunny settings at noon or the afternoon, they contain significant illumination variation that is both global and local (shadows). With this dataset, all testing is performed by cross subject test settings, where the data from one subject is used for testing and the rest for training. This ensures generalization of the learned models.

5.4.2 Evaluation of Hand and Head Integration

Capturing the temporal dynamics of head and hand cues is evaluated in terms of activity classification out of a three class problem: 1) Wheel region interaction with two hands on the wheel; 2) Gear region interaction; 3) Instrument cluster interaction. Hand cues may be used alone, with results shown in Fig. 5.6(a). The results are promising, but instrument cluster and gear classification are sometimes confused due to the arm presence in the gear region while interaction occurs with the instrument cluster. Furthermore, under volatile illumination changes the method may also fail.

Incorporating head cues is shown to resolve some of the challenges, as depicted in Fig. 5.6(b). In order to capture head and hand cue dynamics, head and eye cues are calculated over a temporal window in order to generate $\boldsymbol{\phi}(t)$, the final head and eye feature vector at time t. The effect of changing the time window are shown in Fig. 5.5. We notice how increasing the window size of up to two seconds improves performance, after which results decline. With a large temporal window, the cue becomes less discriminative and also higher in dimensionality, which explains the decline. Nonetheless, we expect a peak in results for a window size larger than one entry, as head and hand cues may be temporally delayed. For example, a driver may look first and then reach towards the instrument cluster or gear shift.

Fig. 5.7 visualizes some example cases where hand cues provide ambiguous activity classification due to visually challenging settings, yet these are resolved after the predictions are rescored with the second stage hierarchical SVM and head and eye cues. For each of the depicted scenarios, the hand view, head view, and the fitted head models are shown. Using the hand cue prediction (shown in the purple probabilities) would have resulted in an incorrect activity classification. For instance, some of the hand enters the gear shift while still interacting with the instrument cluster in the top figure. This leads to a wrong prediction using hand cues, but pitch and head information rescore the probabilities and correctly classify the activity (final classification after integration is visualized with a red transparent patch). Illumination variation may also cause incorrect activity classification based on hand cues alone, as shown in Fig. 5.7.

For the three region classification problem, head pose and landmark cues exhibit a distinctive pattern over the temporal window. A large window to include the initial glance before reaching to the instrument cluster or the gear shift as well as any head motions during the interaction significantly improves classification as shown in Fig. 5.6. Mainly, the gear shift and instrument cluster benefit from the integration.

5.5 Modeling Driver, Vehicle, and Surround for Holistic On-road Maneuver Prediction

We study techniques for monitoring and understanding real-world human activities, in particular of drivers, from distributed vision sensors. Thus far, we proposed a framework for leveraging both a hand and head view in order to provide activity recognition for interactivity. Integration provided improved activity recognition results and allowed for a more complete semantic description of the driver's activity state. A set of in-vehicle secondary tasks performed during on-road driving was utilized to demonstrate the benefit for such an approach, with promising results.

Next, we extend our study to include additional temporal cues, including driver foot movement cues, scene information, and ego-vehicle sensors, for real-time and early prediction of maneuvers, specifically overtake and brake events. Our study in this particular domain is motivated by the fact that early knowledge of driver behavior, in concert with the dynamics of the vehicle and surrounding agents, can help to recognize dangerous situations. Furthermore, it can assist in developing effective warning and driver assistance systems. Multiple perspectives and modalities are captured and fused in order to achieve a comprehensive representation of the scene. Temporal activities are learned from a multi-camera head pose estimation module, hand and foot tracking, ego-vehicle parameters, lane and road geometry analysis, and surround vehicle trajectories. The system is evaluated on a challenging dataset of naturalistic driving in real-world settings.

Distributed camera and sensor networks are needed for studying and monitoring agent activities in many domains of application [334]. Algorithms that reason over the multiple perspectives and fuse information have been developed with applications to outdoor or indoor surveillance [335]. In this work, multiple real-time systems are integrated in order to obtain temporal activity classification of video from a vehicular platform. The problem is related to other applications of video event recognition, as it requires a meaningful representation of the scene. Specifically, event definition and techniques for temporal representation, segmentation, and multi-modal fusion will be studied. These will be done with an emphasis on speed and reliability, which are necessary for the real-world, challenging application of preventing car accidents and making driving and roads safer. Furthermore, in the process of studying the usability and discriminative power of each of different cues, we gain insight into the underlying processes of driver behavior.

In 2012 alone, 33,561 people died in motor vehicle traffic crashes in the United States [336]. A majority of such accidents occurred due to an inappropriate maneuver or a distracted driver. In this work, we propose a real-time holistic framework for on-road analysis of driver behavior in naturalistic settings. Knowledge of the surround and vehicle dynamics, as well as the driver's state will allow the development of more efficient driver assistance systems. As a case study, we look into two specific maneuvers in order to evaluate the proposed framework. First, overtaking maneuvers will be studied. Lateral control maneuvers such as overtaking and lane changing represent a significant portion of the total accidents each year. Between 2004 and 2008, 336,000 such crashes occurred in the US [337]. Most of these

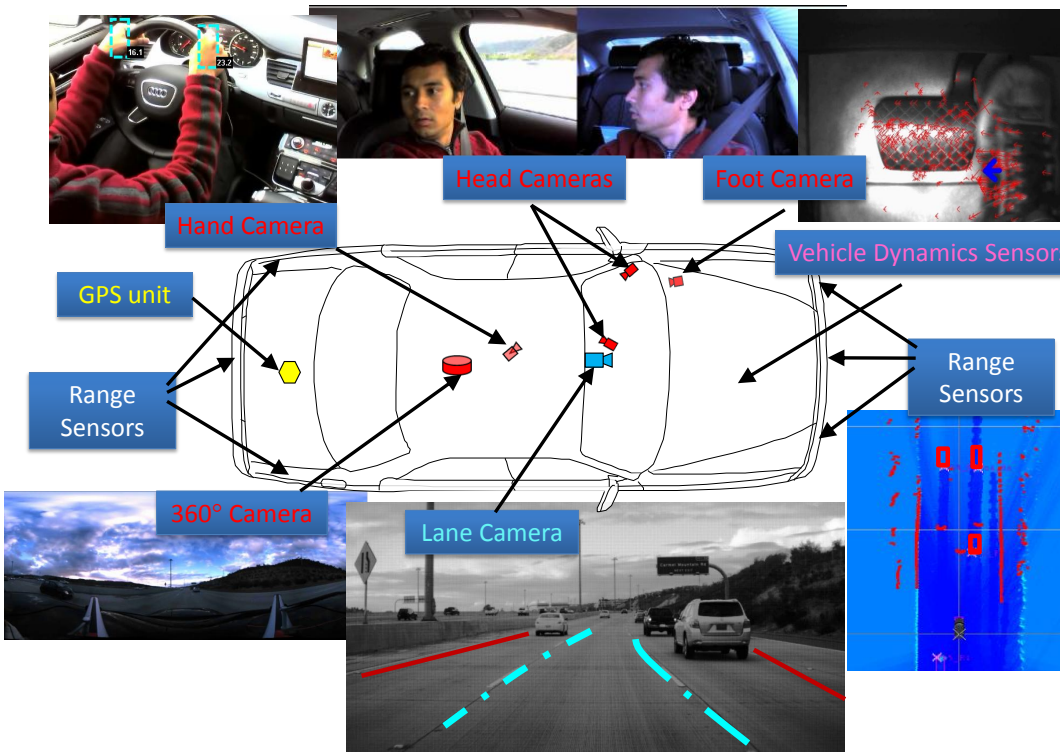


Figure 5.8: Distributed, synchronized network of sensors used in this study. A holistic representation of the scene allows for prediction of driver maneuvers. Knowledge of events a few seconds before occurrence and the development of effective driver assistance systems could make roads safer and save lives.

occurred on a straight road at daylight, and most of the contribution factors were driver related (i.e. due to distraction or inappropriate decision making). Second, we look at braking events, which are associated with longitudinal control and their study also plays a key role in preventing accidents. Early recognition of dangerous events can aid in the development of effective warning systems. In this work we emphasize that the system must be extremely robust in order to: 1) Engage only when it is needed by maintaining a low rate of false alarm rate, 2) Function at a high true positive rate so that critical events, as rare as they may be, are not missed. In order to understand what the driver intends to do, a wide range of vision and vehicle sensors are employed to develop techniques that can satisfy real-world requirements.

The requirement for robustness and real-time performance motivates us to study feature *representation* as well as techniques for *recognition* of temporal events. The study will focus on three main components: the vehicle, the driver, and the surround. The implications of this study are numerous. In addition to early warning systems, knowledge of the state of driver allows for customization of the system to the driver's needs, thereby mitigating further distraction caused by the system and easing user acceptance. On the contrary, a system which is not aware of the driver may cause annoyance. Additionally, under a dangerous situation (e.g. overtaking without turning on the blinker), a warning could be conveyed

Table 5.3: Overview of selected studies performed in real-world driving settings (i.e. as opposed to simulator settings) for maneuver analysis.

Study	Maneuvers	Inputs*	Method
McCall and Trivedi [107] (2007)	Brake	E,He,R,F	Relevance Vector Machine (RVM)
Doshi *et al.* [105] (2011)	Lane-change†	E,He,L,R	RVM
Tran *et al.* [2] (2012)	Brake	F	Hidden Markov Model (HMM)
Cheng *et al.* [146]	Turns	E,He,Ha	HMM
Pugeault and Bowden [67] (2010)‡	Brake, acceleration, clutch, steering	V	GIST+GentleBoost
Mori *et al.* [112] (2012)	Awareness during lane-change	R,Gaze	Correlation Index
Liebner *et al.* [59] (2012)	Intersection turns and stop	GPS	Bayesian Network (BN)
Berndt and Dietmayer [60] (2009)	Lane change and turns	E,L,GPS,Map	HMM
This study‡	Overtake, Brake	E,He,Ha,L,R,F,V	Latent-Dynamic Conditional Random Field (LDCRF) and Multiple Kernel Learning (MKL)

*Input types: E=Ego-Vehicle Parameters, He=Head, Ha=Hand, L=Lane, R=Radar/Lidar Objects, F=Foot, V=Visual cues not included in previous types, such as break lights and pre-attentive cues.
†: Defined lane-change at lane crossing. ‡: Explicitly models pre-intent cues.

to other approaching vehicles. For instance the blinker may be turned on automatically.

Our goal is defined as follows: The prediction and early detection of overtaking and braking intent and maneuvers using driver, vehicle, and surround information.

In the vehicle domain, a few hundred milliseconds could signify an abnormal or dangerous event. To that end, we aim to model every piece of information suggesting an upcoming maneuver. In order detect head motion patterns associated with visual scanning [338–340] under settings of occlusion and large head motion, a two camera system for head tracking is employed. Subtle preparatory motion is studied using two additional cameras monitoring hand and foot motion. In addition to head, hand, and foot gesture analysis, sensors measuring vehicle parameters and surrounding vehicles are employed (Fig. 5.8). A gray-scale camera is placed in order to observe lane markings and road geometry, and a 360° color camera on top of the vehicle allows for panoramic analysis. Because visual challenges that are encountered in different surveillance domains, such as large illumination changes and occlusion, are common in our data, the action analysis modules studied in this work are generalizable to other domains of application as well.

We first perform a review of related literature in Section 5.6, while making a case for holistic understanding of multi-sensory fusion for the purpose of driver understanding and prediction. Event definition and testbed setup will be discussed in Sections 5.12 and 5.8, respectively. The different signals and feature extraction modules are detailed in Section 5.9. Two temporal modeling approaches for maneuver representation and fusion will be discussed in Section 5.10, and the experimental evaluation (Section 5.12) demonstrates analysis of different cues and modeling techniques in terms of their predictive power.

5.6 Related Research Studies

In our specific application, prediction involves recognition of distinct temporal cues not found in the large, 'normal' driving class. Related research may fall into three categories, which are roughly aligned with different temporal segments of the maneuver: trajectory estimation, inference, and intent

prediction,with the first being the most common. In trajectory estimation, the driver is usually not observed, but IMU, GPS [341] and maps [342], vehicle dynamics [59], and surround sensors [343] play a role. These attempt to predict the trajectory of the vehicle given some observed evidence (i.e. the beginning of significant lateral motion) and the probability of crossing the lane marking [344, 56]. A thorough recent review can be found in [302].

In intent inference approaches, the human is brought in as an additional cue, which may allow for earlier prediction. For instance, Doshi *et al.* [105] uses head pose, among other cues, in order to predict the probability that the vehicle will cross the lane marking in a two second window before the actual event. Several recent simulator studies have been performed using a variety of cues for intent inference. In [345], driver intent to perform overtaking was investigated using gaze information and an Artificial Neural Network (ANN). Vehicle dynamics, head, gaze, and upper body tracking cues were used in [346] with a rule-based approach for the analysis of driver intent to perform a variety of maneuvers. Even EEG cues may be used, as was done in [347] for emergency brake application prediction. Table 5.3 lists related research based on the maneuver studied, the learning approach, and the cues used for comparison with this work. Table 5.3 lists related studies done in naturalistic driving settings, as in our experiments. These present additional challenges to vision-based approaches.

Intent prediction corresponds to the earliest temporal prediction, and is rare in literature. Generally, existing studies do not look back in the prediction beyond 2-3 seconds before the event (e.g. the lane marker crossing for lane change maneuver). Intent prediction implies scene representation that may attempt to imitate human perception of the scene in order to produce a prediction for an intended maneuver. For instance, in [67] pre-attentive visual cues from a front camera are learned for maneuver prediction. An example would be a brake light appearing in front of the ego-vehicle, causing the driver to brake.

In our objective to preform early prediction, we study a wide array of cues as shown in Table 5.3. In particular, we attempt to characterize maneuvers completely from beginning to end using both driver-based cues and surround-based cues. We point out that a main contribution comes from analysis of a large number of modalities combined, while other studies usually focused on a subset of the signals in this work (Table 5.3). Furthermore, the detection and tracking modules are all kept in real-time. Training and testing of models for intention prediction, inference, and trajectory estimation will be done. Furthermore, we study additional cues (hand, foot, visual pre-attentive cues) which were little studied in previous work. Studying driver, surround, and vehicle cues allows for gaining insight into how these are related throughout a maneuver (Fig. 5.9).

5.7 Event Definition

Commonly, a lane change event or an overtake event (which includes a lane-change) are defined to begin at the lane marker crossing. On the contrary, in this work the beginning of an overtake event is defined earlier when the lateral motion started. We note that there are additional ways to define a maneuver such as an overtake or a lane-change (see [340]), and that our definition is significantly earlier

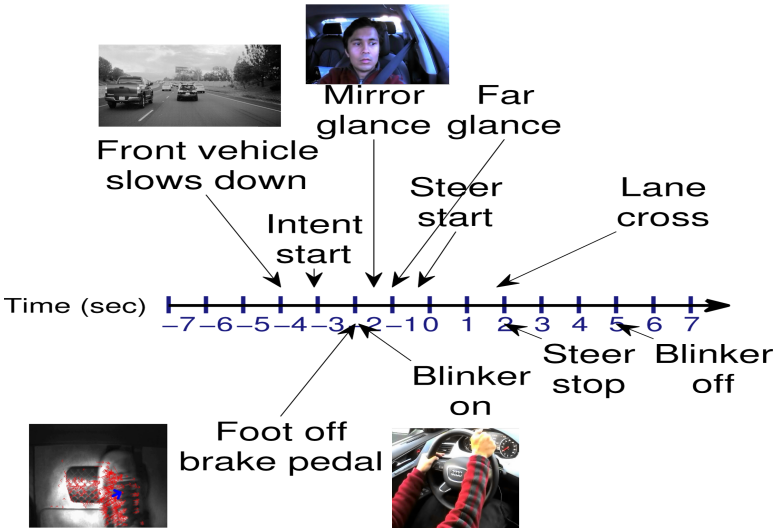


Figure 5.9: Timeline of an example overtake maneuver. Our algorithm analyzes cues for intent prediction, intent inference, and trajectory estimation towards the end of the maneuver.

than those in several related studies in Table 5.3. For instance, techniques focusing on trajectory-based prediction define lane-change at the lane marker crossing.

Nonetheless, as shown in (Fig. 5.9), the driver had the intent to change lanes much earlier, even before any lane deviation occurred. We wish to study how well can we observe such intent. By annotating events at the beginning of the lateral motion following the steering cue, the task of prediction becomes significantly more challenging. Under such a definition, lane deviation and vehicle dynamics are weak cues for prediction, while human-centered cues play a bigger role. Some examples are cues for visual scanning, as well as preparatory movements with foot and hands.

In addition to studying overtake maneuvers, which involve lateral control of the vehicle, we study a longitudinal control maneuver which is also essential in preventing accidents and monitoring for driver assistance. These are events where the driver chose to brake due to a situational need. While brakes are more easily defined (by pedal engagement), they allow us to evaluate the ability of the framework to generalize to other maneuvers. Any brake event (both harsh and weak) is kept in the data. This is done in order to emphasize analysis of key elements in the scene which cause drivers to brake.

5.8 Instrumented Mobile Testbed and Dataset

A uniquely instrumented testbed vehicle is used in order to holistically capture the dynamics of the scene: the vehicle dynamics, a panoramic view of the surround, and the driver. Built on a 2011 Audi A8, the automotive testbed is outfitted with extensive auxiliary sensing for the research and development of advanced driver assistance technologies. Fig. 5.8 shows a visualization of the sensor array, consisting of vision, radar, lidar, and vehicle (CAN) data. The goal of the testbed buildup is to provide a near-panoramic

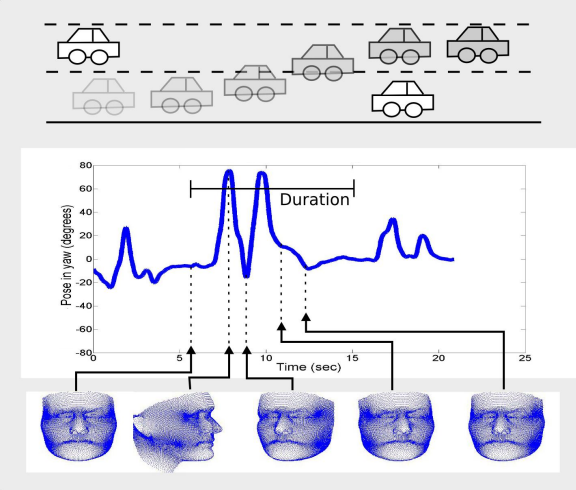


Figure 5.10: An example overtake maneuver. Head cues are important for capturing visual scanning and observing intent. The output of the head pose tracker as the maneuver evolves are shown using a 3D model.

sensing field of view for experimental data capture. Currently, the experimental testbed features robust computation in the form of a dedicated PC for development, which taps all available data from the on-board vehicle systems, excluding some of the camera systems which are synchronized using UDP/TCP protocols. Sensor data from the radars and lidars are fused into a single object list, with object tracking and re-identification handled by a sensor fusion module developed by Audi. On our dataset, the sensors are synchronized up to 22ms (on average). The sensor list is as follows:

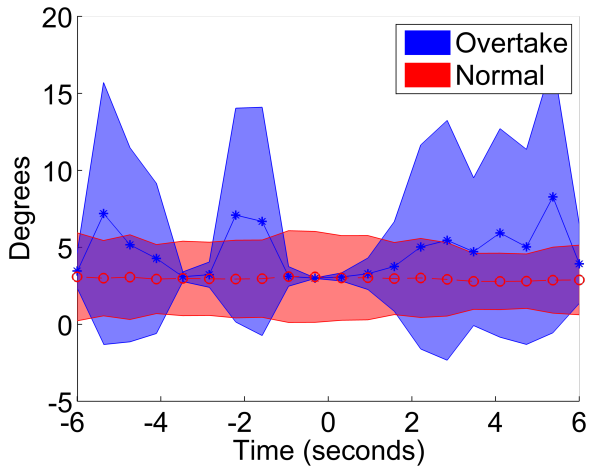
Looking into the vehicle:

- Two cameras for head pose tracking.
- One camera for hand detection and tracking.
- One camera for foot motion analysis.

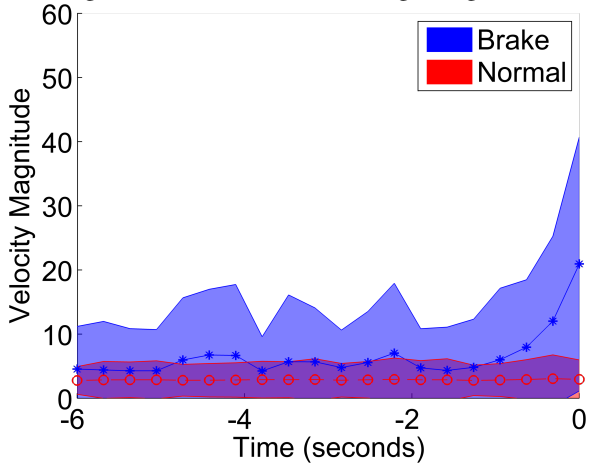
Looking outside of the vehicle:

- Forward looking camera for lane tracking.
- Two lidar sensors, one forward and one facing backwards.
- Two radar sensors on either side of the vehicle.
- A Ladybug2 360° video camera (composed of an array of 6 individual rectilinear cameras) on top of the vehicle.

The sensors are integrated into the vehicle body or placed in non-distracting regions to ensure minimal distraction while driving. Finally, information is captured from the CAN bus providing 13 mea-



(a) Head yaw in degrees during an overtake event (t=0 at beginning of lateral motion of the vehicle).



(b) Velocity magnitude of the foot during a braking event.

Figure 5.11: Mean and standard deviation of signals from the head pose and foot motion tracking modules during the two maneuvers studied in this work.

surements of the vehicle's dynamic state and controls, such as steering angle, throttle and brake, and vehicle's yaw rate.

