UNIVERSITY OF CALIFORNIA, SAN DIEGO

Contextual 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

2016

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

2016

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 Dignazio, 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.

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 1 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. Chapter 2 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, and Mohan M. Trivedi. Chapter 3 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.

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
2011-2016	Graduate Student Researcher, University of California, San Diego
2016	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? 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 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, 2016

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, and multi-modal data with applications to human-robot interactivity. With the goal of developing contextual systems for interactivity, several key contributions are proposed (1) Context for robust image-level object detection, including vehicles, pedestrians, hands, and faces in on-road setting. (2) Context for spatio-temporal multi-modal and multi-cue driver behavior representation. Finally, the thesis develops a human-centric framework for object recognition by analyzing a notion of object importance and relevance, as measured in a spatio-temporal context of navigation/driving a vehicle. The framework 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 case studies will emphasize the safety-critical application of autonomous vehicles, the contributions of this research could be extended to applications in other domains of human-robot interactivity.

Chapter 1

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 required 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 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

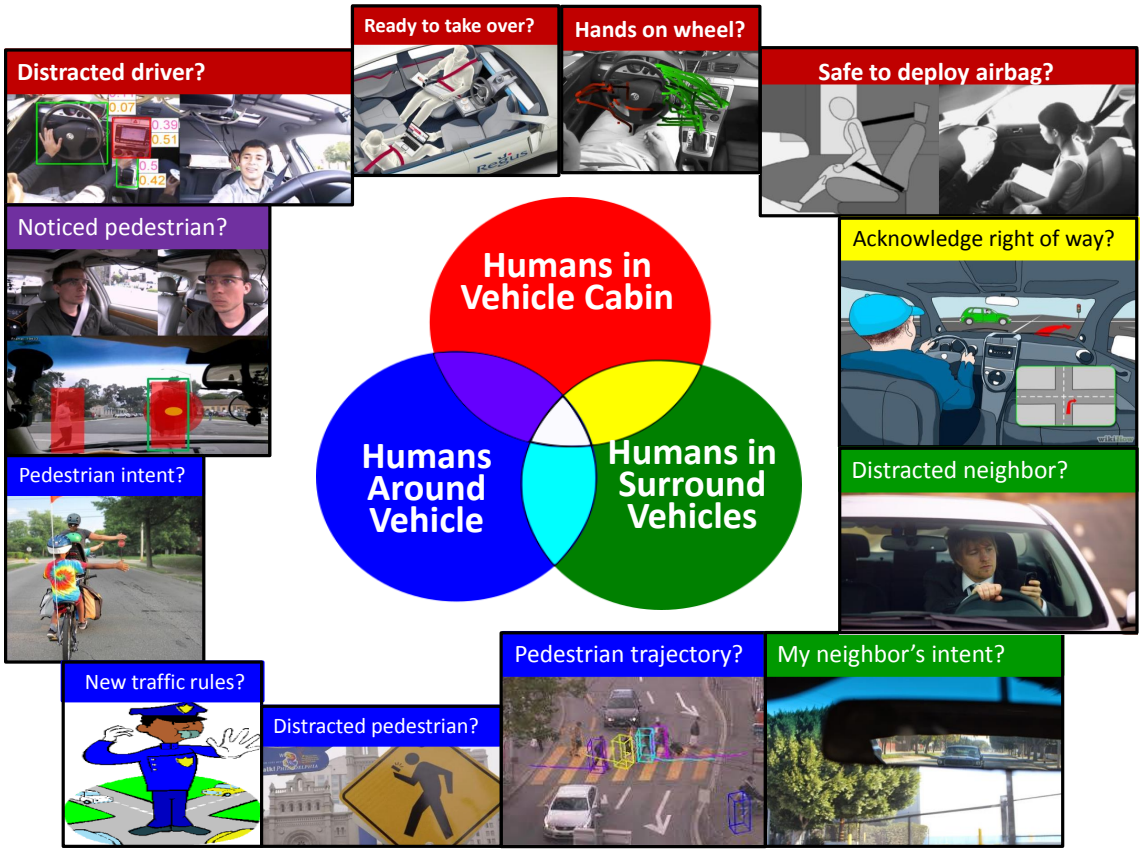


Figure 1.1: Intricate roles of humans to be considered in the development of highly automated and self-driving vehicles. For a safe and comfortable ride, intelligent vehicles must observe, understand, model, infer, and predict behavior of occupants inside the vehicle cabin, pedestrians around the vehicle, and humans in surrounding vehicles.