With this testbed, a dataset composed of three continuous videos with three different subjects for a total of about 110 minutes (over 165,000 video frames at 25 frames per second were used) was collected. Each driver was requested to drive as they would in naturalistic settings to a set of pre-determined set of destinations. Training and testing is done using a 3-fold cross validation over the different subjects, with two of the subjects used for training and the rest for testing. Overall, we randomly chose 3000 events of 'normal' driving with no brake or overtake events, 30 overtaking instances, and 87 brake events. Braking events may be harsh or soft, as any initial engagement of the pedal is used.

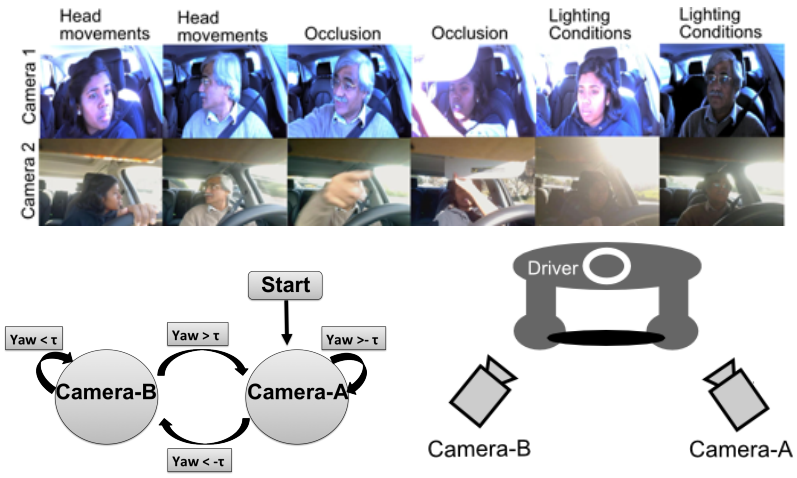


Figure 5.12: A two camera system overcomes challenges in head pose estimation and allow for continuous tracking even under large head movements, varying illumination conditions, and occlusion.

5.9 Maneuver Representation

In this section we detail the vision modules used in order to extract useful signals for analysis of activities.

5.9.1 Signals

Head: Head dynamics are an important cue in prediction. The head differs from the other body parts since the head is used by drivers for information retrieval from the environment. For instance, head motion may precede an overtaking maneuver in order to scan for other vehicles

Multiple cameras for human activity analysis [348] and face analysis [349] have been shown to reduce occlusion-related failures. In [330], a multi-perspective framework increased the operational range of monitoring head pose by mitigating failures under large head turns. In our setup, one camera is mounted on the front windshield near the A-pillar and another camera is mounted on the front windshield near the rear-view mirror to minimize intrusiveness.

First, head pose is estimated independently on each camera perspective using some of the least deformable facial landmarks (i.e. eye corners, nose tip), which are detected using supervised descent method [329], and their corresponding points on a 3D mean face model [323]. The system runs at 50Hz. It is important to note that head pose estimation from each camera perspective is with respective to the camera coordinates. One-time calibration is performed to transform head pose estimation from respective camera coordinates to a common coordinate where a yaw rotation angle equal to, less than and greater than 0° represent the driver looking forward, rightward and leftward, respectively.

Second, head pose is tracked over a wide operational range in the yaw rotation angle using both camera perspectives as shown in Fig. 5.12. In order to handle camera selection and hand-off, multiple

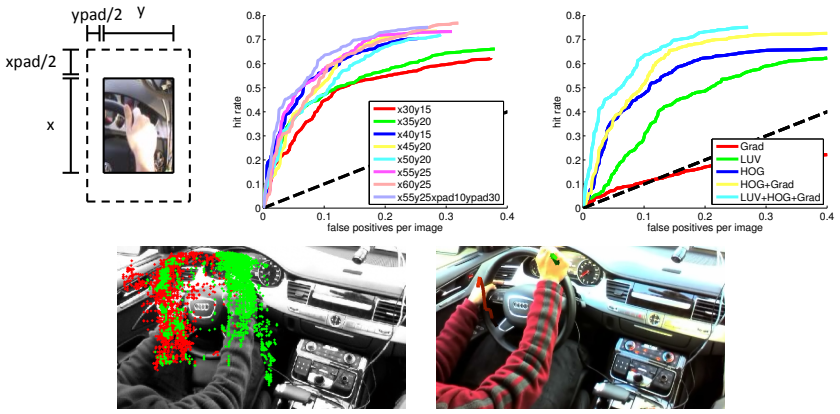


Figure 5.13: Analysis of the hand localization module. Top: Hand detection results with varying patch size and features; MAG gradient magnitude, HOG gradient orientation, and LUV color channels. Bottom: Scatter plot of left (in red) and right (in green) hand detection for the entire drive. A hand trajectory of reaching towards the signal before an overtake is shown (brighter is later in time).

techniques have been proposed in literature (a survey of different methods can be found at [334]). We had success with using the yaw as the camera hand-off cue. Assuming, without loss of generality, that at time $t = 0$ camera A is used to estimate head pose, then the switch to using camera B happens from when yaw rotation angle is greater than τ. Similarly the switch from B to A happens when yaw rotation angle is less than $-\tau$. If there is little to no spatial overlap in camera selection (i.e. $\tau = 0$), then noisy head pose measurements at the threshold will result in switching between the two camera perspectives needlessly. To avoid unnecessary switching between cameras, a sufficiently overlapping region is employed.

Hand: The hand signal will be used to study preparatory motions before a maneuver is performed. Below, we specify the hand detection and tracking module. Hand detection is a difficult problem in computer vision, due to the hand's tendency to occlude itself, deform, and rotate, producing a large variability in its appearance [127]. We use aggregate channel features [215] which are fast to extract. Specifically, for each patch extracted from a color image, gradient channels (six gradient orientation channels and normalized gradient magnitude) and color channels (CIE-LUV color channels were experimentally validated to work best compared to RGB and HSV) were extracted. 2438 instances of hands were annotated, and an AdaBoost classifier with decision trees as the weak classifiers is used for learning [350, 23]. The hand detector runs at 30 fps on a CPU. We noticed many of the false detections occurring in the proximity of the actual hand (the arm, or multiple detections around the hand), hence we used a non-maximal suppression with a 0.2 threshold. Because of this, window size and padding had a significant effect on the results (Fig. 5.13). In order to differentiate the left from the right hand, we train a histogram of oriented gradients (HOG) with a support vector machine (SVM) detector. A Kalman filter is used for tracking.

Foot: One camera is used to observe the driver's foot behavior near the brake and throttle pedal, and an illuminator is also used due to lack of lighting in the pedal region. While embedded pedal sensors

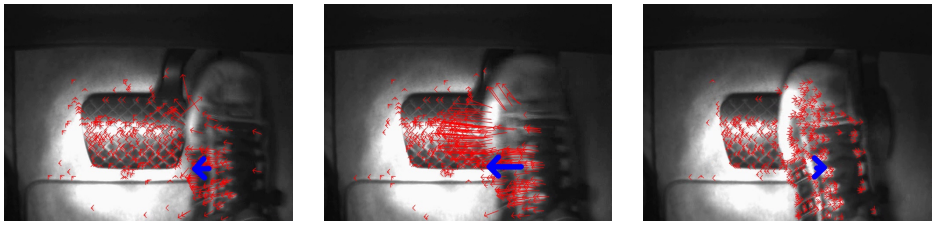


Figure 5.14: Foot tracking using iterative pyramidal Lucas-Kanade optical flow. Majority vote produces location and velocity.

already exist to indicate when the driver is engaging any of the pedals, vision-based foot behavior analysis has additional benefits of providing foot movements before and after pedal press. Such analysis can be used to predict a pedal press before it is registered by the pedal sensors.

An optical flow (iterative pyramidal Lucas-Kanade [351], running at 30Hz) based motion cue is employed to determine the location and magnitude of the foot and its velocity (Fig. 5.14). Optical flow is sufficiently robust for analyzing foot behavior due to little illumination changes and the lack of other moving objects in the region. First, optical flow vectors are computed over sparse interest points, which are detected using Harris corner detection. Second, a majority vote over the computed flow vectors reveals an approximate location and magnitude of the global flow vector.

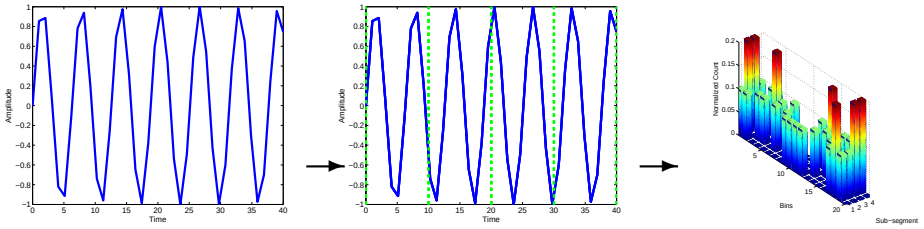
Optical flow based motion cues have been used in literature for analyzing head [352] and foot [2] gestures. Tran *et al.* [2] showed promising results where 74% of the pedal presses were correctly predicted $133ms$ before the actual pedal press.

Lidar/Radar: The maneuvers we study correlate with surrounding events. For instance, a driver may brake because of a forward vehicle slowing down or choose to overtake a vehicle in its proximity. Such cues are studied using an array of range sensors that track vehicles in term of their position and relative velocity. The sensor-fusion module, developed by Audi, tracks and re-identifies vehicles across the lidar and radar systems in a consistent global frame of reference. In this work we only consider trajectory information (longitudinal and lateral position and velocity) of the forward vehicle.

Lane: A front-observing gray-scale camera (see Fig. 5.8) is used for lane marker detection and tracking using a built-in system. The system can detect up to four lane boundaries. This includes the ego-vehicle's lanes and two adjacent lanes to those. The signals we consider are the vehicle's lateral deviation (position within the lane) and lane curvature.

Vehicle: The dynamic state of the vehicle is measured using a CAN bus, which supplies 13 parameters such as blinker state and vehicle's yaw rate. In understanding and predicting the maneuvers in this work, we only steering wheel angle information (important for analysis of overtake events), vehicle velocity, and brake and throttle pedal information.

Surround Visual: The 360° panoramic camera outputs the composed view of six cameras. The view is used for annotation, offline analysis, as well as extracting color and visual information from the scene. The front vehicle, detected by the lidar sensor, is projected to the panorama image using an offline



Raw Signal | Split into k sub-segments ($k = 4$ in this case) | $k \times (\#bins)$ histogram descriptor

Figure 5.15: Two features used in this work: raw trajectory features outputted by the detection and tracking, and histograms of sub-segments.

calibration. The projected vehicle box is padded, and a 50-bin histogram of the LUV channels is used as a descriptor for each frame. We also experimented with other scene descriptors, such as the GIST descriptor as done in [67]. GIST was shown to benefit cues that were not surround-observing (such as vehicle dynamics), yet the overall contribution after fusion of all of the sensors was not significant and so a detailed study of such features is left for future work.

5.9.2 Temporal Features

We compare two temporal features for each of the signals outputted by any one of the sensors described above at each time, f_t. First, we simply use the signal in a time window of size L,

$$F_t = (f_{t-L+1}, \ldots, f_t) \tag{5.5}$$

The time window in our experiments is fixed at three seconds. These will be referred to as 'raw' features, as they simply involve a concatenation of the time series in the window.

A second set of features studied involves quantization of the signal into bins (states) in order to produce histograms (depicted in Fig. 5.15). The temporal feature is a normalized count of the states that occurred in the windowed signal. In this scheme, temporal information is preserved by a split of the signal into k equal sub-signals and histogram each of these sub-signals separately. We experimented with different choices for k, and found $k = 1, 2, 4$ to work well with no advantage in increasing the number of sub-segments further. This was used in all of the experiments. The number of bins was kept fixed at 20.

5.10 Temporal Modeling

A model for the signals extracted by the modules in Section 5.9 must address several challenges. First, signal structure must be captured efficiently in order to produce a good modeling of maneuvers. Second, the role of different modalities should be studied with an appropriate fusion technique. Two types of modeling schemes are studied in this work, one using a Conditional Random Field (CRF) [353] and the other using Multiple Kernel Learning (MKL) [354]. The limitations and advantageous of these

two schemes will be discussed, with the overarching goal of understanding the evolution and role of different signals in maneuver representation.

Given a sequence of observations from Eq. 5.5, $\mathbf{x} = \{F_t^{(1)}, \ldots, F_t^{(s)}\}$, where s is the total number of signals, the goal is to learn a mapping to a label space, $\mathcal{Y}$, of different maneuver labels. This can be done using a conditional random field.

Conditional Random Field: Temporal dynamics are often modeled using a graphical model which reasons over the temporal structure of the signal. This can be done by learning a generative model, such as a Markov Model (MM) [146], or a discriminative model such as a Conditional Random Field (CRF) [353]. Generally, CRF has been shown to significantly outperform its generative counterpart, the MM. Furthermore, CRF can be modified to better model latent temporal structures, which is essential for our purposes.

The Hidden CRF (HCRF) [355] introduces hidden states that are coupled with the observations for better modeling of parts in the temporal structure of a signal with a particular label. A similar mechanism is employed by the Latent-Dynamic CRF (LDCRF) [353], with the advantage of also providing a segmentation solution for a continuous data stream. Defining a latent conditional model and assuming that each class label has a disjoint set of associated hidden states **h** gives

$$P(\mathbf{y}|\mathbf{x};\Lambda) = \sum_{\mathbf{h}} P(\mathbf{y}|\mathbf{h},\mathbf{x},\Lambda)P(\mathbf{h}|\mathbf{x},\Lambda) = \sum_{\mathbf{h}:\forall h_i \in H_{y_i}} P(\mathbf{h}|\mathbf{x};\Lambda) \tag{5.6}$$

where Λ is the set of model parameters and **y** is a label or a sequence of labels. In a CRF with a simple chain assumption, this joint distribution over **h** has an exponential form,

$$P(\mathbf{h}|\mathbf{x};\Lambda) = \frac{\exp\left(\sum_k \Lambda_k \cdot \mathbf{T}_k(\mathbf{h},\mathbf{x})\right)}{\sum_{\mathbf{h}} \exp\left(\sum_k \Lambda_k \cdot \mathbf{T}_k(\mathbf{h},\mathbf{x})\right)} \tag{5.7}$$

We follow [353], where the function $\mathbf{T}_k$ is defined as a sum of state (vertex) or binary transition (edge) feature functions,

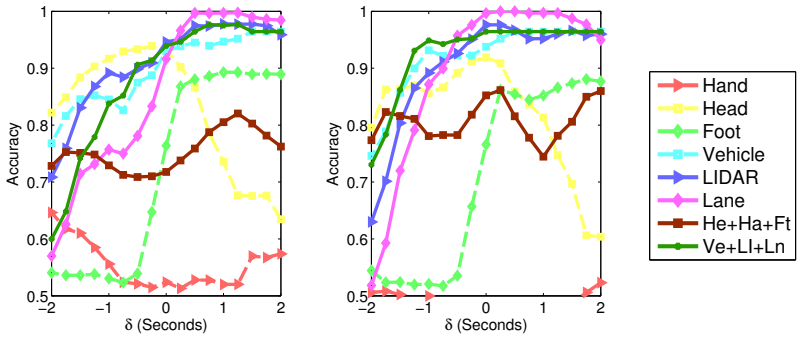
$$\mathbf{T}_k(\mathbf{h},\mathbf{x}) = \sum_{i=1}^{m} l_k(h_{i-1}, h_i, \mathbf{x}, i) \tag{5.8}$$

The model parameters are learned with gradient ascent over the training data using the objective function,

$$L(\Lambda) = \sum_{i}^{n} \log P(\mathbf{y}_i|\mathbf{x}_i,\Lambda) - \frac{1}{2\sigma^2}||\Lambda||^2 \tag{5.9}$$

where $P(\Lambda) \sim \exp(\frac{1}{2\sigma^2}||\Lambda||^2)$. In inference, the most probable sequence of labels is the one that maximizes the conditional model (Eqn. 5.6). Marginalization over the hidden states is computed using belief propagation.

With LDCRF, early-fusion is used for fusion of the temporal signal features. When considering



(a) LDCRF (b) MKL

Figure 5.16: Classification and prediction of overtake-late/brake (Experiment 1a) maneuvers using raw trajectory features. He+Ha+Ft stands for the driver observing cues head, hand, and foot. Ve+Li+La is vehicle (CAN), lidar, and lane. MKL is shown to handle integration of multiple cues better.

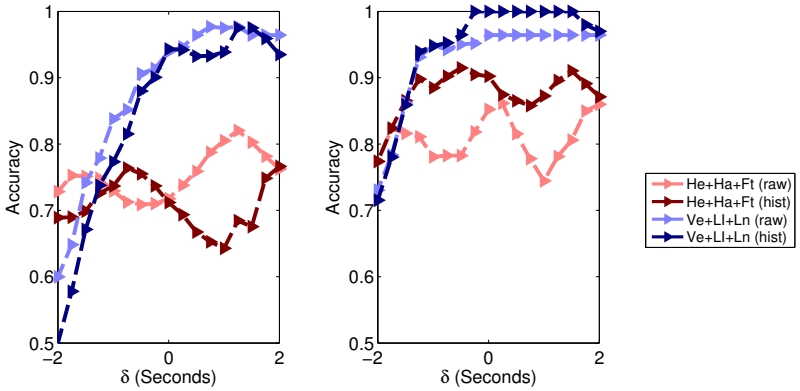
the histogram features studied in this work, each bin in the histogram is associated with an observation vector of size k (where k is illustrated in Fig. 5.15). In this case, temporal structure is measured by the evolution of each bin over time. Possibly due to the increase in dimensionality and the already explicit modeling of temporal structure in the LDCRF model, using raw features was shown to work as good or better than the sub-segment histogram features.

Multiple Kernel Learning: A second approach for constructing a maneuver model is motivated by the need for fusion of the large number of incoming signals from a variety of modalities. Given a set of training instances and signal channel c_l (i.e. brake pedal output), a kernel function is calculated for the signal, $\kappa_{c_l}(\mathbf{x}_i, \mathbf{x}_j) : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ (d is the feature dimension and $\mathbf{x}_i, \mathbf{x}_j$ are two data points). This produces a set of s kernel matrices for the n data points in the training set, $\{\mathbf{K}^{c_l} \in \mathbb{R}^n \times \mathbb{R}^n, l = 1, \ldots, s\}$, so that $K_{ij}^{c_l} = \kappa_{c_l}(\mathbf{x}_i, \mathbf{x}_j)$. s stands for the total number of outputs provided by the modules in Section 5.9. In our implementation, Radial Basis Function (RBF) kernels are derived for each of the signals using $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp(-||\mathbf{x}_i - \mathbf{x}_j||/\gamma)$. The cost and spread parameters are found for each signal separately using grid search.

The kernels are combined by learning a probability distribution $\mathbf{p} = (p^1, \ldots, p^s)$, with $p \in \mathbb{R}_+$ and $\mathbf{p}^T\mathbf{1} = 1$, such that the combination of kernel matrices,

$$\mathbf{K}(\mathbf{p}) = \sum_{l=1}^{s} p^l \mathbf{K}^{c_l} \tag{5.10}$$

is optimal. In this work, the weights are learned using stochastic approximation [354]. LIBSVM [245] is used as the final classifier. The histogram features were shown to work well with MKL, performing better than simply using the raw temporal signal features [239].



(a) LDCRF (b) MKL

Figure 5.17: Comparison of the two temporal features (see Section 5.9.2) studied in this work, raw temporal features and sub-segments histogram features, using overtake-late/brake (Experiment 1a) maneuvers. MKL benefits from the histogram features, especially in fusion of multiple types of modalities.

5.11 Experimental Setup

Several experiments are conducted in order to test the proposed framework for recognition of intent and prediction of maneuvers. As mentioned in Section 5.7, we experiment with two definitions for the beginning of an overtake event. An overtake event may be marked when the vehicle crossed the lane marking or when the lateral movement began. These are referred to as **overtake-late** and **overtake-early**, respectively. Normal driving is defined as events when the brake pedal was not engaged and no significant lane deviation occurred, but the driver was simply keeping within the lanes. A brake event is any event in which the brake pedal became engaged. Furthermore, we do not require a minimum speed for the events, so normal, brake, and overtake events may occur at any speed. Brake events may be in any magnitude of pedal press.

Initially, the proposed framework is evaluated by studying the question of whether a driver is about to overtake of brake due to a leading vehicle, as both are possible maneuvers. These experiments provide analysis on the temporal features and modeling. Once these initial experiments are complete, this allows us to move further to more complicated scenarios. Below, we detail the reference system to each experiment that will be performed in the experimental evaluation (Section 5.12).

- **Experiment 1a**: Overtake-late events vs. brake events (overtake-late/brake).
- **Experiment 1b**: Overtake-early events vs. brake events (overtake-early/brake).

Next, we are concerned with how each of the above events is characterized compared to normal driving.

- **Experiment 2a**: Overtake-late events vs. normal driving events (overtake-late/normal).
- **Experiment 2b**: Overtake-early events vs. normal driving events (overtake-early/normal).

Finally, we study the framework under a different maneuver,

- **Experiment 3**: Brake events vs. normal driving (brake/normal).

5.12 Experimental Evaluation

Temporal modeling: The first set of evaluations is concerned with comparison among the choices for the temporal features and temporal modeling. Each cue is first modeled independently in order to study its predictive power. The results for LDCRF and MKL under experiment 1a, ovetake-late/brake are shown in Fig. 5.16 for raw trajectory features. LDCRF demonstrates better predictive power using each modality independently when compared to MKL. For instance, lane information provides better prediction at $\delta = -2$ (2 seconds before the event start definition) with the LDCRF model. Similar conclusion holds for the head pose signal as well. As LDCRF explicitly reasons over temporal structure in the signal, these results are somewhat expected.

Temporal features and fusion: Fig. 5.16 also shows the results of fusion of multiple modalities with one model learned over the multiple types of signals. For clarity, we only show fusion of driver-based cues (head, hand, and foot) and surround cues (vehicle parameters, lidar, and lane). MKL is shown to perform better, as it is designed for fusion of multiple sources of signals. On the other hand, with the increase in dimensionality, the LDCRF model is shown to be limited. This is further studied in Fig. 5.17, where the MKL scheme demonstrates further gains due to the temporal structure encoded by the histogram descriptor. This is not the case for LDCRF, as it already explicitly reasons over temporal structure in the data. Therefore, for the rest of the section, LDCRF is joined with raw temporal features and the MKL with the temporal histogram features. Next, the more challenging experiments of early prediction are performed. As specific events are studied against a large 'normal' events dataset which includes naturalistic variation in each cue, the prediction task becomes more challenging. Furthermore, prediction much earlier in the maneuver of overtake-early events is also challenging.

2gray!25white

The results are summarized in Fig. 5.18 for experiments 2 and 3, where the entire set of signals described in Section 5.9 is used. For each experiment, the predictive power of the learned model is measured by making a prediction of a maneuver earlier in time, at increments of one second. At $\delta = -2$, a prediction is made two seconds before the actual event definition. Fig. 5.18(b) demonstrates the challenging task of prediction of overtake-early events, which mostly involve recognition of scanning and preparatory movement together with the surround cues. In this scenario of intent inference, lane deviation or steering angle info (which are strong cues for prediction in overtake-late events) are less informative. On the other hand, prediction of two seconds before an overtake-late maneuver is well defined in the feature space. Generally, the MKL is shown better results due to better fusion of the multiple signal sources, yet the prediction trends are consistent with the two temporal modeling schemes.

Insights into the maneuvers: Next, we consider the trade-off and value in sensor addition to an

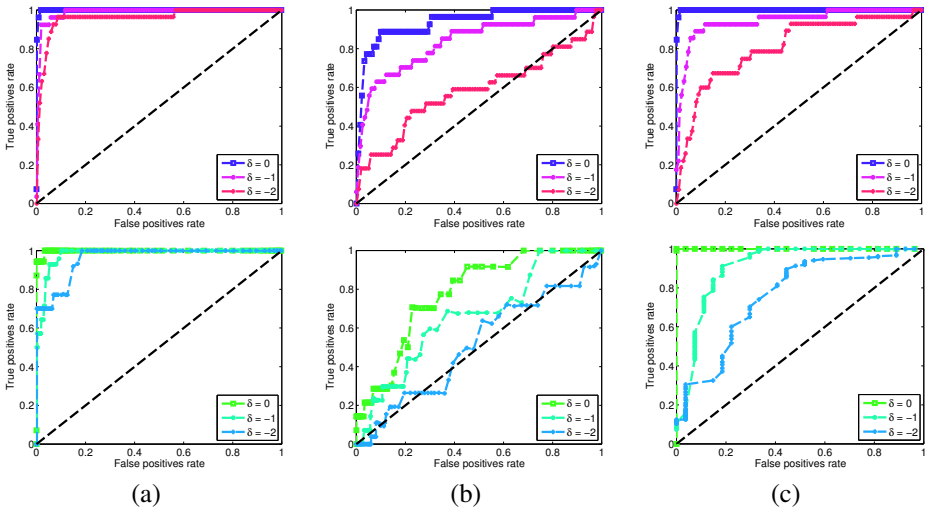


Figure 5.18: Measuring prediction by varying the time in seconds before an event, δ. **Top**: MKL results. **Bottom**: LDCRF results. (a) Experiment 2a: Overtake-late vs. normal (b) Experiment 2b: Overtake-early vs. normal (c) Experiment 3: Brake vs. normal. Note how prediction of overtake-early events, which occur seconds before the beginning of an overtake-late events, is more difficult.

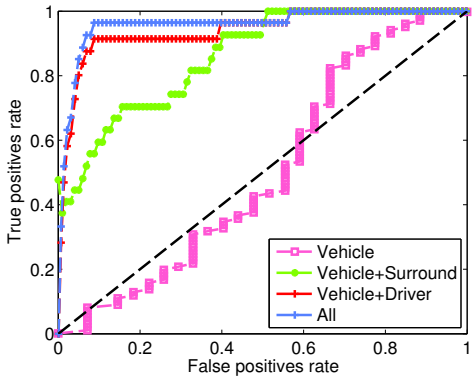
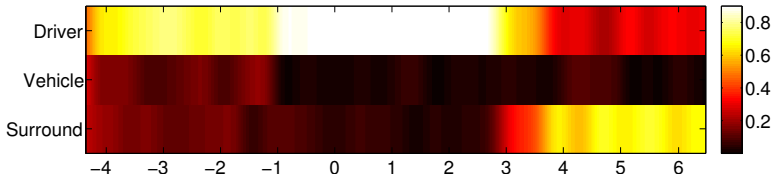


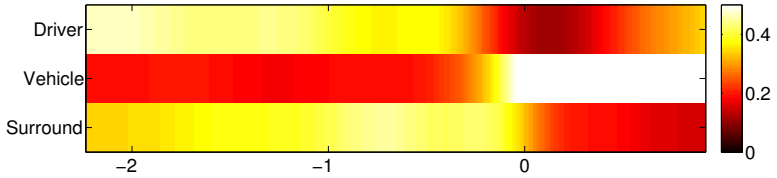
Figure 5.19: For a fixed prediction time of $\delta = -2$ seconds, we show the effects of appending cues to the vehicle dynamics under overtake-late/normal (experiment 2a). The surround cues utilize lidar, lane, and visual data. Driver cues include the hand, head, and foot signals.

existing vehicle system. Suppose that vehicle dynamics are provided, we quantify the benefit of adding a surround sensor capturing system for the prediction compared to a driver sensing system. The results are depicted in Fig. 5.19. Although both systems provide an advantage, most gains for early prediction come for prediction by observing driver related cues.

Fig. 5.20 shows the temporal evolution of cue importance using the weight output **p** from the MKL framework. Effective kernels will correspond to a heavier weight, and kernels with little discriminative value will be associated a smaller weight. Fig. 5.20 demonstrates how the entire maneuver can now be characterized in terms of the dynamics and evolution of different cue over the maneuver. For overtake events, driver-related cues of head, hand, and foot are strongest around the time that the lateral



(a) Overtake-early/normal (Experiment 2b)



(b) Brake/normal (Experiment 3)

Figure 5.20: Kernel weight associated with each cue learned from the dataset with MKL (each column sums up to one). Each manuever was learned against a set of normal events without the maneuver. Characterizing a maneuver requires cues from the human (hand, head, and foot), vehicle (CAN), and the environment (lidar, lane, visual-color changes). Time 0 for overtake is at the beginning of the lateral motion.

motion begins (t=0) in Fig. 5.20(a). Surround cues include lane, lidar, and visual surround cues. After the steering began, the lane deviation cue becomes a strong indicator for the activity. Similarly, the temporal evolution of the cues is shown for brake/normal event classification in Fig. 5.20(b). We see that driver cues (i.e. foot), and surround cues (i.e. visual cues, lidar) are best for early prediction, and a sharp increase in the kernel weight associated with vehicle dynamics occurs around the time of the pedal press.

5.13 Chapter Concluding Remarks

The chapter dealt with leveraging both a hand and head view in order to provide activity recognition for interactivity. Integration provided improved activity recognition results and allowed for a more complete semantic description of the human's activity state. A set of in-vehicle secondary tasks performed during on-road driving was utilized to demonstrate the benefit of the proposed approach, with promising results. Furthermore, the research task was extended to additional driver and scene cues, with the aim of holistic, spatio-temporal representation of activity. This is of particular importance to the application of automotive driver assistance systems, as these must perform under time-critical constraints, where even tens of milliseconds are essential. A holistic and comprehensive understanding of the driver's intentions can help in gaining crucial time and save lives. Prediction of human activities was studied using information fusion from an array of sensors in order to fully capture the development of complex temporal interdependencies in the scene. Evaluation was performed on a rich and diverse naturalistic driving dataset showing promising results for prediction of both overtaking and braking maneuvers. The framework allowed the study of the different types of signals over time in terms of predictive importance. In the future, additional maneuver types, such as those performed when approaching to and at intersections

will be studied.

This chapter is in part a reprint of material that is published in the International Conference on Pattern Recognition (2014), by Eshed Ohn-Bar, Sujitha Martin, Ashish Tawari, and Mohan M. Trivedi. The dissertation author was the primary investigator and author of this paper.

This chapter is in part a reprint of material that is published in the journal of Computer Vision and Image Understanding (2015), by Eshed Ohn-Bar, Ashish Tawari, Sujitha Martin, and Mohan M. Trivedi. The dissertation author was the primary investigator and author of this paper.

Chapter 6

Towards Human-Centric Scene Understanding in Video

This chapter provides a unified analysis of object recognition, behavior modeling, and human perception in a combined research task. Understanding intent and relevance of surrounding agents from video is an essential task for many applications in robotics and computer vision, suitable for studying spatio-temporal context modeling techniques. The modeling and evaluation of contextual, spatio-temporal situation awareness is particularly important in the domain of intelligent vehicles, where a robot is required to smoothly navigate in a complex environment while also interacting with humans. In this thesis, we address these issues by studying the task of on-road object importance ranking from video. First, human-centric object importance annotations are employed in order to analyze the relevance of a variety of multi-modal cues for the importance prediction task. A deep convolutional neural network model is used for capturing video-based contextual spatial and temporal cues of scene type, driving task, and object properties related to intent. Second, the proposed importance annotations are used for producing novel analysis of error types in image-based object detectors. Specifically, we demonstrate how cost-sensitive training, informed by the object importance annotations, results in improved detection performance on objects of higher importance. This insight is essential for an application where navigation mistakes are safety-critical, and the quality of automation and human-robot interaction is key.

6.1 Introduction

There is a great need for smarter and safer vehicles [356, 122]. Large resources in both industry and academia have been allocated for the development of vehicles with a higher level of autonomy and advancement of human-centric artificial intelligence (AI) for driver assistance. Understanding, modeling, and evaluation of situational awareness tasks, in particular the understanding of the behavior and intent of

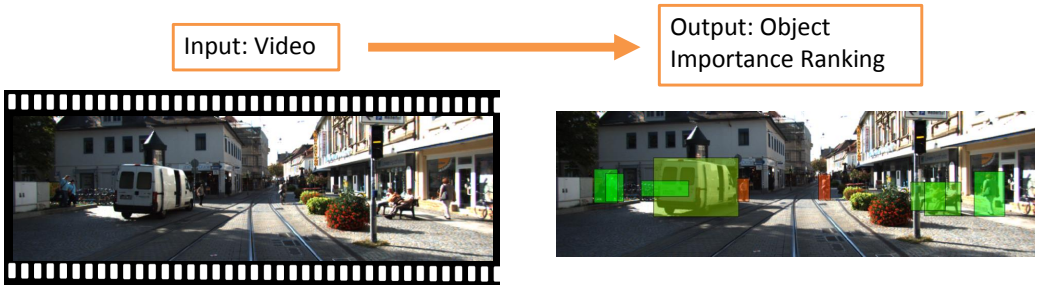


Figure 6.1: What makes an object salient in the spatio-temporal context of driving? Given a video, this work aims to rank agents in the surrounding scene by relevance to the driving task. Furthermore, the notion of importance defined in this work allows a novel evaluation of vision algorithms and their error types. The importance score (averaged over subjects' annotations) for each object are shown, colored from red (high) to amber (moderate) to green (low).

agents surrounding a vehicle, is an essential component in the development of such systems [357–360]. Human drivers continuously depend on situation awareness when making decisions. In particular, the observation that attention given by human drivers to surrounding road occupants varies based on a task-related, scene-specific, and object-level cues motivates our study of human-centric object recognition.

A model of driver perception of the scene requires reasoning over spatio-temporal saliency, agent intent, potential risk, as well as past and possible future events. For instance, consider the on-road scene in Fig. 6.1. Obstacle avoidance requires robust recognition of all obstacles in the scene, yet surrounding obstacles are not all equal in terms of relevance to the driving task and attention required by a driver. Given the specific scene in Fig. 6.1, a subset of the road occupants (remote, occluded, or low-relevance objects) was consistently annotated at a lower importance level by human annotators when considering the driving task. On the other hand, a pedestrian intending to cross and a cyclist at the ego-lane were consistently annotated at higher importance levels for the driving task. The input to the modeling/annotation task is a video, and the output is a per-frame, object-level importance score. This level of contextual reasoning is essential for an intelligent robot required to navigate in the world, as well as communicate with and understand humans. This work is concerned with training recognition algorithms that can perform such complex reasoning. In order to better understand the aforementioned observations and issues, we propose to study a notion of on-road object importance, as measured in a spatio-temporal context of driving a vehicle. The contributions of our study are as follows.

6.1.1 Contributions

Modeling object importance: The main contribution of this work is in the study of which cues are useful for on-road object importance ranking. Specifically, a set of spatio-temporal object attributes are proposed for capturing attention, agent intent, and scene context. The analysis is performed in the context of autonomous driving on KITTI videos [195], but may also be useful to other application domains in computer vision requiring spatio-temporal analysis and human perception modeling, including saliency

modeling [361, 362], robotics [363], and ego-centric vision [364].

Importance-guided performance metrics: The collected dataset is used to produce new evaluation insights for vision tasks. In particular, the annotations are used to highlight dataset bias in object detection for autonomous driving. As highly important objects are rare, we experimentally demonstrate existing training and testing procedures to be biased towards certain object characteristics, thereby hindering insights from comparative analysis. Furthermore, the object importance annotations are used to train cost-sensitive, attention-aware object detection models. The proposed importance-guided training procedure is shown to result in models which produce less errors when objects of higher importance are concerned - a useful insight for the safety-critical application considered in this study.

6.2 Motivation and Related Research Studies

Importance analysis: Importance ranking essentially involves modeling context. Capturing spatial image context has been heavily studied [365, 366]. Berg *et al.* [367] measure object-level importance in an image by the likelihood of the object to be mentioned by a person describing it. Temporal context implies movement modeling [368], understanding of what an agent can do, intends to do, or how multiple agents may interact [369]. Lee *et al.* [370] studies object importance regression in long-term ego-centric videos using gaze, hand-object interaction, and occurrence frequency cues, but no human importance annotations are employed. Mathialagan *et al.* [371] performs single image importance prediction of people with linear regression over pose, occlusion, and distance features. On the other hand, we pursue spatio-temporal importance ranking as it relates to a perceived driving environment by a driver. The task of on-road object importance modeling may also be somewhat correlated with general visual saliency [361, 372], but the latter is often not studied for a driving task.

Human-centric evaluation: It is known that driver experience level (usually measured in years) significantly impacts safe driving partly due to improved identification and prediction of other road occupants' intentions [122]. As computer vision datasets become more realistic and complex, one way to evaluate such prior knowledge and complex modeling of spatio-temporal events (involving object recognition, scene context modeling, etc.) is using the proposed set of importance metrics (similar metrics have been devised for other machine learning and vision tasks, such as object segmentation and image captioning [373, 374]). Human-centric metrics provide a rich tool for understanding the human in the loop, from modeling human drivers in general to a specific driver perception and style, and is of great use to development effective driver assistance and human-computer cooperation. Conveying intents by autonomous driving vehicles to other road occupants is also an important task relevant to our study, as it may require understanding of how humans perceive a scene.

Importance metrics for on-road object detection: We employ the importance annotations in order to perform a finer-grained evaluation of object detection. At a high level, two object detectors may potentially have similar detection performance while differing in ability to detect important objects. A dataset bias could further hinder such an insight. Algorithms for visual recognition of objects has seen

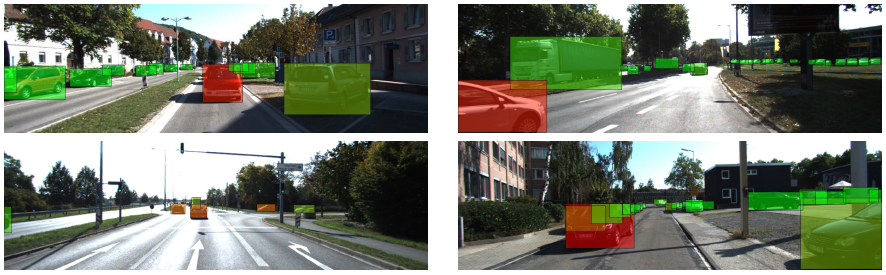


Figure 6.2: This study is motivated by the fact that not all objects are equally relevant to the driving task. As shown in example frames from the dataset with overlaid object-level importance score (averaged over subjects), drivers' attention to road occupants varies based on task-related, scene-specific, and object-level cues.

tremendous progress in recent years, most notably on the ILSVRC [375–377], PASCAL [222, 161], Caltech [215], and KITTI datasets [141], yet low cost, camera-based object detection with low false positives over many hours of video in a wide variety of possible environmental conditions is still not solved. Therefore, better understanding and evaluation of the limitations of state-of-the-art object detection algorithms is essential. We believe current metrics employed for generic object detection are limited for the study of on-road object detection as detailed below. We emphasize that this study is not concerned with ethical issues in autonomous driving, but instead with deeper understanding of requirements and limitations for safe navigation and human-centric AI on an object detection and classification level.

Are all objects equal? It may not be surprising that the **answer is no**, even in existing evaluation protocols for object detection. Some objects posing certain visual challenges are notoriously more difficult to detect than others. Objects of small size, heavy occlusion, or large truncation are partially or entirely excluded from existing evaluation (and training) on PASCAL, Caltech, and KITTI. Yet in the context of driving, such instances may be the most relevant under safety-critical events! Existing evaluation metrics are often inconsistent regarding these visual challenges, and reflect a certain bias [378–380, 183] where importance is measured differently from in the driving domain. We experimentally demonstrate the impact of such bias in evaluation on KITTI (Section 6.5). Furthermore, importance-based metrics normalize evaluation curves differently than ones based on object appearance properties (properties which may be distributed differently across datasets), and so it has the potential of offering complementary insights. For instance, consider scenarios of dense scenes with tens of road occupants that are heavily occluded or are across a barrier (e.g. highway settings). As annotation of such scenes is challenging and evaluation of objects across a barrier may not be necessary for development and evaluation of algorithmic recognition performance, the importance-centric framework only consider a handful of agents which are of higher importance. As large numbers of objects in KITTI (Fig. 6.2) were generally annotated at low relevance to the driving task, the proposed annotations could be used to provide deeper understanding of existing object detectors in a domain where errors are costly and the type of errors made should well understood. It will be shown in Section 6.5 that training detectors without a notion of importance can have a biasing effect on the output of the detector itself. Our approach is also biologically plausible, as human

Figure 6.3: The interface used to obtain object-level importance ranking annotations. The cyclist is highlighted as it is the currently queried object to annotate, colored boxes have already been annotated with an importance level by the annotator, and blue boxes are to be annotated.

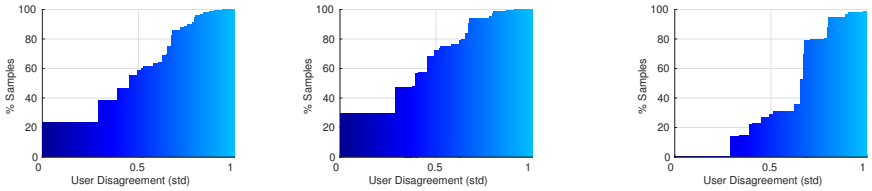
drivers do not generally pay attention to all objects in the scene (Fig. 6.2), but are skillful at recognition and analysis of only a subset of relevant objects. On the other hand, vision algorithms are evaluated on a large portion of low importance vehicle samples, which may skew analysis and insights.

6.3 Importance Annotation Dataset

The KITTI dataset [141, 195] was chosen due to richness of object-level annotation and sensor data. As video data is essential for the notion of importance, we utilize a subset of the raw data recordings with the provided 3D annotations of pedestrians, cyclists, and vehicles. The annotations include bird's eye view orientation and tracklet IDs. The dataset contains synchronized GPS, LIDAR, and vehicle dynamics, useful for studying the dynamics of a variety of cues as they relate to perceived object importance.

Importance annotations: Experiments were done in a driving simulator with KITTI videos shown on a large screen using the interface in Fig. 6.3. Subjects watched each short video twice, and every 10^{th} frame was annotated by querying for an integer between 1-3 (1 being high and 3 being low importance). Subjects were asked to imagine driving under similar situations, and mark objects by the level of attention and relevance they would've given the object under real driving. Three levels were chosen for simplifying the annotation process - two levels of importance (yes or no) is too restrictive as there is no way of handling ambiguous cases. On the other hand, a continuous ranking score may have been used, but such a task may lead to a large confusion among subjects and for guessing, which we aimed to reduce.

Although subjective in nature, the task of importance ranking is performed by all drivers every day. Out of a total of 18 subject, high correlation between subject driving experience, age, and annotation output was demonstrated. Interestingly, the annotation task resulted in a clear relationship between annotation output and subject driving experience (measured in years). Subject analysis can be found in the Fig. 6.5. Consistency analysis (Fig. 6.4) of the annotators output demonstrates that many instances in the



(a) Entire Dataset (b) Low Importance Objects (c) High Importance Objects

Figure 6.4: A cumulative histogram obtained by varying the disagreement requirement (standard deviation among subject labels), until 100% of the data is included. While disagreement exists, a subset of highly important and highly non-important objects shows consistency (see Sec. 6.3 for discussion).

low importance class have high agreement among the subjects. On the other hand, the moderate and high importance classes contain higher variation.

The overall dataset used in the experiments contains 17,635 object annotations, including 15,057 vehicles (cars, vans, and trucks), 1,452 pedestrians, and 562 cyclists. In the existing metrics on KITTI for object detection, test samples are categorized into three levels of difficulty based on object properties of height, occlusion, and truncation. 'Easy' test settings include non-occluded samples with height above 40 pixels and truncation under 15%, 'moderate' settings include partially-occluded samples with height above 25 pixels and truncation under 30%, and 'hard' settings include heavy occlusion samples with height above 25 pixels and truncation under 50%. In the same spirit, we introduce three importance classes by taking the median vote among subjects for each object instance, from high, moderate, to low importance. Out of the totals, there were high/moderate/low importance 293/2159/12,605 vehicles, 143/524/785 pedestrians, and 267/147/148 cyclists. Subjects reported a variety of reasons for importance annotations, from the existence of a barrier in traffic, head orientation cues for pedestrians (also studied in [381–383]), and spatio-temporal relationships between different objects. The annotations and code will be made publicly available. In addition to the three importance class, regression of the average importance score will also be studied.

Dataset properties: The dataset statistics are depicted in Fig. 6.6. When analyzing highly important objects, these are shown to be non-occluded samples within 40 meters or less of the ego-vehicle. Most vehicles are categorized as moderate or low importance, which is to be expected as KITTI contains many parked and stationary vehicles. Truncation percentage statistics binned to a histogram are approximately evenly distributed. Fig. 6.7(a-c) demonstrates that objects in the proximity of the vehicle may have any level of importance annotation, suggesting other cues besides position alone are necessary for the importance ranking task. In the image plane, Fig. 6.7(e) demonstrates the distribution of the position in the image plane for high importance objects.

6.4 Object Importance Model

In this section, we formulate the object importance models which will provide insights into what causes some objects to be perceived as more important than others. To that end, we propose two types of

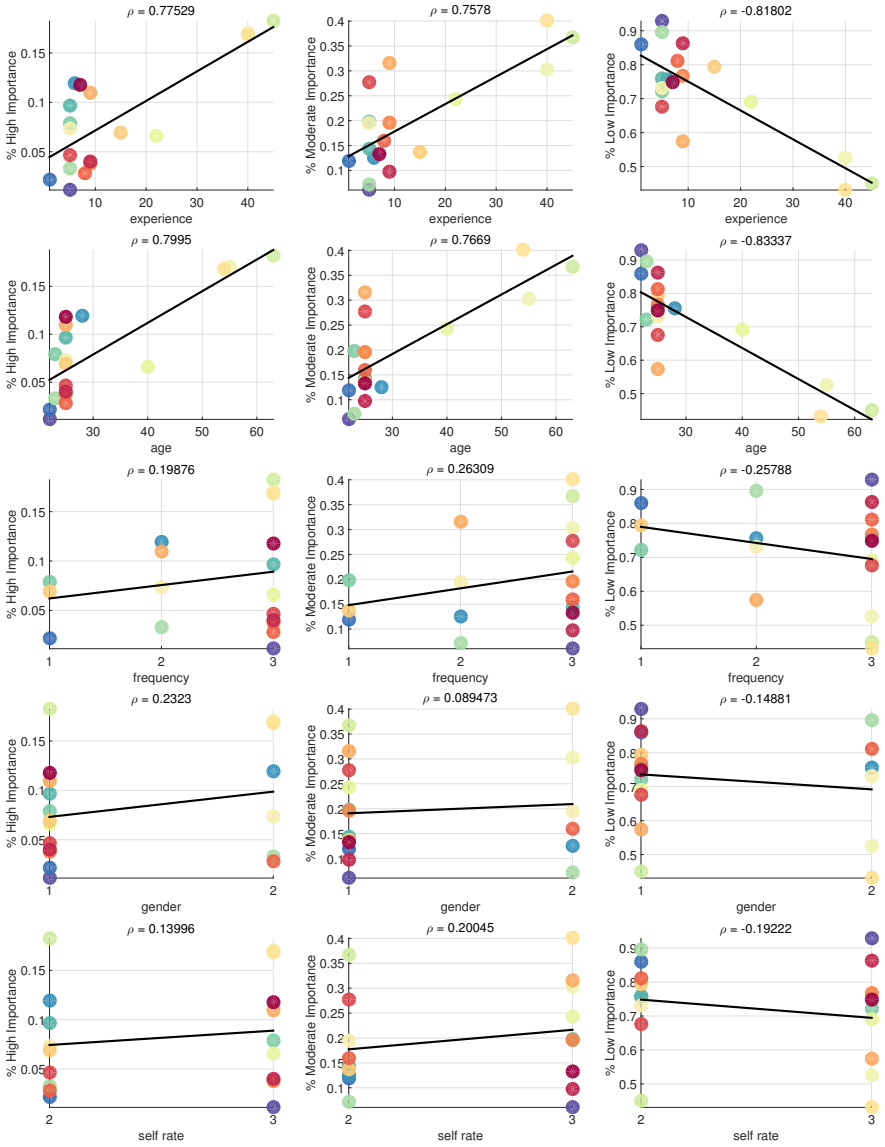


Figure 6.5: Relationship between importance level (grouped by columns) and subject personal information (grouped by rows). Each subject has been assigned a unique color, and is represented in each figure by a dot. From top row: (1) driving experience in years, (2) age in years, (3) frequency of driving, either 1-rarely, less than once a month, 2-occasionally, about once a week, 3-frequently, more than three times a week, (4) gender 1-male, 2-female, (5) rating of driving skill, 2-intermediate, 3-advanced. We observed a strong relationship between experience in years and importance ranking annotations.

models, differing by the type of features employed for scoring an object instance importance level. All model weights are learned using a logistic regression model.

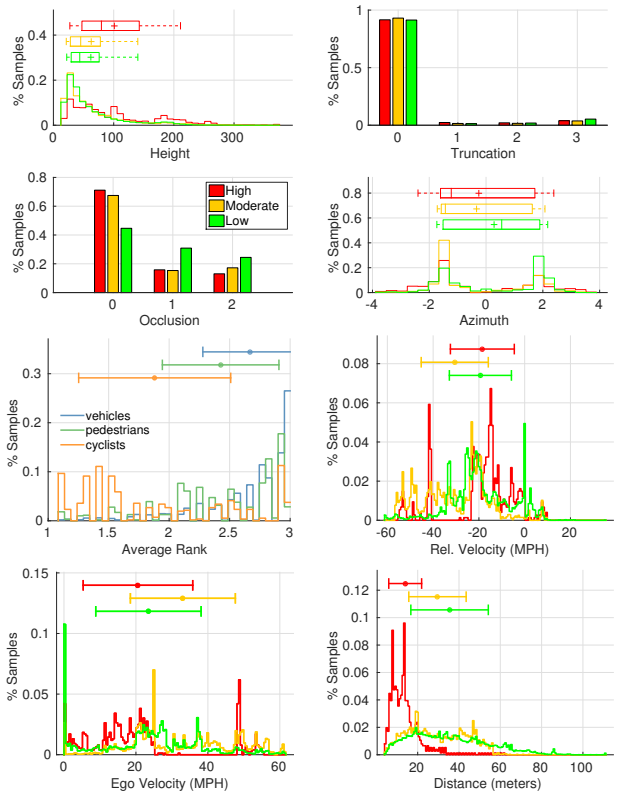


Figure 6.6: Object statistics corresponding to three classes of object importance in the dataset.

6.4.1 Object attributes model, $\mathbf{M}_{attributes}$

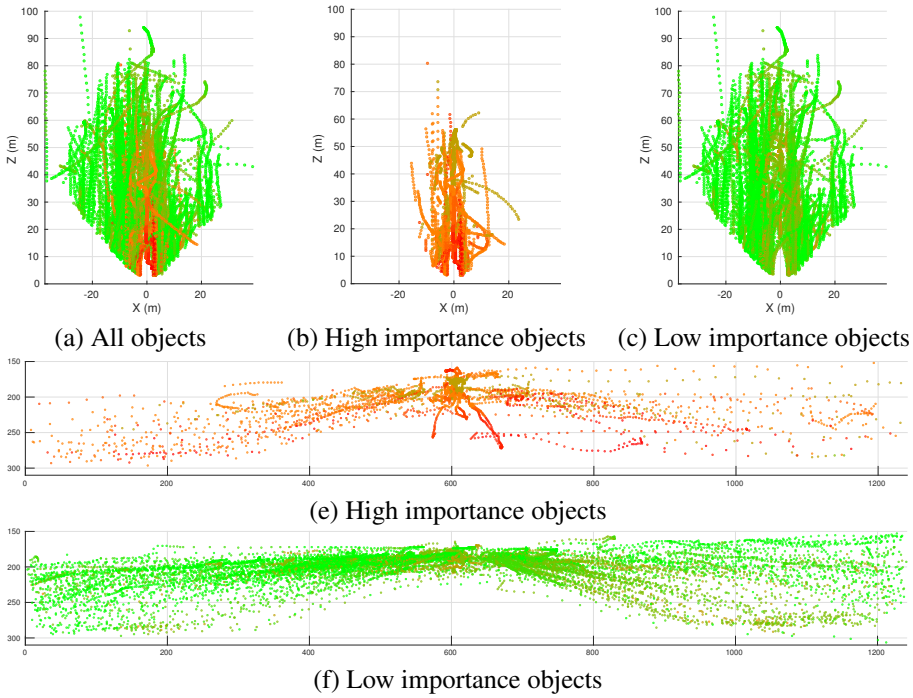
KITTI provides several high quality object-level attributes extracted from ground truth information and multi-modal sensor data. The attributes allow for an explicit analysis of the relationship between different object properties and importance ranking. For an instance s and class importance c, we train the following prediction model,

$$M_{attributes}(s) = \boldsymbol{w}_{c,2D-obj}^{T}\phi_{2D-obj}(s) + \boldsymbol{w}_{c,3D-obj}^{T}\phi_{3D-obj}(s) + \boldsymbol{w}_{c,ego}^{T}\phi_{ego}(s) + \boldsymbol{w}_{c,temporal}^{T}\phi_{temporal}(s) \tag{6.1}$$

where the features used in the $M_{attributes}$ model are defined below.

2D object features: For each sample, the 3D object box annotation is projected to the image plane for obtaining a set of 2D object properties. The $\phi_{2D-obj} \in \mathbb{R}^4$ features are the concatenation of the height in pixels, aspect ratio, occlusion state (either none, partial, and heavy occlusion) and truncation percentage.

3D object features: As shown in Fig. 6.7, distance from the ego-vehicle is correlated with annotated importance levels. Other 3D object properties, such as orientation, may provide hints as to what an on-road occupant is doing or intends to do. The $\phi_{3D-obj} \in \mathbb{R}^6$ features are composed of the



(a) All objects (b) High importance objects (c) Low importance objects

(e) High importance objects

(f) Low importance objects

Figure 6.7: Dataset distribution of object positions in top-down view (a)-(c) and image plane (d)-(e). Each instance is colored according to average importance ranking, from red (high) to amber (moderate) to green (low) importance.

left-right (lateral) and forward-backward (longitudinal) range coordinates (x, z) given by the LIDAR, Euclidean distance from the ego-vehicle, orientation in bird's eye view, and object velocity components, $|V|$ and $\angle V$.

Ego-vehicle features: Ego-vehicle parameters can be used in order to capture contextual settings relevant to the importance ranking task. For instance, if the ego-vehicle is traveling at low speeds, the surrounding radius in which objects may be considered relevant decreases. For that reason, ego-vehicle speed information is displayed during the annotation process as shown in Fig. 6.3. Hence, the attribute model includes ego-vehicle velocity magnitude and orientation features, $\phi_{ego} = [\text{ego}|V|,\ \text{ego}\angle V]$.

Temporal attributes: The total aforementioned 2D object, 3D object, and ego-vehicle features can be used to represent an object and certain contextual information in a given frame. Nonetheless, the temporal evolution of such properties may also provide useful information in representing past, present, and potential future actions, and consequently impact importance ranking. This assumption is captured in $\phi_{temporal}$, which is computed using the aforementioned object and ego-vehicle attributes but over a past time window. Specifically, $\phi_{temporal}$ is obtained by concatenating the attributes over the time window. In addition, we add the values after a max-pooling operation over the time window, as well as the Discrete Cosine Transform (DCT) coefficients [368].

We note that $M_{attributes}$, while utilizing the extensive KITTI multi-modal data and annotations, is not intended to be exhaustive. Additional attributes can potentially be considered, such as

object-object relationships attributes, object-lane relationship attributes, as well as scene-type attributes (although these are not currently provided with KITTI and will need to be extracted/annotated). The objective of $M_{attributes}$ is in gaining explicit insight into the role of object attributes which are known to contain little noise on importance ranking. Furthermore, $M_{attributes}$ is of use when comparing to a visual, video-only importance prediction model, which will be presented next. For instance, limitations in the visual prediction model will be analyzed using $M_{attributes}$. On the other hand, the visual model can implicitly encode attributes missing from $M_{attributes}$, such as spatial relationships among objects, scene types, and more.

6.4.2 Visual prediction model, $\mathbf{M}_{visual}$

Our main task is the visual prediction of object importance. Given a 2D bounding box annotation, M_{visual} learns a mapping from an image region to an importance class using

$$M_{visual}(s) = \boldsymbol{w}_{c,obj}^T \phi_{obj}(s) + \boldsymbol{w}_{c,spatial}^T \phi_{spatial}(s) + \boldsymbol{w}_{c,temporal}^T \phi_{temporal}(s) \tag{6.2}$$

where the feature components of the visual prediction model are defined next.

Object visual features: For $\phi_{obj} \in \mathbb{R}^{4096}$ features, we employ the activations of the last fully connected layer of the OxfordNet (VGG-16) [384] convolutional network. The network was pre-trained on the ImageNet dataset [376] and fine-tuned on KITTI using Caffe [385].

Spatial context features: In order to capture spatial context, such as relationship with other objects in the scene, lane information, scene type information, or better capture object properties (e.g. occlusion, truncation, orientation), each object instance is padded by a factor of $\times 1.75$ for generating $\phi_{spatial} \in \mathbb{R}^{4096}$.

Temporal context features: Similarly to in $M_{attributes}$, we hypothesize the human annotators reason over spatio-temporal cues in the videos shown to them when determining object relevance to a driving task. In order to test the hypothesis and provide insights into the importance ranking task, the per-frame visual descriptors, ϕ_{obj} and $\phi_{spatial}$, are employed for computing a $\phi_{temporal}$ component. Specifically, given an object tracklet with 2D box positions and a temporal window, the previous object and spatial context features are computed over a time window, concatenated, and max-pooled.

6.5 Importance Metrics for Object Detection

As described in Section 6.2, there are potential issues with applying traditional object detection metrics to on-road object detection analysis. In addition to the importance ranking task described in Sections 6.4.1 and 6.4.2, we provide further insights into the proposed importance dataset by studying the importance annotations in the context of object detection. Specifically, we study the usefulness of importance-based metrics in evaluating object detectors. For instance, as the majority of vehicles in KITTI were consistently ranked with lower importance to the immediate driving task, the rarity of objects

of higher importance may result in a bias both in training and evaluation. First, training may rather emphasize visual attributes found in the most common objects. Second, evaluation using traditional metrics may not reveal such a bias. In order to demonstrate this phenomenon and motivated by work on specializing convolutional networks (ConvNets) [386], we train object detectors which are specialized at detecting objects of higher importance.

The experiments employ the Faster R-CNN framework [8] with two training procedures, one importance-agnostic and one importance-guided. Following Fast R-CNN [194], the framework trains a network with two sibling output layers. The first output layer predicts a discrete probability distribution per each image region, $p = (p_0, \ldots, p_K)$ over $K + 1$ object categories, using a softmax over the $K + 1$ outputs of a fully connected layer. The second layer outputs bounding-box regression offsets for the 4 coordinates of the image region. For each training region labeled with a ground-truth class u and a ground-truth bounding-box regression target v, we use the following multi-task loss

$$L(p, u, \gamma, t^u, v) = L_{cls}^{IG}(p, u, \gamma) + \lambda_{loc}[u \geq 1]L_{loc}(t^u, v) \tag{6.3}$$

such that $L_{cls}^{IG}(p, u, \gamma) = -\alpha_\gamma \log p_u$ is the log loss for true class u. The weight factor α_γ is added, defined as

$$\alpha_\gamma = \begin{cases} \lambda & \gamma \leq 2.25 \\ 1/\lambda & \text{otherwise} \end{cases} \tag{6.4}$$

to allow cost-sensitive importance-guided training, where γ is the average importance score of the current sample. The cost-sensitive training allows steering the objective function optimization by increasing mis-classification penalty on objects with higher importance. The second task loss, L_{loc}, is the sum of the smooth L1 loss function over the 4 box coordinates as defined in [194]. L_{loc} is computed for samples of non-background class ($[u \geq 1]$) only. In the experiments, we set $\lambda_{loc} = 1$ and $\lambda = 10$. We note that setting $\alpha = 1$ for all γ results in the commonly used, importance-agnostic training procedure.

6.6 Experimental Evaluation

6.6.1 Importance Prediction Models

A total of 8 videos is employed in the experiments, with a 2-fold validation split. Results using the two importance models are shown in Table 6.1. In each experiment, classification is done for each importance class, a Precision-Recall (PR) curve is calculated, and the area under the curve (AP) is averaged (mAP) over the classes for an overall performance summary, so that higher mAP value implies better classification performance. For a second evaluation metric, we regress the average importance score for each object instance and compute the mean absolute error (MAE). Due to the large imbalance in the distribution of the importance scores, we show overall MAE on all samples as well as MAEγ which is computed

Table 6.1: Summary of the classification experiments using the two proposed importance prediction models.

Model	mAP (%)	MAE	$\text{MAE}_{\gamma=2.25}$
$M_{visual}(\phi_{obj})$	51.06	0.2648	0.5392
$M_{visual}(\phi_{obj} + \phi_{spatial})$	55.53	0.2611	0.5007
$M_{visual}(\phi_{obj} + \phi_{temporal})$	53.30	0.2507	0.4765
$M_{visual}(\phi_{obj} + \phi_{spatial} + \phi_{temporal})$	56.34	0.2447	0.4625
$M_{attributes}$ (without $\phi_{temporal}$)	53.70	0.2440	0.3853
$M_{attributes}$ (with $\phi_{temporal}$)	**60.35**	**0.2148**	**0.2914**

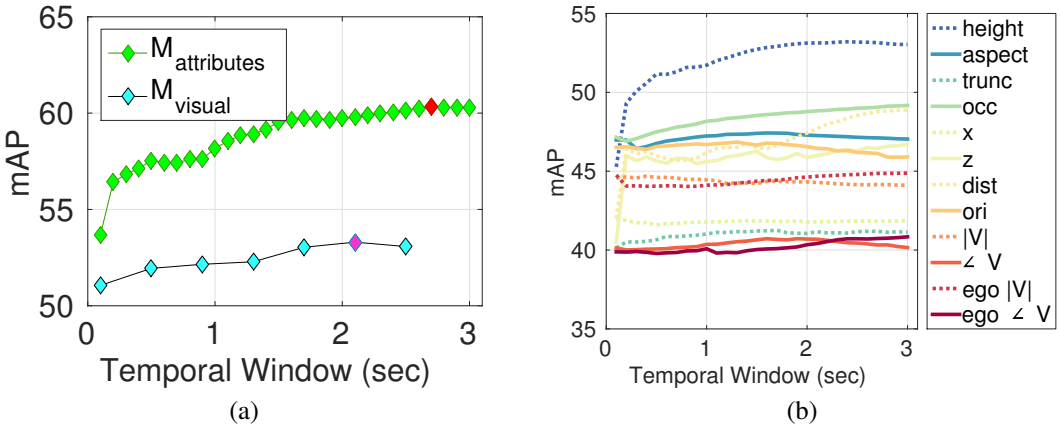


Figure 6.8: Cue analysis with the importance models. (a) Classification accuracy when varying the time window used for computing $\phi_{temporal}$ in both models. (b) Classification accuracy with each of the attributes in $M_{attributes}$ with an increasing temporal window used for a temporal feature extraction.

over a subset of samples with an average importance score less than or equal to γ. Setting $\gamma = 2.25$ allows for computing the MAE only on objects of higher importance, excluding objects considered of lower importance (with average importance score of more than 2.25).

Evaluation of $M_{attributes}$: Table 6.1 shows the performance of the attributes-based model. We note that for the experiments in Table 6.1, training and evaluation is done in an object class agnostic manner, only considering the importance class/score of samples. We note that due to the high-level features used in $M_{attributes}$, it should be considered as a strong baseline, achieving mAP of 53.70% and 60.35% without and with temporal features extraction, respectively. Temporal features are shown to be essential for both importance classification and regression of objects of higher importance. As shown in Fig. 6.8, a past time window of up to 2.7 seconds is shown to contain beneficial information for importance classification with $M_{attributes}$, while performance saturates for M_{visual} with a $\sim$2 seconds window.

Next, we further analyze the impact of different components in $M_{attributes}$ in order to better understand what makes an on-road object important. Fig. 6.9 depicts the relationship between individual attributes and importance prediction. Performance using the combination of all of the object attributes is shown as 'comb', which provides the best importance ranking results. Analysis is shown for each object

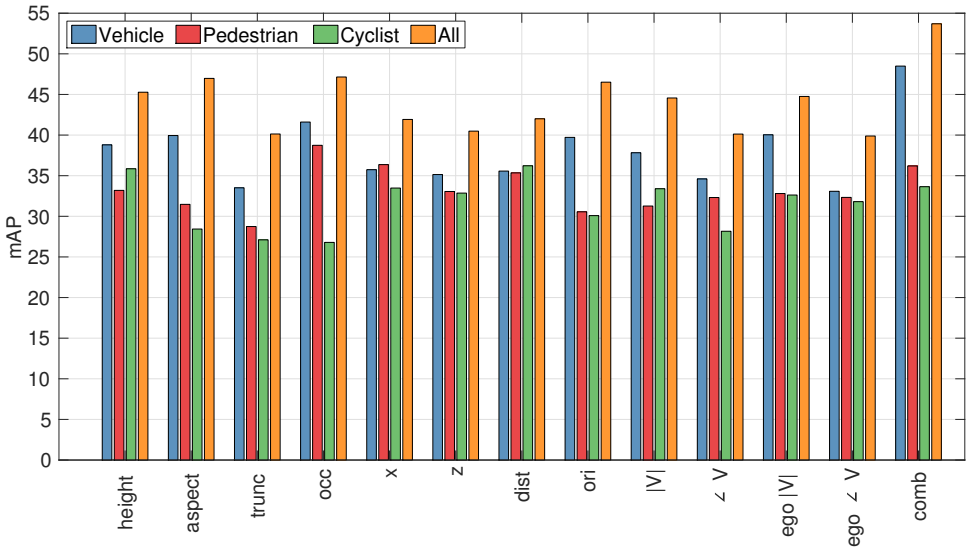


Figure 6.9: Object importance classification results using each attribute in $M_{attributes}$ separately, as well as with a combination of all attributes ('comb'). Results are shown for training and evaluation on each object class separately, as well as in an object class agnostic manner ('All'). No temporal feature extraction is used in these experiments.

category separately, as well as for training a single importance prediction model over all object types in an object class agnostic manner. Highest mAP for importance classification of vehicles is achieved using the object attributes of occlusion, aspect ratio, orientation, and height in the image plane. Because occlusion by another object often implies lower relevance to the driving task, occlusion state is shown to be a particularly useful cue. Similarly, orientation and aspect ratio may capture traffic flow direction and planned future actions. Ego-vehicle velocity magnitude is also shown to have high relationship with importance ranking, serving as a frame-level contextual cue. For the pedestrian object class, high impact attributes are also the distance and position in 3D. The cyclist object class follows similar trends, yet reliable conclusions are more difficult to draw as it contains a small number of samples.

Fig. 6.8(b) isolates the benefit that each individual attribute provides as the time window for the feature computation increases. Results are shown when considering an object class agnostic model. Fig. 6.8(b) highlights the importance of temporal feature extraction for several high-level semantic cues, including past occlusion and truncation, distance change from the ego-vehicle, lateral movement, and object size in the image plane. Certain attributes, such as ego-vehicle parameters, are shown to benefit from a larger past temporal window. This is to be expected, as ego-vehicle information serves as a general frame-level contextual cue. Fig. 6.10 shows the PR curves used to compute the final performance summary in Table 6.1. Fig. 6.10 demonstrates the significant impact of temporal attribute cues in classifying importance class for different object types, improving classification performance in almost every case. The smaller, cyclist object class contains large annotation inconsistencies, in particular within the moderate importance class, leading to poor performance for all of the importance prediction models. A larger

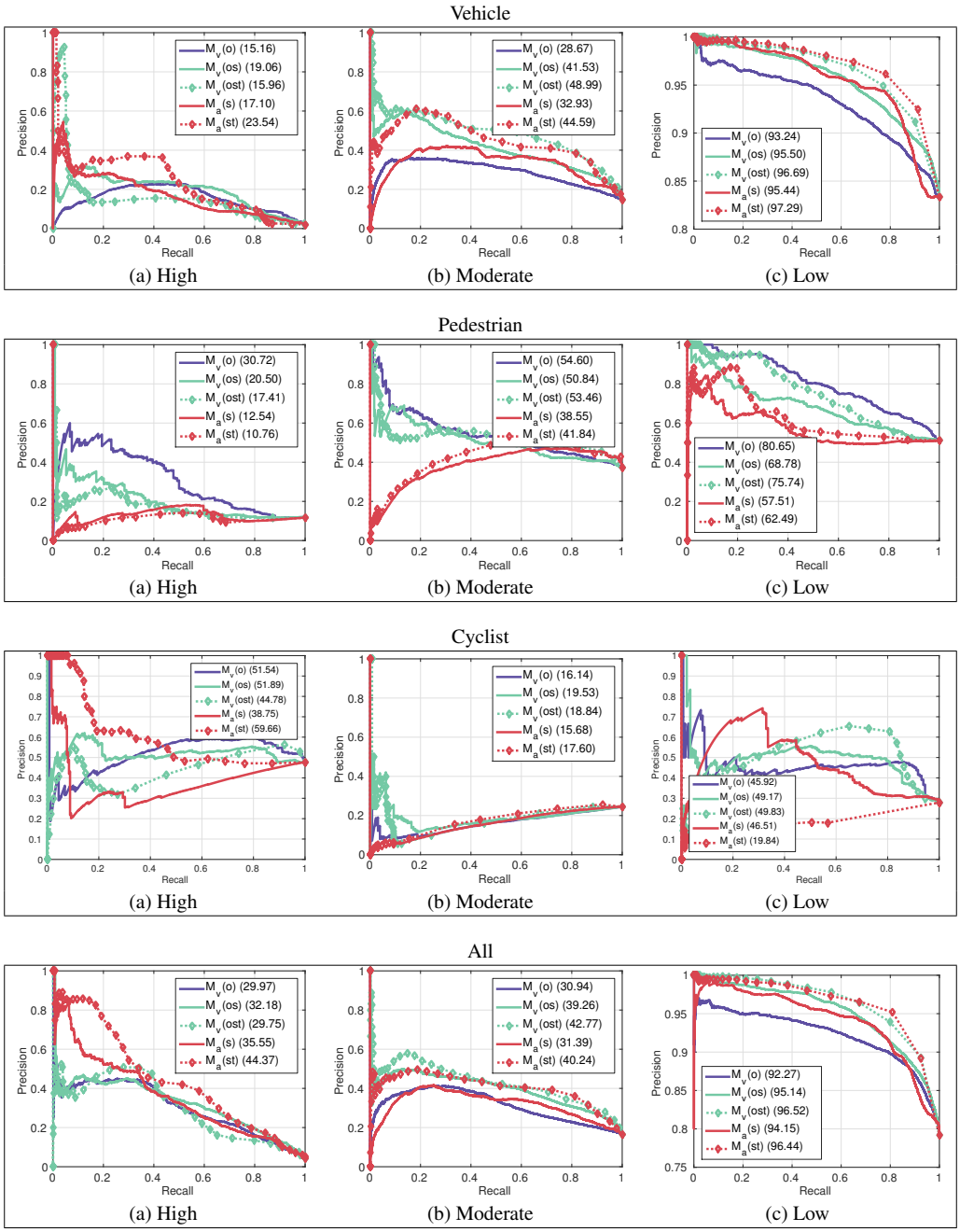


Figure 6.10: For each object class (rows) and object importance level (columns), we show performance precision-recall curves when employing different models and cue types. For the attributes model (M_a), performance without and with temporal features is shown as 's' and 'st', respectively. Similarly, for the visual model (M_v) performance with ϕ_{obj}, $\phi_{obj} + \phi_{spatial}$, and $\phi_{obj} + \phi_{spatial} + \phi_{temporal}$ is shown as 'o', 'os', and 'ost', respectively. In parenthesis is the area under the curve.

dataset could resolve such issues. Furthermore, additional insights may be gained by subject-specific modeling and evaluation, which is left for future work.

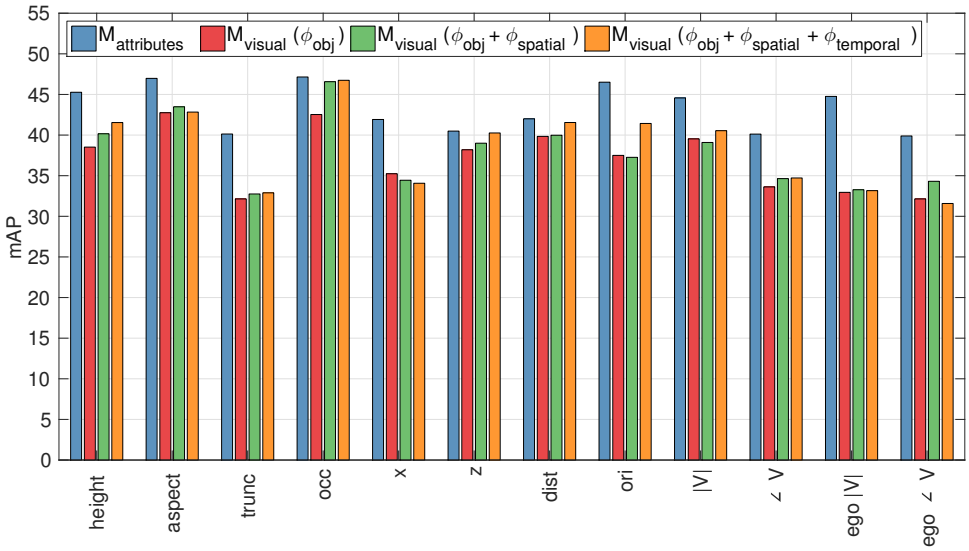


Figure 6.11: Regressing each attribute using various feature combinations in M_{visual} and consequently using the attribute for importance class classification allows for explicit analysis of the limitations of M_{visual}.

Evaluation of M_{visual}: Table 6.1 shows the performance summary of different components in the visual importance prediction model. Contrasting with $M_{attributes}$, simply using the object region features ϕ_{obj} results in a reduction of 2.64% mAP points to 51.06% mAP. This is expected, as $M_{attributes}$ employs clean annotation and other sensor data. The MAE in prediction average importance score also suffers, in particular on objects of higher importance. Addition of the spatial context component, $\phi_{spatial}$, results in a large performance improvement of 4.47% mAP points, as well as a noticeable reduction in MAEγ. The analysis demonstrates the importance of contextual information in modeling object importance. We've also experimented with schemes of feature extraction from the entire image for capturing scene information, but no additional benefit was shown.

As with $M_{attributes}$, incorporation of a temporal feature extraction component, $\phi_{temporal}$, to M_{visual} results in a further performance improvement, although to a lesser extent (56.34% mAP). As shown in Fig. 6.8, the improvement plateaus beyond a $\sim$2 seconds past window. When comparing performance among the two models, both in classification and regression, the M_{visual} model is significantly outperformed by $M_{attributes}$ (in particular on objects of higher importance). The results in Table 6.1 motivate further study of models suitable for capturing spatio-temporal visual cues [387–389, 24], which can be a future study.

Limitation analysis of M_{visual}: Comparing the visual-only ranking against the strong baseline $M_{attributes}$ of object attributes reveals insights as to the current limitations in representing object properties with the VGG network. This motivates an explicit limitation study, as shown in Fig. 6.11. In this experiment, the VGG network is used to regress each object attribute in $M_{attributes}$, and consequently the regressed value is used for importance ranking instead of the original value from $M_{attributes}$. The ex-

Table 6.2: Evaluation of object detection (AP) using the proposed set of importance metrics and the Faster-RCNN framework (FRCN) [8]. 'IG' refers to importance-guided fine-tuning, where correct classification of samples with higher importance annotations is weighted heavier in the training loss.

	Traditional Test Settings			**Importance Test Settings**		
Method	Easy	Mod.	Hard	High	High+Mod.	Low
FRCN-ZF	89.26	79.70	64.96	66.89	82.80	58.85
FRCN-ZF-IG	91.09	80.86	66.18	73.00	87.19	59.90
ΔAP	+1.83	+1.16	+1.22	+6.11	+4.39	+1.05
FRCN-VGG	95.63	88.98	74.65	81.73	91.60	69.54
FRCN-VGG-IG	94.54	88.71	74.01	85.13	91.67	69.09
ΔAP	-1.09	-0.27	-0.64	+3.40	+0.07	-0.45

(a) Vehicle, height 25 pixels and up

	Traditional Test Settings			**Importance Test Settings**		
Method	Easy	Mod.	Hard	High	High+Mod.	Low
FRCN-ZF	89.26	85.69	72.68	71.27	84.46	65.11
FRCN-ZF-IG	91.09	86.74	73.75	76.01	87.59	65.88
ΔAP	+1.83	+1.05	+1.07	+4.74	+3.13	+0.77
FRCN-VGG	95.63	92.74	80.90	85.56	92.29	74.53
FRCN-VGG-IG	94.54	91.70	79.56	86.73	91.44	73.40
ΔAP	-1.09	-1.04	-1.34	+1.17	-0.85	-1.13

(b) Vehicle, height 40 pixels and up

	Traditional Test Settings			**Importance Test Settings**		
Method	Easy	Mod.	Hard	High	High+Mod.	Low
FRCN-ZF	50.30	45.66	42.91	21.88	30.45	35.03
FRCN-ZF-IG	62.43	57.07	51.97	34.29	47.15	37.67
ΔAP	+12.13	+11.41	+9.06	+12.41	+16.70	+2.64
FRCN-VGG	66.71	61.23	57.96	22.48	44.91	48.67
FRCN-VGG-IG	70.67	64.81	59.47	33.01	53.76	43.53
ΔAP	+3.96	+3.58	+1.51	+10.53	+8.85	-5.14

(c) Pedestrian, height 25 pixels and up

	Traditional Test Settings			**Importance Test Settings**		
Method	Easy	Mod.	Hard	High	High+Mod.	Low
FRCN-ZF	50.30	47.59	44.75	22.57	32.13	36.45
FRCN-ZF-IG	62.43	58.12	52.98	34.61	48.13	38.39
ΔAP	+12.13	+10.53	+8.23	+12.04	+16.00	+1.94
FRCN-VGG	66.71	63.09	59.74	16.81	47.18	49.91
FRCN-VGG-IG	70.67	67.23	61.80	27.19	56.61	45.46
ΔAP	+3.96	+4.14	+2.06	+10.38	+9.43	-4.45

(d) Pedestrian, height 40 pixels and up

periment is repeated for different feature combinations in M_{visual}, providing insight into the benefit that different features provide and assist in explaining the current limitations in M_{visual}. Fig. 6.11 demonstrates that while some object attributes as they relate to object importance are predicted well (such as occlusion state), others (such as orientation, object velocity, or truncation) are lacking. The incorporation of the spatial and temporal context features significantly improves the ability to capture object state, in particular object occlusion state, range, and orientation. On the other hand, explicit regression of object velocity, ego-vehicle parameters, or truncation value is challenging.

6.6.2 Importance-Guided Object Detection

In the detection experiments, we follow the KITTI evaluation protocol of correct detection at 0.7 overlap for vehicles, and 0.5 for pedestrians and cyclists. All models are first fine-tuned for object detection on KITTI using the publicly available detection benchmark, but excluding frames from videos used in the importance experiments. Next, for each fold in the 2-fold cross validation, we fine-tune faster R-CNN (FRCN) [8] in an importance-agnostic manner and importance-guided manner, as described in Section 6.5. Results are shown in Table 6.2 for both the ZF [179] and the VGG [384] network architectures. Table 6.2 depicts the complementary relationship between the proposed set of importance metrics and traditional test settings (defined in Section 6.3). For instance, AP values differ among the easy/hard test settings when comparing to high/low importance test settings. In particular, as the low importance class isolates many instances with challenging settings of larger occlusion and smaller height, it exhibits the lowest performance across all metrics. Another observation is the impact of importance-guided training, in particular when performance is measured with importance-based metrics. For instance, importance-guided training with ZF results in a significant 6.11% AP improvement in detection of objects of the high importance class, while such an improvement is not visible in traditional metrics based on object height, occlusion, or truncation. This is due to a dataset bias, as most vehicles in the dataset are of lower importance ranking. A similar observation holds for results using VGG, but to a lesser extent as the larger and deeper VGG model is better at general object detection.

When analyzing results on KITTI, we observed a large number of false positives occurring for both the ZF and VGG models on objects of small height. In addition to the challenge in detecting small objects, we also observed inaccurate annotations in KITTI on small objects. Furthermore, the importance-guided training may be simply emphasizing large objects which are generally of higher importance. Therefore, Table 6.2 shows results on objects of 25 pixels and up (as proposed by KITTI), as well as on objects of 40 pixels and up. The latter corresponds to varying only occlusion/truncation in the 'moderate' and 'hard' traditional test settings. Comparing the two test settings on objects of 40 pixels and up, we can see that while importance-guided training indeed emphasizes correct detection on larger objects, the importance-based metrics are still able to capture complementary insights to the importance-agnostic metrics. For the pedestrian object class, there is a stronger correlation between the two types of metrics due to a higher proportion of high and moderate importance classes samples. Nonetheless, the general trends of improved performance due to importance-guided training still hold. Due to the small number of cyclists, only the vehicles and pedestrian categories are analyzed. The results demonstrate the feasibility of the proposed metrics both for the training and testing of vision tasks, in particular object detection. We note that as mentioned in [178], training task-specific ConvNets (e.g. for occlusion) does not necessarily result in improvement (and may even reduce overall detection performance). As shown in Table 6.2, this is not the case with importance classes.

6.7 Chapter Concluding Remarks

This chapter developed a human-centric framework for analyzing driving videos. Object recognition was analyzed under a notion of importance, as measured in a spatio-temporal context of driving a vehicle. Given a driving video, our main research aim was to model which of the surrounding vehicles are most important to the immediate driving task. Employing human-centric annotations allowed for gaining insights as to how drivers perceive different on-road objects. Although perception of surrounding agents is influenced by previous experience and driving style, we demonstrated a consistent human-centric framework for importance ranking. Extensive experiments showed a wide range of spatio-temporal cues to be essential when modeling object-level importance. Furthermore, the importance annotations proved useful when evaluating vision algorithms designed for on-road applications and autonomous driving. Future work includes studying the relationship between gaze dynamics, saliency, and object importance ranking. Furthermore, the dataset can be used in order to study subject-specific modeling which is relevant to cooperative driving and control transitions [388, 58, 390, 356]. Further investigation of the cost-sensitive training procedure [391, 392, 389] may lead to additional insights in the future. Appropriate temporal metrics, such as how quickly an object was classified as important in the video, can also be useful for comparing methods in importance prediction. Cross-dataset generalization and annotations on additional datasets [393, 394, 152] can provide further understanding into models and evaluations for importance prediction. Ideally, annotation of additional datasets can be done more efficiently by employing lessons learned from this work. Evaluation of the sensitivity of the importance models on different times of day, night, weather condition, and diverse traffic scenes are also important next steps. We hope that this study will motivate further developments in spatio-temporal object detection and importance modeling, essential for real-world video applications.

This chapter is in part a reprint of material that will be published in the journal of Pattern Recognition (2017), by Eshed Ohn-Bar, and Mohan M. Trivedi. The dissertation author was the primary investigator and author of this paper.

Chapter 7

Conclusions

With the goal of developing tools for human-robot interactivity, this dissertation made contributions to contextual object recognition and human behavior modeling. The chapters in the thesis were organized to follow their semantic modeling level, gradually and rigorously increasing the complexity of the discussed research task.

First, related research studies for looking at humans, particularly in the automotive domain, were surveyed. In particular, we find that there are many emerging research tasks for studying humans and behavior both in the vehicular environment, and for human-robot interaction environments. Although the different research tasks surveyed are often treated independently in literature, we drew upon the underlying connecting theme of studying human behavior in order to gain an overview of the vast research landscape.

The thesis touched upon different elements required in materializing the final goal of human-machine interactivity. We began with robust object detection and image-level contextual reasoning, necessary for performing additional in-depth visual analysis of behavior. Next, we turned to studying hand gestures, as well as coordination of human cues (hand, head, eye, and foot) with contextual scene and surround agents behavior. For interactivity, a robot must model human behavior in a scene (i.e. for a navigation task), as well as human perception of a scene (i.e. for an assistance task). This provides a rich research problem, with opportunities for developing efficient multi-modal fusion, studying temporal evolution of cues, discovering better modeling techniques, and gaining novel experimental insights. Throughout these diverse but connected research tasks, the theme of context (scene, parts, spatio-temporal relationships, etc.) was repeated as a crucial step towards useful human-robot interactivity systems.

Significant amount of work remains to be done for each of the aforementioned research tasks. Specifically for this thesis, we have presented a need for improved visual temporal modeling, approaches for better fusion, more in-depth analysis of style and behavior, and studying issues in human-robot cooperation. The opportunities to improve safety and quality of lives through contextual robotics systems which leverage behavior models are many. Specifically in the vehicular domain, through the use of advanced

sensor-based intelligence and interactivity, the next generation of transportation systems will ultimately strive for the goal of safer and accident-free roadways. Generally, the vehicular domain is one out of many domains of everyday life which will see a disruption due to highly intelligent and autonomous robotic systems.

Bibliography

[1] P. Molchanov, S. Gupta, K. Kim, and K. Pulli, "Multi-sensor system for drivers hand-gesture recognition," in *IEEE Intl. Conf. Automatic Face and Gesture Recognition*, 2015.

[2] C. Tran, A. Doshi, and M. M. Trivedi, "Modeling and prediction of driver behavior by foot gesture analysis," *Computer Vision and Image Understanding*, vol. 116, pp. 435–445, 2012.

[3] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila, "Context-based pedestrian path prediction," in *European Conf. Computer Vision*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., 2014.

[4] S. Sivaraman, B. Morris, and M. M. Trivedi, "Learning multi-lane trajectories using vehicle-based vision," in *IEEE Intl. Conf. Computer Vision Workshops-CVVT*, 2011.

[5] B. Fröhlich, M. Enzweiler, and U. Franke, "Will this car change the lane? - turn signal recognition in the frequency domain," in *IEEE Intelligent Vehicles Symposium*, 2014.

[6] S. Ullman, "Against direct perception," *Behavioral and Brain Sciences*, vol. 3, no. 03, pp. 373–381, 1980.

[7] E. Ohn-Bar and M. M. Trivedi, "Are all objects equal? deep spatio-temporal importance prediction in driving videos," *Pattern Recognition*, 2016.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.

[9] U. Ozguner, T. Acarman, and K. Redmill, *Autonomous ground vehicles*. Artech House, 2011.

[10] S. Martin, A. Tawari, and M. M. Trivedi, "Toward privacy-protecting safety systems for naturalistic driving videos," *IEEE Trans. Intelligent Transportation Systems*, 2014.

[11] M.-I. Toma, L. J. Rothkrantz, and C. Antonya, "Driver cell phone usage detection on strategic highway research program (SHRP2) face view videos," in *IEEE Intl. Conf. on Cognitive Infocommunications*, 2012.

[12] K. Behn, A. Pavelkov, and A. Herout, "Implicit hand gestures in aeronautics cockpit as a cue for crew state and workload inference," in *IEEE Intl. Conf. Intelligent Transportation Systems*, 2015.

[13] A. Fuentes, R. Fuentes, E. Cabello, C. Conde, and I. Martin, "Videosensor for the detection of unsafe driving behavior in the proximity of black spots," *Sensors*, vol. 14, no. 11, 2014.

[14] F. Attal, A. Boubezoul, L. Oukhellou, and S. Espi, "Riding patterns recognition for powered two-wheelers users' behaviors analysis," in *IEEE Conf. Intelligent Transprtation Systems*, 2013.

[15] A. Bender, G. Agamennoni, J. R. Ward, S. Worrall, and E. M. Nebot, "An unsupervised approach

for inferring driver behavior from naturalistic driving data," *IEEE Trans. Intelligent Transportation Systems*, vol. 16, no. 6, pp. 3325–3336, 2015.

[16] A. Sathyanarayana, S. O. Sadjadi, and J. H. L. Hansen, "Leveraging sensor information from portable devices towards automatic driving maneuver recognition," in *IEEE Conf. Intelligent Transportation Systems*, 2012.

[17] L. M. Bergasa, D. Almera, J. Almazn, J. J. Yebes, and R. Arroyo, "Drivesafe: An app for alerting inattentive drivers and scoring driving behaviors," in *IEEE Intelligent Vehicles Symposium*, 2014.

[18] E. Ohn-Bar, S. Martin, A. Tawari, and M. M. Trivedi, "Head, eye, and hand patterns for driver activity recognition," in *IEEE Intl. Conf. Pattern Recognition*, 2014.

[19] F. Parada-Loira, E. Gonzlez-Agulla, and J. L. Alba-Castro, "Hand gestures to control infotainment equipment in cars," in *IEEE Intelligent Vehicles Symposium*, 2014.

[20] C. Ahlstrom, T. Victor, C. Wege, and E. Steinmetz, "Processing of eye/head-tracking data in large-scale naturalistic driving data sets," *IEEE Trans. Intelligent Transportation Systems*, 2012.

[21] A. Rangesh, E. Ohn-Bar, and M. M. Trivedi, "Hidden hands: Tracking hands with an occlusion aware tracker," in *IEEE Conf. Computer Vision and Pattern Recognition Workshops-HANDS*, 2016.

[22] A. Tawari, K. H. Chen, and M. M. Trivedi, "Where is the driver looking: Analysis of head, eye and iris for robust gaze zone estimation," in *IEEE Intl. Conf. Intelligent Transportation Systems*, 2014.

[23] E. Ohn-Bar and M. M. Trivedi, "Beyond just keeping hands on the wheel: Towards visual interpretation of driver hand motion patterns," in *IEEE Intl. Conf. Intelligent Transportation Systems*, 2014.

[24] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3D convolutional neural networks," *CVPRW*, 2015.

[25] S. Cheng and M. M. Trivedi, "Vision-based infotainment user determination by hand recognition for driver assistance," *IEEE Trans. Intelligent Transportation Systems*, 2010.

[26] A. D. lvarez, Francisco, S. Garca, J. E. Naranjo, J. J. Anaya, and F. Jimnez, "Modeling the driving behavior of electric vehicles using smartphones and neural networks," *IEEE Trans. Intelligent Transportation Systems Magazine*, 2012.

[27] M. Wllmer, C. Blaschke, T. Schindl, B. Schuller, B. Frber, S. Mayer, and B. Trefflich, "Online driver distraction detection using long short-term memory," *IEEE Trans. Intelligent Transportation Systems*, 2011.

[28] P. Jimnez, L. M. Bergasa, J. Nuevo, N. Hernndez, and I. G. Daza, "Gaze fixation system for the evaluation of driver distractions induced by ivis," *IEEE Trans. Intelligent Transportation Systems*, 2012.

[29] R. O. Mbouna, S. G. Kong, and M.-G. Chun, "Visual analysis of eye state and head pose for driver alertness monitoring," *IEEE Trans. Intelligent Transportation Systems*, 2013.

[30] T. Liu, Y. Yang, G.-B. Huang, Y. K. Yeo, and Z. Lin, "Driver distraction detection using semi-supervised machine learning," *IEEE Trans. Intelligent Transportation Systems*, 2015.

[31] F. Vicente, Z. Huang, X. Xiong, F. D. la Torre, W. Zhang, and D. Levi, "Driver gaze tracking and

eyes off the road detection system," *IEEE Trans. Intelligent Transportation Systems*, 2015.

[32] A. Tawari and M. M. Trivedi, "Robust and continuous estimation of driver gaze zone by dynamic analysis of multiple face videos," in *IEEE Intelligent Vehicles Symposium*, 2014.

[33] A. Witayangkurn, T. Horanont, Y. Sekimoto, and R. Shibasaki, "Anomalous event detection on large-scale gps data from mobile phones using hidden markov model and cloud platform," in *Pervasive and Ubiquitous Computing*, 2013.

[34] M.-I. Toma, L. J. Rothkrantz, and C. Antonya, "Car driver skills assessment based on driving postures recognition," in *IEEE Intl. Conf. on Cognitive Infocommunications*, 2012.

[35] M. V. Ly, S. Martin, and M. Trivedi, "Driver classification and driving style recognition using inertial sensors," in *IEEE Intelligent Vehicles Symposium*, 2013.

[36] D. Johnson and M. Trivedi, "Driving style recognition using a smartphone as a sensor platform," in *IEEE Intl. Conf. Intelligent Transportation Systems*, 2011.

[37] S. Lefèvre, A. Carvalho, Y. Gao, H. E. Tseng, and F. Borrelli, "Driver models for personalised driving assistance," *Vehicle System Dynamics*, vol. 53, no. 12, pp. 1705–1720, 2015.

[38] D. Drr, D. Grabengiesser, and F. Gauterin, "Online driving style recognition using fuzzy logic," in *IEEE Intl. Conf. Intelligent Transportation Systems*, 2014.

[39] A. S. Zeeman and M. J. Booysen, "Combining speed and acceleration to detect reckless driving in the informal public transport industry," in *IEEE Intl. Conf. Intelligent Transportation Systems*, 2013.

[40] A. Aljaafreh, N. Alshabatat, and M. S. N. Al-Din, "Driving style recognition using fuzzy logic," in *IEEE Intl. Conf. Vehicular Electronics and Safety*, 2012.

[41] H. Eren, S. Makinist, E. Akin, and A. Yilmaz, "Estimating driving behavior by a smartphone," in *IEEE Intelligent Vehicles Symposium*, 2012.

[42] J. Dai, J. Teng, X. Bai, Z. Shen, and D. Xuan, "Mobile phone based drunk driving detection," in *Intl. Conf. Pervasive Computing Technologies for Healthcare*, 2010.

[43] D. W. Koh and H. B. Kang, "Smartphone-based modeling and detection of aggressiveness reactions in senior drivers," in *IEEE Intelligent Vehicles Symposium*, 2015.

[44] R. Arajo, . Igreja, R. de Castro, and R. E. Arajo, "Driving coach: A smartphone application to evaluate driving efficient patterns," in *IEEE Intelligent Vehicles Symposium*, 2012.

[45] G. Castignani, R. Frank, and T. Engel, "Driver behavior profiling using smartphones," in *IEEE Intl. Conf. Intelligent Transportation Systems*, 2013.

[46] erman Castignani, T. Derrmann, R. Frank, and T. Enge, "Driver behavior profiling using smartphones: A low-cost platform for driver monitoring," *IEEE Intelligent Transportation Systems Magazine*, 2015.

[47] J.-H. Hong, B. Margines, and A. K. Dey, "A smartphone-based sensing platform to model aggressive driving behaviors," in *ACM Conf. Human Factors in Computing Systems*, 2014.

[48] H. Eren, S. Makinist, E. Akin, and A. Yilmaz, "Estimating driving behavior by a smartphone," in

IEEE Intelligent Vehicles Symposium, 2012.

[49] J. Gonalves, J. S. V. Gonalves, R. J. F. Rossetti, and C. Olaverri-Monreal, "Smartphone sensor platform to study traffic conditions and assess driving performance," in *IEEE Conf. on Intelligent Transportation Systems*, 2014.

[50] C. Gold, D. Dambck, L. Lorenz, and K. Bengler, "take over! how long does it take to get the driver back into the loop?" *Human Factors and Ergonomics*, vol. 57, no. 1, pp. 1938–1942, 2013.

[51] V. A. Shia, Y. Gao, R. Vasudevan, K. D. Campbell, T. Lin, F. Borrelli, and R. Bajcsy, "Semiautonomous vehicular control using driver modeling," *IEEE Trans. Intelligent Transportation Systems*, vol. 15, no. 6, pp. 2696–2709, 2014.

[52] V. A. Banks and N. A. Stanton, "Keep the driver in control: Automating automobiles of the future," *Applied Ergonomics*, vol. 53, Part B, pp. 389–395, 2016.

[53] J. Koo, J. Kwac, W. Ju, M. Steinert, L. Leifer, and C. Nass, "Why did my car just do that? explaining semi-autonomous driving actions to improve driver understanding, trust, and performance," *Interactive Design and Manufacturing*, vol. 9, no. 4, pp. 269–275, 2014.

[54] M. Walch, K. Lange, M. Baumann, and M. Weber, "Autonomous driving: Investigating the feasibility of car-driver handover assistance," in *Intl. Conf. AutomotiveUI*, 2015.

[55] C. Braunagel, W. Stolzmann, E. Kasneci, and W. Rosenstiel, "Driver-activity recognition in the context of conditionally autonomous driving," in *IEEE Conf. Intelligent Transportation Systems*, 2015.

[56] S. Lefèvre, J. Ibañez-Guzmán, and C. Laugier, "Context-based estimation of driver intent at road intersections," in *IEEE Symposium on Computational Intelligence in Vehicles and Transportation Systems*, 2011.

[57] A. Nakano, H. Okuda, T. Suzuki, S. Inagaki, and S. Hayakawa, "Symbolic modeling of driving behavior based on hierarchical segmentation and formal grammar," *Intl. Conf. Intelligent Robots and Systems*, pp. 5516–5521, 2009.

[58] A. Jain, H. S. Koppula, B. Raghavan, S. Soh, and A. Saxena, "Car that knows before you do: Anticipating maneuvers via learning temporal driving models," in *IEEE Intl. Conf. Computer Vision*, 2015.

[59] M. Liebner, M. Baumann, F. Klanner, and C. Stiller, "Driver intent inference at urban intersections using the intelligent driver model," in *IEEE Intelligent Vehicles Symposium*, 2012.

[60] H. Berndt and K. Dietmayer, "Driver intention inference with vehicle onboard sensors," in *IEEE Conf. Vehicular Electronics and Safety*, 2009.

[61] S. Lefèvre, C. Laugier, and J. Ibañez-Guzmán, "Evaluating risk at road intersections by detecting conflicting intentions," in *IEEE Conf. Intelligent Robots and Systems*, 2012.

[62] T. Streubel and K. H. Hoffmann, "Prediction of driver intended path at intersections," in *IEEE Intelligent Vehicles Symposium*, 2014, pp. 134–139.

[63] J. Krumm, "A markov model for driver turn prediction," *SAE World Congress*, 2008.

[64] H. Berndt, J. Emmert, and K. Dietmayer, "Continuous driver intent recognition with hidden markov

models," *IEEE Intl. Conf. Intelligent Transportation Systems*, pp. 1189–1194, 2008.

[65] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey, "Maximum entropy inverse reinforcement learning," in *AAAI Conference on Artificial Intelligence*, 2008, pp. 1433–1438.

[66] B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey, "Human behavior modeling with maximum entropy inverse optimal control," *AAAI Conference on Artificial Intelligence*, 2009.

[67] N. Pugeault and R. Bowden, "Learning pre-attentive driving behaviour from holistic visual features," in *European Conf. Computer Vision*, 2010.

[68] B. Tang, S. Khokhar, and R. Gupta, "Turn prediction at generalized intersections," in *IEEE Intelligent Vehicles Symposium*, 2015.

[69] S. Ferguson, B. Luders, R. C. Grande, and J. P. How, "Real-time predictive modeling and robust avoidance of pedestrians with uncertain, changing intentions," in *Intl. Workshop Algorithmic Foundations of Robotics*, 2014.

[70] M. Goldhammer, M. Gerhard, S. Zernetsch, K. Doll, and U. Brunsmann, "Early prediction of a pedestrian's trajectory at intersections," in *IEEE Intl. Conf. Intelligent Transportation Systems*, 2013.

[71] S. Khler, M. Goldhammer, S. Bauer, K. Doll, U. Brunsmann, and K. Dietmayer, "Early detection of the pedestrian's intention to cross the street," in *IEEE Intl. Conf. Intelligent Transportation Systems*, 2012.

[72] F. Madrigal, J.-B. Hayet, and F. Lerasle, "Intention-aware multiple pedestrian tracking," in *IEEE Intl. Conf. Pattern Recognition*, 2014.

[73] R. Quintero, I. Parra, D. Llorca, and M. Sotelo, "Pedestrian path prediction based on body language and action classification," in *IEEE Intl. Conf. Intelligent Transportation Systems*, 2014.

[74] A. Møgelmose, M. Trivedi, and T. Moeslund, "Trajectory analysis and prediction for improved pedestrian safety: Integrated framework and evaluations," in *IEEE Intelligent Vehicles Symposium*, 2015.

[75] C. Keller and D. Gavrila, "Will the pedestrian cross? a study on pedestrian path prediction," *IEEE Trans. Intelligent Transportation Systems*, vol. 15, no. 2, 2014.

[76] T. Gandhi and M. Trivedi, "Image based estimation of pedestrian orientation for improving path prediction," in *IEEE Intelligent Vehicles Symposium*, 2008, pp. 506–511.

[77] A. T. Schulz and R. Stiefelhagen, "Pedestrian intention recognition using latent-dynamic conditional random fields," in *IEEE Intelligent Vehicles Symposium*, 2015.

[78] ——, "A controlled interactive multiple model filter for combined pedestrian intention recognition and path prediction," in *IEEE Intl. Conf. Intelligent Transportation Systems*, 2015.

[79] T. Bandyopadhyay, C. Z. Jie, D. Hsu, M. H. Ang, D. Rus, and E. Frazzoli, "Intention-aware pedestrian avoidance," in *Intl. Symposium on Experimental Robotics*, P. J. Desai, G. Dudek, O. Khatib, and V. Kumar, Eds., 2013.

[80] J. F. P. Kooij, N. Schneider, and D. M. Gavrila, "Analysis of pedestrian dynamics from a vehicle perspective," in *IEEE Intelligent Vehicles Symposium*, 2014.

[81] J. F. P. Kooij, G. Englebienne, and D. M. Gavrila, "Mixture of switching linear dynamics to discover behavior patterns in object tracks," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 322–334, 2016.

[82] W. Choi and S. Savarese, "Understanding collective activities of people from videos," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 36, no. 6, pp. 1242–1257, 2014.

[83] M. Goldhammer, A. Hubert, S. Koehler, K. Zindler, U. Brunsmann, K. Doll, and B. Sick, "Analysis on termination of pedestrians gait at urban intersections," in *IEEE Intl. Conf. Intelligent Transportation Systems*, 2014.

[84] W. Choi, K. Shahid, and S. Savarese, "What are they doing? : Collective activity classification using spatio-temporal relationship among people," in *IEEE Intl. Conf. Computer Vision Workshops*, 2009.

[85] H. Kataoka, Y. Aoki, Y. Satoh, S. Oikawa, and Y. Matsui, "Fine-grained walking activity recognition via driving recorder dataset," in *IEEE Intl. Conf. Intelligent Transportation Systems*, 2015.

[86] J. Kooij, M. Liem, J. Krijnders, T. Andringa, and D. Gavrila, "Multi-modal human aggression detection," *Computer Vision and Image Understanding*, vol. 144, pp. 106–120, 2016.

[87] D. Llorca, R. Quintero, I. Parra, R. Izquierdo, C. Fernandez, and M. Sotelo, "Assistive pedestrian crossings by means of stereo localization and rfid anonymous disability identification," in *IEEE Intl. Conf. Intelligent Transportation Systems*, 2015.

[88] A. Flores and S. Belongie, "Removing pedestrians from google street view images," in *IEEE Conf. Computer Vision and Pattern Recognition Workshops*, 2010.

[89] P. Agrawal and P. Narayanan, "Person de-identification in videos," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 21, 2011.

[90] B. Li, T. Wu, C. Xiong, and S.-C. Zhu, "Recognizing car fluents from video," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2016.

[91] A. Jahangiri, H. A. Rakha, and T. A. Dingus, "Adopting machine learning methods to predict redlight running violations," in *IEEE Intl. Conf. Intelligent Transportation Systems*, 2015.

[92] C. L. Azevedo and H. Farah, "Using extreme value theory for the prediction of head-on collisions during passing maneuvres," in *IEEE Intl. Conf. Intelligent Transportation Systems*, 2015.

[93] T. Gindele, S. Brechtel, and R. Dillmann, "Learning context sensitive behavior models from observations for predicting traffic situations," in *IEEE Intl. Conf. Intelligent Transportation Systems*, 2013.

[94] R. Graf, H. Deusch, F. Seeliger, M. Fritzsche, and K. Dietmayer, "A learning concept for behavior prediction at intersections," in *IEEE Intelligent Vehicles Symposium*, 2014.

[95] G. S. Aoude, B. D. Luders, K. K. H. Lee, D. S. Levine, and J. P. How, "Threat assessment design for driver assistance system at intersections," in *IEEE Conf. Intelligent Transportation Systems*, 2010.

[96] C. Laugier, I. E. Paromtchik, M. Perrollaz, M. Yong, J. D. Yoder, C. Tay, K. Mekhnacha, and A. Ngre, "Probabilistic analysis of dynamic scenes and collision risks assessment to improve driving safety," *IEEE Intelligent Transportation Systems Magazine*, vol. 3, no. 4, pp. 4–19, 2011.

[97] G. Agamennoni, J. I. Nieto, and E. M. Nebot, "A bayesian approach for driving behavior inference," in *IEEE Intelligent Vehicles Symposium*, 2011.

[98] R. K. Satzoda and M. M. Trivedi, "Looking at vehicles in the night: Detection dynamics of rear lights," *IEEE Trans. Intelligent Transportation Systems*, 2016.

[99] M. P. Philipsen, M. B. Jensen, R. K. Satzoda, M. M. Trivedi, A. Møgelmose, and T. B. Moeslund, "Day and night-time drive analysis using stereo vision for naturalistic driving studies," in *IEEE Intelligent Vehicles Symposium*, 2015.

[100] H. Zhang, A. Geiger, and R. Urtasun, "Understanding high-level semantics by modeling traffic patterns," in *IEEE Intl. Conf. on Computer Vision*, 2013.

[101] M. T. Phan, V. Fremont, I. Thouvenin, M. Sallak, and V. Cherfaoui, "Recognizing driver awareness of pedestrian," in *IEEE Intl. Conf. Intelligent Transportation Systems*, 2014, pp. 1027–1032.

[102] R. Tanishige, D. Deguchi, K. Doman, Y. Mekada, I. Ide, and H. Murase, "Prediction of driver's pedestrian detectability by image processing adaptive to visual fields of view," in *IEEE Intl. Conf. Intelligent Transportation Systems*, 2014.

[103] A. Tawari, A. Mogelmose, S. Martin, T. Moeslund, and M. Trivedi, "Attention estimation by simultaneous analysis of viewer and view," in *IEEE Intl. Conf. Intelligent Transportation Systems*, 2014.

[104] B. Morris, A. Doshi, and M. Trivedi, "Lane change intent prediction for driver assistance: On-road design and evaluation," in *IEEE Intelligent Vehicles Symposium*, 2011.

[105] A. Doshi, B. T. Morris, and M. M. Trivedi, "On-road prediction of driver's intent with multimodal sensory cues," *IEEE Pervasive Computing*, vol. 10, pp. 22–34, 2011.

[106] E. Ohn-Bar, A. Tawari, S. Martin, and M. M. Trivedi, "On surveillance for safety critical events: In-vehicle video networks for predictive driver assistance systems," *Computer Vision and Image Understanding*, vol. 134, pp. 130–140, 2015.

[107] J. McCall and M. M. Trivedi, "Driver behavior and situation aware brake assistance for intelligent vehicles," *Proceedings of the IEEE*, vol. 95, pp. 374–387, 2007.

[108] M. Bahram, C. Hubmann, A. Lawitzky, M. Aeberhard, and D. Wollherr, "A combined model- and learning-based framework for interaction-aware maneuver prediction," *IEEE Trans. Intelligent Transportation Systems*, 2016.

[109] A. Doshi and M. M. Trivedi, "Attention estimation by simultaneous observation of viewer and view," in *IEEE Conf. Computer Vision and Pattern Recognition Workshops*, 2010.

[110] ——, "Investigating the relationships between gaze patterns, dynamic vehicle surround analysis, and driver intentions," *IEEE Intelligent Vehicles Symposium*, June 2009.

[111] T. Bar, D. Linke, D. Nienhuser, and J. Zollner, "Seen and missed traffic objects: A traffic object-specific awareness estimation," in *IEEE Intelligent Vehicles Symposium*, 2013.

[112] M. Mori, C. Miyajima, P. Angkititrakul, T. Hirayama, Y. Li, N. Kitaoka, and K. Takeda, "Measuring driver awareness based on correlation between gaze behavior and risks of surrounding vehicles," in *IEEE Conf. Intelligent Transportation Systems*, 2012.

[113] M. Rezaei and R. Klette, "Look at the driver, look at the road: No distraction! no accident!" in *IEEE Conf. Computer Vision and Pattern Recognition*, 2014.

[114] K. Takagi, H. Kawanaka, M. Bhuiyan, and K. Oguri, "Estimation of a three-dimensional gaze point and the gaze target from the road images," in *IEEE Intl. Conf. Intelligent Transportation Systems*, 2011.

[115] A. Tawari, S. Sivaraman, M. M. Trivedi, T. Shannon, and M. Tippelhofer, "Looking-in and looking-out vision for urban intelligent assistance: Estimation of driver attentive state and dynamic surround for safe merging and braking," in *IEEE Intelligent Vehicles Symposium*, 2014.

[116] A. Jain, H. S. Koppula, S. Soh, B. Raghavan, A. Singh, and A. Saxena, "Brain4cars: Car that knows before you do via sensory-fusion deep learning architecture," *CoRR*, vol. abs/1601.00740, 2016.

[117] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *IEEE Intl. Conf. on Computer Vision*, 2009.

[118] S. Martin, E. Ohn-Bar, and M. M. Trivedi, "Automatic critical event extraction and semantic interpretation by looking-inside," in *IEEE Intl. Conf. Intelligent Transportation Systems*, 2015.

[119] O. Kumtepe, G. B. Akar, and E. Yuncu, "Driver aggressiveness detection using visual information from forward camera," in *IEEE Intl. Conf. Advanced Video and Signal Based Surveillance*, 2015, pp. 1–6.

[120] S. Hamdar, "Driver behavior modeling," in *Handbook of Intelligent Vehicles*, 2012, pp. 537–558.

[121] E. Ohn-Bar and M. M. Trivedi, "The power is in your hands: 3D analysis of hand gestures in naturalistic video," in *IEEE Conf. Computer Vision and Pattern Recognition Workshops-AMFG*, 2013.

[122] M. Sivak and B. Schoettle, "Road safety with self-driving vehicles: General limitations and road sharing with conventional vehicles," University of Michigan Transportation Research Institute, Tech. Rep. UMTRI-2015-2, 2015.

[123] K. Bengler, K. Dietmayer, B. Farber, M. Maurer, C. Stiller, and H. Winner, "Three decades of driver assistance systems: Review and future perspectives," *IEEE Intelligent Transportation Systems Magazine*, vol. 6, no. 4, pp. 6–22, 2014.

[124] A. Ess, B. Leibe, K. Schindler, and L. Van Gool, "Robust multiperson tracking from a mobile platform," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1831–1846, 2009.

[125] A. Robicquet, A. Alahi, A. Sadeghian, B. Anenberg, J. Doherty, E. Wu, and S. Savarese, "Forecasting social navigation in crowded complex scenes," *CoRR*, vol. abs/1601.00998, 2016.

[126] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 478–500, 2010.

[127] E. Ohn-Bar, S. Martin, and M. M. Trivedi, "Driver hand activity analysis in naturalistic driving studies: Issues, algorithms and experimental studies," *Journal of Electronic Imaging*, vol. 22, 2013.

[128] S. G. Klauer, T. A. Dingus, V. L. Neale, J. D. Sudweeks, and D. J. Ramsey, "Road safety with self-driving vehicles: General limitations and road sharing with conventional vehicles," Virginia

Tech Transportation Institute, Tech. Rep. DOT HS 810 594, 2006.

[129] A. Rangesh, E. Ohn-Bar, and M. M. Trivedi, "Long-term, multi-cue tracking of hands in vehicles," *IEEE Trans. Intelligent Transportation Systems*, 2016.

[130] E. Ohn-Bar and M. M. Trivedi, "Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations," *IEEE Trans. Intelligent Transportation Systems*, vol. 15, no. 6, pp. 2368–2377, Dec 2014.

[131] D. Tang, H. J. Chang, A. Tejani, and T. K. Kim, "Latent regression forest: Structured estimation of 3d articulated hand posture," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2014.

[132] J. S. Supancic, G. Rogez, Y. Yang, J. Shotton, and D. Ramanan, "Depth-based hand pose estimation: Data, methods, and challenges," in *IEEE Intl. Conf. on Computer Vision*, 2015.

[133] E. Ohn-Bar and M. M. Trivedi, "In-vehicle hand activity recognition using integration of regions," in *IEEE Intelligent Vehicles Symposium*, 2013.

[134] B. D. Ziebart, A. Maas, A. K. Dey, and J. A. Bagnell, "Navigate like a cabbie: probabilistic reasoning from observed context-aware behavior," *in Proceedings of the 10th International Conference on Ubiquitous Computing*, 2008.

[135] I. Nizetic, K. Fertalj, and D. Kalpic, "A prototype for the short-term prediction of moving object's movement using markov chains," *Intl. Conf. Information Technology Interfaces*, pp. 559–564, 2009.

[136] J. Krumm, "Where will they turn: predicting turn proportions at intersections," *Personal and Ubiquitous Computing*, vol. 14, pp. 591–599, 2010.

[137] F. Flohr, M. Dumitru-Guzu, J. Kooij, and D. Gavrila, "Joint probabilistic pedestrian head and body orientation estimation," in *IEEE Intelligent Vehicles Symposium*, 2014.

[138] ——, "A probabilistic framework for joint pedestrian head and body orientation estimation," *IEEE Trans. Intelligent Transportation Systems*, vol. 16, no. 4, pp. 1872–1882, 2015.

[139] E. Rehder, H. Kloeden, and C. Stiller, "Head detection and orientation estimation for pedestrian safety," in *IEEE Intl. Conf. Intelligent Transportation Systems*, 2014.

[140] D. Hall and P. Perona, "Fine-grained classification of pedestrians in video: Benchmark and state of the art," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2015.

[141] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2012.

[142] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," in *Intl. Journal of Robotics Research*, 2013.

[143] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2015.

[144] M. S. Kristoffersen, J. V. Dueholm, R. Satzoda, M. Trivedi, A. Møgelmose, and T. Moeslund, "Understanding surrounding vehicular maneuvers: A panoramic vision-based framework for real-world highway studies," in *IEEE Conf. Computer Vision and Pattern Recognition Workshops-ATS*, 2016.

[145] A. Carvalho, S. Lefèvre, G. Schildbach, J. Kong, and F. Borrelli, "Automated driving: The role of forecasts and uncertaintya control perspective," *European Journal of Control*, vol. 24, pp. 14–32, 2015.

[146] S. Y. Cheng, S. Park, and M. M. Trivedi, "Multi-spectral and multi-perspective video arrays for driver body tracking and activity analysis," *Computer Vision and Image Understanding*, vol. 106, pp. 245–257, 2007.

[147] N. Das, E. Ohn-Bar, and M. M. Trivedi, "On performance evaluation of driver hand detection algorithms: Challenges, dataset, and metrics," in *IEEE Conf. Intelligent Transportation Systems*, 2015.

[148] "VIVA: Vision for intelligent vehicles and applications challenge," http://cvrr.ucsd.edu/vivachallenge/.

[149] M. Cordts, M. Omran, S. Ramos, T. Scharwächter, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes dataset," in *CVPR Workshop on The Future of Datasets in Vision*, 2015.

[150] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[151] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2009.

[152] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes dataset for semantic urban scene understanding," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2016.

[153] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Intl. Conf. Multimedia*, 2009.

[154] A. Ess, B. Leibe, and L. V. Gool, "Depth and appearance for mobile scene analysis," in *IEEE Intl. Conf. on Computer Vision*, 2007.

[155] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "Motchallenge 2015: Towards a benchmark for multi-target tracking," *arXiv:1504.01942 [cs]*, 2015.

[156] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3D pose estimation and tracking by detection," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2010.

[157] E. Ohn-Bar and M. M. Trivedi, "Learning to detect vehicles by clustering appearance patterns," *IEEE Trans. Intelligent Transportation Systems*, 2015.

[158] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Data-driven 3D voxel patterns for object category recognition," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2015.

[159] T. Wu, B. Li, and S. C. Zhu, "Learning and-or models to represent context and occlusion for car detection and viewpoint estimation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2015.

[160] Q. Hu, S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Fast detection of multiple objects in traffic scenes with a common detection framework," *IEEE Trans. Intell. Transp. Syst.*, 2015.

[161] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object

detection and semantic segmentation," in *CVPR*, 2014.

[162] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *Intl. Conf. Learning Representations*, 2014.

[163] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Neural Information Processing Systems*, 2012.

[164] X. Chen, K. Kundu, Y. Zhu, A. Berneshawi, H. Ma, S. Fidler, and R. Urtasun, "3D object proposals for accurate object class detection," in *NIPS*, 2015.

[165] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2016.

[166] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, 1997.

[167] D. Pomerleau, "Alvinn: an autonomous land vehicle in a neural network," in *Neural Information Processing Systems*, 2016.

[168] T. Jochem, D. Pomerleau, B. Kumar, and J. Armstrong, "Pans: A portable navigation platform," in *IEEE Intelligent Vehicles Symposium*, 1995.

[169] D. Pomerleau, "Neural network vision for robot driving," 1995.

[170] C. J. C. H. Watkins, "Learning from delayed rewards," Ph.D. dissertation, King's College, Cambridge, 1989.

[171] M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, "End to end learning for self-driving cars," *CoRR*, vol. arXiv:1604.07316, 2016.

[172] U. Muller, J. Ben, E. Cosatto, B. Flepp, and Y. L. Cun, "Off-road obstacle avoidance through end-to-end learning," in *Neural Information Processing Systems*, 2006.

[173] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[174] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in *Intl. Conf. Learning Representations*, 2016.

[175] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *Intl. Conf. Machine Learning*, 2016.

[176] B. Shi, X. Bai, and C. Yao, "Script identification in the wild via discriminative convolutional neural network," *Pattern Recognition*, vol. 52, pp. 448–458, 2016.

[177] Z. Zuo, G. Wang, B. Shuai, L. Zhao, and Q. Yang, "Exemplar based deep discriminative and shareable feature learning for scene image classification," *Pattern Recognition*, vol. 48, no. 10, pp. 3004–3015, 2015.

[178] B. Pepik, R. Benenson, T. Ritschel, and B. Schiele, "What is holding back convnets for detection?" in *GCPR*, 2015.

[179] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, 2014.

[180] D. Park, D. Ramanan, and C. Fowlkes, "Multiresolution models for object detection," in *European Conf. Computer Vision*, 2010.

[181] Y. Ding and J. Xiao, "Contextual boost for pedestrian detection," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2012.

[182] R. Benenson, M. Mathias, R. Timofte, and L. V. Gool, "Pedestrian detection at 100 frames per second," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2012.

[183] R. N. Rajaram, E. Ohn-Bar, and M. M. Trivedi, "Looking at pedestrians at different scales: A multiresolution approach and evaluations," *TITS*, 2016.

[184] W. Zhang, G. Zelinsky, and D. Samaras, "Real-time accurate object detection using multiple resolutions," in *IEEE Intl. Conf. on Computer Vision*, 2007.

[185] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[186] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," DTIC Document, Tech. Rep., 1985.

[187] ——, "Learning representations by back-propagating errors," *Cognitive modeling*, vol. 5, p. 3, 1988.

[188] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[189] S. Xie and Z. Tu, "Holistically-nested edge detection," in *IEEE Intl. Conf. on Computer Vision*, 2015.

[190] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional networks," in *British Machine Vision Conf.*, 2014.

[191] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Intl. Conf. Learning Representations*, 2015.

[192] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Intl. Journal Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[193] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European Conf. Computer Vision*, 2014.

[194] R. Girshick, "Fast R-CNN," in *Intl. Conf. on Computer Vision*, 2015.

[195] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[196] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Convolutional channel features," in *IEEE Intl. Conf. on Computer Vision*, 2015.

[197] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, "Pedestrian detection with unsupervised multi-stage feature learning," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2013.

[198] D. Osaku, R. Nakamura, L. Pereira, R. Pisani, A. Levada, F. Cappabianco, A. Falco, and J. P. Papa, "Improving land cover classification through contextual-based optimum-path forest," *Information Sciences*, vol. 324, pp. 60–87, 2015.

[199] C. Desai, D. Ramanan, and C. C. Fowlkes, "Discriminative models for multi-class object layout," *Intl. Journal Computer Vision*, vol. 95, no. 1, pp. 1–12, 2011.

[200] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Neural Information Processing Systems*, 2014.

[201] M. Hoai, L. Torresani, F. D. la Torre, and C. Rother, "Learning discriminative localization from weakly labeled data," *Pattern Recognition*, vol. 47, no. 3, pp. 1523–1534, 2014.

[202] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1915–1929, 2013.

[203] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[204] N. Zhang, R. Farrell, F. Iandola, and T. Darrell, "Deformable part descriptors for fine-grained recognition and attribute prediction," in *IEEE Intl. Conf. on Computer Vision*, 2013.

[205] C. Long, X. Wang, G. Hua, M. Yang, and Y. Lin, "Accurate object detection with location relaxation and regionlets relocalization," in *Asian Conf. Computer Vision*, 2014.

[206] D. Hoiem, A. Efros, and M. Hebert, "Putting objects in perspective," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2006.

[207] G. Chen, Y. Ding, J. Xiao, and T. X. Han, "Detection evolution with multi-order contextual co-occurrence," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2013.

[208] Z. Tu and X. Bai, "Auto-context and its application to high-level vision tasks and 3D brain image segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 10, 2010.

[209] L. Quannan, J. Wang, Z. Tu, and D. P. Wipf, "Fixed-point model for structured labeling," in *Intl. Conf. Machine Learning*, 2013.

[210] M. A. Sadeghi and A. Farhadi, "Recognition using visual phrases," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2011.

[211] B. Li, T. Wu, and S.-C. Zhu, "Integrating context and occlusion for car detection by hierarchical And-Or model," in *European Conf. Computer Vision*, 2014.

[212] P. Sermanet and Y. LeCun, "Traffic sign recognition with multi-scale convolutional networks," in *Intl. Joint Conf. Neural Networks*, 2011.

[213] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in *Neural Information Processing Systems*, 2013.

[214] R. Girshick, F. Iandola, T. Darrell, and J. Malik, "Deformable part models are convolutional neural networks," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2015.

[215] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 36, pp. 1532–1545, 2014.

[216] E. Ohn-Bar and M. M. Trivedi, "Fast and robust object detection using visual subcategories," in *CVPRW*, 2014.

[217] M. A. Sadeghi and D. Forsyth, "30Hz object detection with DPM V5," in *European Conf. Computer Vision*, 2014.

[218] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008.

[219] T. Joachims, T. Finley, and C.-N. Yu, "Cutting-plane training of structural SVMs," *Machine Learning*, vol. 77, no. 1, pp. 27–59, 2009.

[220] S. Branson, O. Beijbom, and S. Belongie, "Efficient large-scale structured learning," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2013.

[221] S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher, "Block-coordinate Frank-Wolfe optimization for structural SVMs," in *Intl. Conf. Machine Learning*, 2013.

[222] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *IJCV*, 2010.

[223] M. Blaschko and C. Lampert, "Learning to localize objects with structured output regression," in *European Conf. Computer Vision*, 2008.

[224] D. Hoiem, Y. Chodpathumwan, and Q. Dai, "Diagnosing error in object detectors," in *European Conf. Computer Vision*, 2012.

[225] P.-A. Savalle, S. Tsogkas, G. Papandreou, and I. Kokkinos, "Deformable part models with cnn features," in *ECCVW*, 2014.

[226] L. Wan, D. Eigen, and R. Fergus, "End-to-end integration of a convolutional network, deformable parts model and non-maximum suppression," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2015.

[227] M. S. Kristoffersen, J. V. Dueholm, R. K. Satzoda, M. M. Trivedi, A. Mogelmose, and T. B. Moeslund, "Towards semantic understanding of surrounding vehicular maneuvers: A panoramic vision-based framework for real-world highway studies," in *IEEE Conf. Computer Vision and Pattern Recognition Workshops*, 2016.

[228] V. I. Pavlovic, R. Sharma, and T. S. Huang, "Visual interpretation of hand gestures for human-computer interaction: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 677–695, Jul. 1997.

[229] J. M. Gregers, M. B. Skov, and N. G. Thomassen, "You can touch, but you can't look: interacting with in-vehicle systems," in *SIGCHI Conf. Human Factors in Computing Systems*, 2008.

[230] M. Alpern and K. Minardo, "Developing a car gesture interface for use as a secondary task," in *CHI Human factors in computing systems*, 2003.

[231] F. Parada-Loira, E. González-Agulla, and J. L. Alba-Castro, "Hand gestures to control infotainment equipment in cars," in *IEEE Intell. Veh. Symp.*, 2014.

[232] C. A. Pickering, K. J. Burnham, and M. J. Richardson, "A research study of hand gesture recognition technologies and applications for human vehicle interaction," in *Institution of Engineering and Technology Conference on Automative Electronics*, 2007.

[233] W. J. Horrey, "Assessing the effects of in-vehicle tasks on driving performance," *Ergonomics in Design*, no. 19, pp. 4–7, 2011.

[234] G. Jahn, J. F. Krems, and C. Gelau, "Skill aquisition while operating in-vehicle information systems: Interface design determines the level of safety-relevant distractions," *Human Factors and Ergonomics Society*, no. 51, pp. 136–151, 2009.

[235] S. Klauer, F. Guo, J. Sudweeks, and T. Dingus, "An analysis of driver inattention using a case-crossover approach on 100-car data: Final report," National Highway Traffic Safety Administration, Washington, D.C., Tech. Rep. DOT HS 811 334, 2010.

[236] E. Ohn-Bar, A. Tawari, S. Martin, and M. M. Trivedi, "Predicting driver maneuvers by learning holistic features," in *IEEE Intelligent Vehicles Symposium*, 2014.

[237] C. Tran and M. M. Trivedi, "Vision for driver assistance: Looking at people in a vehicle," in *Visual Analysis of Humans*, 2011, pp. 597–614.

[238] C. Tran and M. Trivedi, "Driver assistance for "keeping hands on the wheel and eyes on the road"," in *IEEE Conf. Veh. Electron. Safety*, 2009.

[239] E. Ohn-Bar, A. Tawari, S. Martin, and M. M. Trivedi, "Vision on wheels: Looking at driver, vehicle, and surround for on-road maneuver analysis," in *Computer Vision and Pattern Recognition Workshops-Mobile Vision*, 2014.

[240] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Efficient model-based 3D tracking of hand articulations using Kinect," in *British Machine Vision Conf.*, 2011.

[241] C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun, "Realtime and robust hand tracking from depth," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2014.

[242] E. Ohn-Bar and M. M. Trivedi, "Joint angles similarities and HOG^2 for action recognition," in *CVPRW-Human Activity Understanding from 3D data*, 2013.

[243] A. Kläser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *British Machine Vision Conf.*, 2008.

[244] W. Heng, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Intl. Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, May. 2013.

[245] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 1–27, 2011.

[246] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund, "Human pose estimation and activity

recognition from multi-view videos: Comparative explorations of recent developments," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 5, pp. 538–552, 2012.

[247] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010.

[248] S. Sathyanarayana, G. Littlewort, and M. Bartlett, "Hand gestures for intelligent tutoring systems: Dataset, techniques evaluation," in *IEEE Intl. Conf. Computer Vision Workshops*, 2013.

[249] S. Hadfield and R. Bowden, "Hollywood 3D: Recognizing actions in 3D natural scenes," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2013.

[250] Y. Zhu, W. Chen, and G. Guo, "Evaluating spatiotemporal interest point features for depth-based action recognition," *Image and Vision Computing*, vol. 32, no. 8, pp. 453–464, 2014.

[251] N. Neverova, C. Wolf, G. Paci, G. Sommavilla, G. W. Taylor, and F. Nebout, "A multi-scale approach to gesture detection and recognition," in *IEEE Intl. Conf. Computer Vision Workshops*, 2013.

[252] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly, "Vision-based hand pose estimation: A review," *Computer Vision and Image Understanding*, vol. 108, no. 12, pp. 52–73, 2007.

[253] C. Keskin, F. Kirac, Y. Kara, and L. Akarun, "Real time hand pose estimation using depth sensors," in *IEEE Intl. Conf. Computer Vision Workshops*, 2011.

[254] G. Evangelidis, G. Singh, and R. Horaud, "Skeletal quads: Human action recognition using joint quadruples," in *IEEE Intl. Conf. Pattern Recognition*, 2014.

[255] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d human skeletons as points in a lie group," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2014.

[256] J. Wang, Z. Liu, and Y. Wu, "Learning actionlet ensemble for 3d human action recognition," in *Human Action Recognition with Depth Cameras*, ser. SpringerBriefs in Computer Science. Springer International Publishing, 2014, pp. 11–40.

[257] I. Kapsouras and N. Nikolaidis, "Action recognition on motion capture data using a dynemes and forward differences representation," *Journal of Visual Communication and Image Representation*, 2014.

[258] N. Pugeault and R. Bowden, "Spelling it out: Real-time ASL fingerspelling recognition," in *IEEE Intl. Conf. Computer Vision Workshops*, 2011.

[259] C. Helen, H. Brian, and B. Richard, "Sign language recognition," in *Visual Analysis of Humans*, 2011, pp. 539–562.

[260] D. Uebersax, J. Gall, M. V. den Bergh, and L. V. Gool, "Real-time sign language letter and word recognition from depth data," in *IEEE Intl. Conf. Computer Vision Workshops*, 2011.

[261] P. Doliotis, A. Stefan, C. McMurrough, D. Eckhard, and V. Athitsos, "Comparing gesture recognition accuracy using color and depth information," in *ACM Conf. Pervasive Technologies Related to Assistive Environments*, 2011.

[262] R.-D. Vatavu, "User-defined gestures for free-hand tv control," in *European conference on Interactive TV and Video*, 2012.

[263] G. Panger, "Kinect in the kitchen: testing depth camera interactions in practical home environments," in *ACM Human Factors in Computing Systems*, 2012.

[264] J. P. Wachs, M. Kölsch, H. Stern, and Y. Edan, "Vision-based hand-gesture applications," *ACM Commun.*, 2011.

[265] C. Kirmizibayrak, N. Radeva, M. Wakid, J. Philbeck, J. Sibert, and J. Hahn, "Evaluation of gesture based interfaces for medical volume visualization tasks," in *Virtual Reality Continuum and Its Applications in Industry*, 2011.

[266] L. Gallo, A. Placitelli, and M. Ciampi, "Controller-free exploration of medical image data: Experiencing the kinect," in *Computer-Based Medical Systems*, 2011.

[267] E. Saba, E. Larson, and S. Patel, "Dante vision: In-air and touch gesture sensing for natural surface interaction with combined depth and thermal cameras," in *IEEE Conf. Emerging Signal Process. Applicat.*, 2012.

[268] C.-Y. Kao and C.-S. Fahn, "A human-machine interaction technique: Hand gesture recognition based on hidden markov models with trajectory of hand motion," *Procedia Engineering*, vol. 15, pp. 3739–3743, 2011.

[269] E. Ozcelik and G. Sengul, "Gesture-based interaction for learning: time to make the dream a reality," *British Journal of Educational Technology*, vol. 43, no. 3, pp. 86–89, 2012.

[270] Z. Ren, J. Meng, and J. Yuan, "Depth camera based hand gesture recognition and its applications in human-computer-interaction," in *IEEE Conf. Inform., Commun. and Signal Process.*, 2011.

[271] J. M. Teixeira, B. Reis, S. Macedo, and J. Kelner, "Open/closed hand classification using kinect data," in *IEEE Symp. on Virtual and Augmented Reality*, 2012.

[272] M. V. den Bergh, D. Carton, R. D. Nijs, N. Mitsou, C. Landsiedel, K. Kuehnlenz, D. Wollherr, L. V. Gool, and M. Buss, "Real-time 3D hand gesture interaction with a robot for understanding directions from humans," in *IEEE RO-MAN*, Aug. 2011.

[273] G. Bailly, R. Walter, J. Mller, T. Ning, and E. Lecolinet, "Comparing free hand menu techniques for distant displays using linear, marking and finger-count menus," in *Human-Computer Interaction-INTERACT*, 2011, vol. 6947, pp. 248–262.

[274] C. Keskin, A. Cemgil, and L. Akarun, "DTW based clustering to improve hand gesture recognition," in *Human Behavior Unterstanding*, 2011, vol. 7065, pp. 72–81.

[275] D. Minnen and Z. Zafrulla, "Towards robust cross-user hand tracking and shape recognition," in *IEEE Intl. Conf. Computer Vision Workshops*, 2011.

[276] O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2013.

[277] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *IEEE Intl. Conf. Computer Vision Workshops*, 2005.

[278] L. Xia and J. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2013.

[279] M. Zobl, R. Nieschulz, M. Geiger, M. Lang, and G. Rigoll, "Gesture components for natural in-

teraction with in-car devices," in *Gesture-Based Communication in Human-Computer Interaction*, 2004, vol. 2915, pp. 367–368.

[280] A. Riener, A. Ferscha, F. Bachmair, P. Hagmüller, A. Lemme, D. Muttenthaler, D. Pühringer, H. Rogner, A. Tappe, and F. Weger, "Standardization of the in-car gesture interaction space," in *ACM Automotive User Interfaces and Interactive Vehicular Applications*, 2013.

[281] F. Althoff, R. Lindl, and L. Walchshaeusl, "Robust multimodal hand and head gesture recognition for controlling automotive infotainment systems," in *VDI-Tagung: Der Fahrer im 21*, 2005.

[282] C. Endres, T. Schwartz, and C. A. Müller, "Geremin": 2D microgestures for drivers based on electric field sensing," in *ACM Conf. Intell. User Interfaces*, 2011.

[283] S. Y. Cheng and M. M. Trivedi, "Vision-based infotainment user determination by hand recognition for driver assistance," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 3, pp. 759–764, Sep. 2010.

[284] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2005.

[285] E. Ohn-Bar, C. Tran, and M. Trivedi, "Hand gesture-based visual user interface for infotainment," in *ACM Automotive User Interfaces and Interactive Vehicular Applications*, 2012.

[286] G. Rogez, J. S. Supančič, and D. Ramanan, "First-person pose recognition using egocentric workspaces," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2015.

[287] S. Wan and J. K. Aggarwal, "Mining discriminative states of hands and objects to recognize egocentric actions with a wearable rgbd camera," in *CVPRW-HANDS*, 2015.

[288] S. Lee, S. Bambach, D. Crandall, J. Franchak, and C. Yu, "This hand is my hand: A probabilistic approach to hand disambiguation in egocentric video," in *CVPRW-Egocentric Vision*, 2014.

[289] S. Bambach, "A survey on recent advances of computer vision algorithms for egocentric video," *arXiv preprint arXiv:1501.02825*, 2013.

[290] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund, "Human action recognition using multiple views: A comparative perspective on recent developments," in *Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding*, 2011, pp. 47–52.

[291] C. Tran and M. Trivedi, "3-D Posture and Gesture Recognition for Interactivity in Smart Spaces," *IEEE Trans. Industrial Informatics*, vol. 8, no. 1, pp. 178–187, Feb 2012.

[292] T. H. Poll, "Most U.S. Drivers Engage in 'Distracting' Behaviors: Poll," no. FMCSA-RRR-09-042, 2011.

[293] "The SHRP2 naturalistic driving study. Transportation Research board. 2012. http://www.shrp2nds.us/."

[294] R. Lockton and A. Fitzgibbon, "Real-time gesture recognition using deterministic boosting," in *British Machine Vision Conf.*, 2002.

[295] R. Wang and J. Popovic, "Real-time Hand-tracking with a Color Glove," *ACM Trans. Graph.*, vol. 28, no. 3, pp. 63:1–63:8, Jul 2009.

[296] A. Mittal, A. Zisserman, and P. Torr, "Hand detection using multiple proposals," in *British Machine*

Vision Conf., 2011.

[297] E.-J. Ong and R. Bowden, "A boosted classifier tree for hand shape detection," in *IEEE Conf. Automatic Face and Gesture Recognition*, May 2004, pp. 889–894.

[298] P. Dollár, "Piotr's Computer Vision Matlab Toolbox (PMT)," http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html.

[299] S. Y. Cheng and M. M. Trivedi, "Vision-based Infotainment User Determination by Hand Recognition for Driver Assistance," *IEEE. Trans. Intell. Transport. Sys.*, vol. 11, no. 3, pp. 759–764, Sep. 2010.

[300] J. Fritsch, T. Kuehnl, and A. Geiger, "A new performance measure and evaluation benchmark for road detection algorithms," in *IEEE Conf. Intelligent Transportation Systems*, 2013.

[301] S. Y. Cheng, S. Park, and M. M. Trivedi, "Multi-spectral and multi-perspective video arrays for driver body tracking and activity analysis," *Computer Vision and Image Understanding*, vol. 106, no. 2, pp. 245–257, 2007.

[302] A. Doshi and M. M. Trivedi, "Tactical driver behavior prediction and intent inference: A review," in *IEEE Conf. Intelligent Transportation Systems*, 2011.

[303] C. Tran and M. M. Trivedi, "3D posture and gesture recognition for interactivity in smart space," *IEEE Trans. on Industrial Informatics*, vol. 8, pp. 178–187, 2012.

[304] C. Li and K. M. Kitani, "Pixel-level hand detection in ego-centric videos," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2013.

[305] B. T. Morris and M. M. Trivedi, "Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2011.

[306] J. Sun, W. Xiao, Y. Shuicheng, C. Loong-Fah, C. Tat-Seng, and L. Jintao, "Hierarchical spatio-temporal context modeling for action recognition," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2009.

[307] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[308] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[309] L. P. Morency, A. Quattoni, and T. Darrel, "Latent-dynamic discriminative models for continuous gesture recognition," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2007.

[310] Y. Song, L. P. Morency, and R. Davis, "Multi-view latent variable discriminative models for action recognition," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2012.

[311] E. Ohn-Bar and M. M. Trivedi, "Hand gesture recognition in real-time for automotive interfaces: A multimodal vision-based approach and evaluations," *IEEE Trans. Intelligent Transportation Systems*, 2014.

[312] R. K. Satzoda and M. M. Trivedi, "Drive analysis using vehicle dynamics and vision-based lane semantics," *IEEE Trans. Intell. Transp. Syst.*, 2014.

[313] J. Tison, N. Chaudhary, and L. Cosgrove, "National phone survey on distracted driving attitudes and behaviors," National Highway Traffic Safety Administration, Washington, D.C., Tech. Rep. DOT HS 811 555, Dec. 2011.

[314] T. H. Poll, "Most U.S. drivers engage in 'distracting' behaviors: Poll," Insurance Institute for Highway Safety, Arlington, Va., Tech. Rep. FMCSA-RRR-09-042, Nov. 2011.

[315] D. D. Waard, T. G. V. den Bold, and B. Lewis-Evans, "Driver hand position on the steering wheel while merging into motorway traffic," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 13, no. 2, pp. 129 – 140, 2010.

[316] M. F. Land and D. N. Lee, "Where we look when we steer." *Nature*, vol. 369, no. 6483, pp. 742–744, 1994.

[317] A. Doshi and M. M. Trivedi, "Head and eye gaze dynamics during visual attention shifts in complex environments," *Journal of Vision*, vol. 12, no. 2, 2012.

[318] A. Inhoff and J. Wang, "Encoding of text, manual movement planning, and eye-hand coordination during copy-typing," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 18, pp. 437–448, 1992.

[319] A. E. Patla and J. Vickers, "Where and when do we look as we approach and step over an obstacle in the travel path?" *Neuroreport*, vol. 8, no. 17, pp. 3661–3665, 1997.

[320] J. Vickers, "Encoding of text, manual movement planning, and eye-hand coordination during copy-typing," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 22, pp. 342–354, 1996.

[321] M. F. Land and P. McLeod, "From eye movements to actions: How batsmen hit the ball." *Nature Neuroscience*, vol. 3, pp. 1340–1345, 2000.

[322] J. Pelz, M. Hayhoe, and R. Loeber, "The coordination of eye, head, and hand movements in a natural task," *Experimental Brain Research*, vol. 139, no. 3, pp. 266–277, 2001.

[323] S. Martin, A. Tawari, E. Murphy-Chutorian, S. Y. Cheng, and M. Trivedi, "On the design and evaluation of robust head pose for visual user interfaces: algorithms, databases, and comparisons," in *ACM Conf. Automotive User Interfaces and Interactive Vehicular Applications*, 2012.

[324] M. V. den Bergh and L. V. Gool, "Combining rgb and tof cameras for real-time 3d hand gesture interaction," in *IEEE Workshop on Applications of Computer Vision*, 2011.

[325] V. Harini, S. Atev, N. Bird, P. Schrater, and N. Papanikolopoulos, "Driver activity monitoring through supervised and unsupervised learning," *IEEE Trans. Intell. Transp. Syst.*, 2005.

[326] A. Kurakin, Z. Zhang, and Z. Liu, "A real time system for dynamic hand gesture recognition with a depth sensor," in *European Signal Processing Conf.*, 2012.

[327] C. Tran and M. M. Trivedi, "Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 5, pp. 538–552, Sep. 2012.

[328] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2012.

[329] X. Xiong and F. D. la Torre, "Supervised descent method and its application to face alignment," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2013.

[330] A. Tawari, S. Martin, and M. M. Trivedi, "Continuous head movement estimator (CoHMET) for driver assistance: Issues, algorithms and on-road evaluations," *IEEE Trans. Intelligent Transportation Systems*, vol. 15, pp. 818–830, 2014.

[331] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[332] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *Journal of Machine Learning Research*, vol. 2, pp. 265–292, 2001.

[333] F. R. Bach, D. Heckerman, and E. Horvitz, "Considering cost asymmetry in learning classifiers," *The Journal of Machine Learning Research*, vol. 7, pp. 1713–1741, 2006.

[334] B. Bhanu, C. V. Ravishankar, A. K. Roy-Chowdhury, H. Aghajan, and D. Terzopoulos, Eds., *Distributed Video Sensor Networks*. Springer, 2011.

[335] S. Calderara, A. Prati, and R. Cucchiara, "Hecol: Homography and epipolar-based consistent labeling for outdoor park surveillance," *Computer Vision and Image Understanding*, vol. 111, pp. 21–42, 2008.

[336] "2012 motor vehicle crashes: overview," National Highway Traffic Safety Administration, Washington, D.C., Tech. Rep. DOT HS 811 856, 2013.

[337] W. G. Najm, R. Ranganathan, G. Srinivasan, J. D. Smith, S. Toma, E. Swanson, and A. Burgett, "Description of light-vehicle pre-crash scenarios for safety applications based on vehicle-to-vehicle communications," National Highway Traffic Safety Administration, Washington, D.C., Tech. Rep. DOT HS 811 731, 2013.

[338] P. M. Valero-Moraa, A. Tontscha, R. Welshb, A. Morrisb, S. Reedb, K. Touliouc, and D. Margaritisc, "Is naturalistic driving research possible with highly instrumented cars? lessons learnt in three research centres," *Accident Analysis and Prevention*, vol. 58, pp. 187–194, 2013.

[339] T. Taylora, A. Pradhanb, G. Divekara, M. Romosera, J. Muttarta, R. Gomeza, A. Pollatsekc, and D. Fisherd, "The view from the road: The contribution of on-road glance-monitoring technologies to understanding driver behavior," *Accident Analysis and Prevention*, vol. 58, pp. 175–186, 2013.

[340] "A comprehensive examination of naturalistic lane-changes," National Highway Traffic Safety Administration, Washington, D.C., Tech. Rep. DOT HS 809 702, 2004.

[341] R. Simmons, B. Browning, Y. Zhang, and V. Sadekar, "Learning to predict driver route and destination intent," in *IEEE Conf. Intelligent Transportation Systems*, 2006.

[342] S. Lefèvre, C. Laugier, and J. Ibañez-Guzmán, "Exploiting map information for driver intention estimation at road intersections," in *IEEE Intelligent Vehicles Symposium*, 2011.

[343] M. Ortiz, F. Kummert, and J. Schmudderich, "Prediction of driver behavior on a limited sensory setting," in *IEEE Conf. Intelligent Transportation Systems*, 2012.

[344] V. Gadepally, A. Krishnamurthy, and U. Ozguner, "A framework for estimating driver decisions near intersections," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 2, 2014.

[345] F. Lethaus, M. R. Baumann, F. Kster, and K. Lemmer, "A comparison of selected simple supervised learning algorithms to predict driver intent based on gaze data," *Neurocomputing*, vol. 121, no. 0, pp. 108–130, 2013.

[346] M.-I. Toma and D. Datcu, "Determining car driver interaction intent through analysis of behavior patterns," in *Technological Innovation for Value Creation*. Springer, 2012, pp. 113–120.

[347] S. Haufe, M. S. Treder, M. F. Gugler, M. Sagebaum, G. Curio, and B. Blankertz, "EEG potentials predict upcoming emergency brakings during simulated driving," *Journal of Neural Engineering*, vol. 8, p. 056001, 2011.

[348] R. Cucchiara, A. Prati, and R. Vezzani, "A multi-camera vision system for fall detection and alarm generation," *Expert Systems*, vol. 24, pp. 334–345, 2007.

[349] L. An, M. Kafai, and B. Bhanu, "Dynamic bayesian network for unconstrained face recognition in surveillance camera networks," *IEEE Trans. Emerging and Selected Topics in Circuits and Systems*, vol. 3, no. 2, pp. 155–164, June 2013.

[350] C. Zhang and P. A. Viola, "Multiple-instance pruning for learning efficient cascade detectors," in *Advances in Neural Information Processing Systems*, 2007.

[351] J. Bouguet, "Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm," *Intel Corporation*, 2001.

[352] S. Martin, C. Tran, and M. Trivedi, "Optical flow based head movement and gesture analyzer (OHMeGA)," in *IEEE Intl. Conf. Pattern Recognition*, 2012.

[353] L. P. Morency, A. Quattoni, and T. Darrell, "Latent-dynamic discriminative models for continuous gesture recognition," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2007.

[354] S. Bucak, R. Jin, and A. K. Jain, "Multi-label multiple kernel learning by stochastic approximation: Application to visual object recognition," in *Advances in Neural Information Processing Systems*, 2010.

[355] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, pp. 1848–1853, 2007.

[356] E. Ohn-Bar and M. M. Trivedi, "Looking at humans in the age of self-driving and highly automated vehicles," *IEEE Transactions on Intelligent Vehicles*, 2016.

[357] A. Doshi and M. M. Trivedi, "Tactical driver behavior prediction and intent inference: A review," in *IEEE Conf. Intell. Transp. Syst.*, 2011.

[358] A. Jain, A. Singh, H. S. Koppula, S. Soh, and A. Saxena, "Recurrent neural networks for driver activity anticipation via sensory-fusion architecture," in *IEEE Intl. Conf. Robotics and Automation*, 2016.

[359] A. Tawari, S. Sivaraman, M. M. Trivedi, T. Shannon, and M. Tippelhofer, "Looking-in and looking-out vision for urban intelligent assistance: Estimation of driver attentive state and dynamic surround for safe merging and braking," in *IEEE Intelligent Vehicles Symposium*, 2014.

[360] E. Ohn-Bar and M. M. Trivedi, "What makes an on-road object important?" in *Intl. Conf. Pattern Recognition*, 2016.

[361] A. Borji, Dicky, N. Sihite, and L. Itti, "Probabilistic learning of task-specific visual attention," in *CVPR*, 2012.

[362] A. Doshi and M. M. Trivedi, "Attention estimation by simultaneous observation of viewer and view," in *CVPRW*, 2010.

[363] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *HRI*, 2013.

[364] G. Rogez, J. S. Supancic, and D. Ramanan, "Understanding everyday hands in action from RGB-D images," in *ICCV*, 2015.

[365] T. Li, T. Mei, I. S. Kweon, and X. S. Hua, "Contextual bag-of-words for visual categorization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 4, pp. 381–392, 2011.

[366] Y. Wang, T. Mei, S. Gong, and X.-S. Hua, "Combining global, regional and contextual features for automatic image annotation," *Pattern Recognition*, vol. 42, no. 2, pp. 259–266, 2009.

[367] A. Berg, T. Berg, H. Daume, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, and K. Yamaguchi, "Understanding and predicting importance in images," in *CVPR*, 2012.

[368] H. Pirsiavash, C. Vondrick, and A. Torralba, "Assessing the quality of actions," in *ECCV*, 2014.

[369] W. Chen, C. Xiong, R. Xu, and J. J. Corso, "Actionness ranking with lattice conditional ordinal random fields," in *CVPR*, 2014.

[370] Y. J. Lee and K. Grauman, "Predicting important objects for egocentric video summarization," *IJCV*, vol. 114, no. 1, pp. 38–55, 2015.

[371] C. S. Mathialagan, A. C. Gallagher, and D. Batra, "Vip: Finding important people in images," in *CVPR*, 2015.

[372] N. Pugeault and R. Bowden, "Learning pre-attentive driving behaviour from holistic visual features," in *ECCV*, 2010.

[373] D. M. Y. Zhu, Y. Tian and P. Dollár, "Semantic amodal segmentation," *CoRR*, vol. abs/1509.01329, 2015.

[374] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. Platt, C. Zitnick, and G. Zweig, "From captions to visual concepts and back," in *CVPR*, 2015.

[375] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[376] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *IJCV*, pp. 1–42, 2015.

[377] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *PAMI*, vol. 35, no. 8, pp. 1915–1929, 2013.

[378] X. Chen and A. Gupta, "Webly supervised learning of convolutional networks," in *ICCV*, 2015.

[379] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR*, 2011.

[380] R. N. Rajaram, E. Ohn-Bar, and M. M. Trivedi, "An exploration of why and when pedestrian detection fails," in *ITSC*, 2015.

[381] F. Flohr, M. Dumitru-Guzu, J. Kooij, and D. Gavrila, "A probabilistic framework for joint pedestrian head and body orientation estimation," *TITS*, vol. 16, no. 4, pp. 1872–1882, 2015.

[382] J. Kooij, N. Schneider, F. Flohr, and D. Gavrila, "Context-based pedestrian path prediction," in *ECCV*, 2014.

[383] T. Gandhi and M. M. Trivedi, "Pedestrian protection systems: Issues, survey, and challenges," *IEEE Trans. Intelligent Transportation Systems*, vol. 8, pp. 413–, 2007.

[384] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[385] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[386] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *AAAI*, 2015.

[387] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural networks," *CVPR*, 2016.

[388] E. Ohn-Bar, A. Tawari, S. Martin, and M. M. Trivedi, "On surveillance for safety critical events: In-vehicle video networks for predictive driver assistance systems," *Computer Vision and Image Understanding*, vol. 134, pp. 130–140, 2015.

[389] A. Jain, A. Singh, H. S. Koppula, S. Soh, and A. Saxena, "Recurrent neural networks for driver activity anticipation via sensory-fusion architecture," *ICRA*, 2016.

[390] A. Doshi and M. M. Trivedi, "Examining the impact of driving style on the predictability and responsiveness of the driver: real-world and simulator analysis," *in Proceedings of the IEEE Intelligent Vehicles Symposium*, pp. 232–237, 2010.

[391] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.

[392] O. Beijbom, M. Saberian, D. Kriegman, and N. Vasconcelos, "Guess-averse loss functions for cost-sensitive multiclass boosting," in *ICML*, 2014.

[393] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *PAMI*, vol. 31, no. 12, pp. 2179–2195, 2009.

[394] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila, "Multi-cue pedestrian classication with partial occlusion handling," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2010.