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, 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

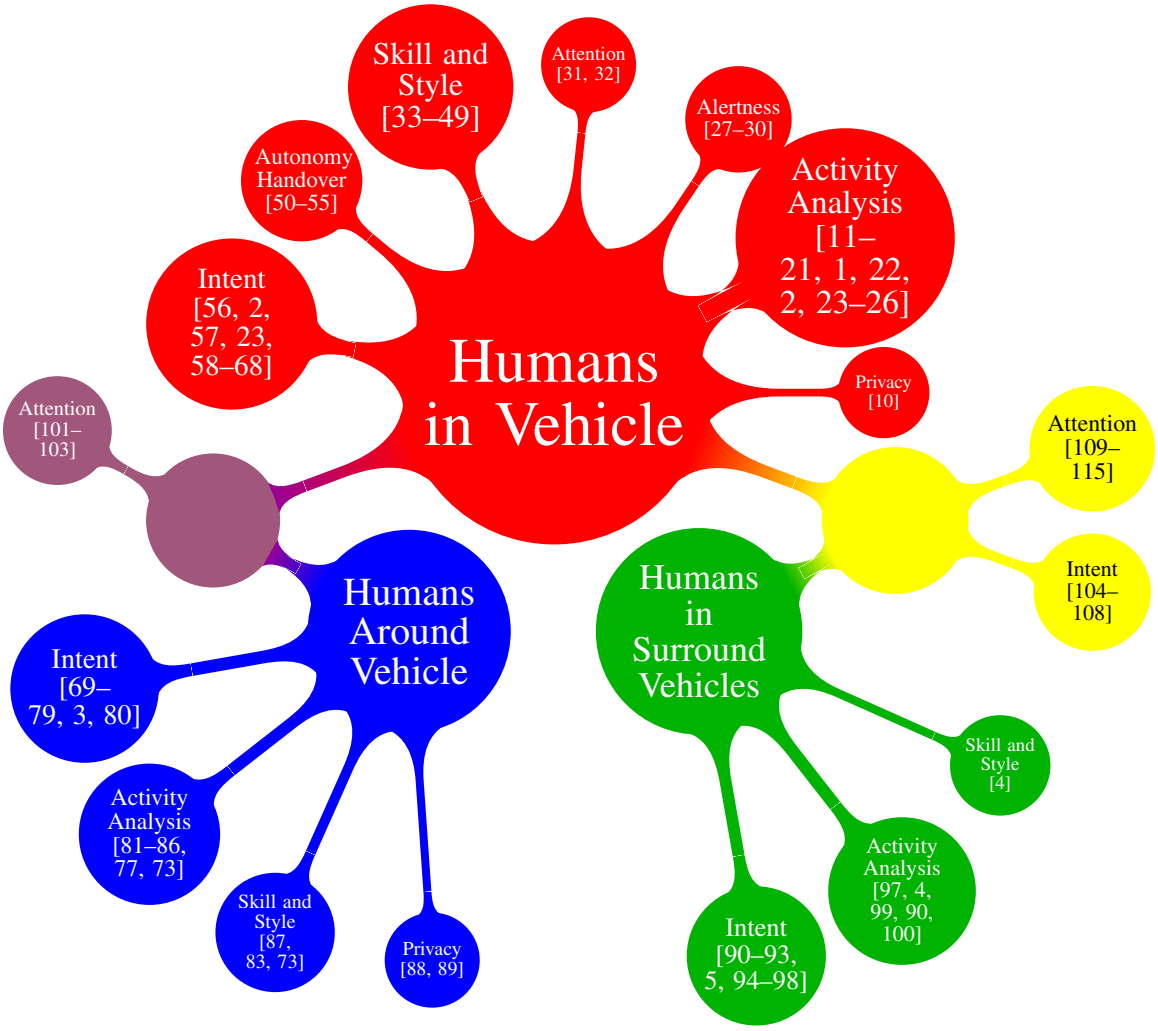


Figure 1.2: Trends in human-centric intelligent vehicle research. The figure visualizes related research studies discussed in this work as they relate to different semantic goals, from maneuver analysis and prediction, to style modeling. Each topic size is proportional the the count of studies surveyed it contains.

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 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. 1.1):

- **Humans in vehicle cabin:** Whether the humans in the vehicle cabin are active drivers, passengers,

Table 1.1: Overview of human-centric related research studies by research goal and human-centric cues employed. Goal types follow Table 2.1, 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.

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

Table 1.2: 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]

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 2.2 for providing a highlight of the current research landscape. Section 2.2 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 1.2 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 1.2.

1.1 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 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. 1.1).

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. 1.1, 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 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 2.1. Furthermore, we detail the specific type of cues employed by selected studies in Table 1.1, which complements the other two clustering techniques described above.

Fig. 1.2 shows a domain-based and research goal-based clustering of the papers listed in the corresponding Table 2.1. An emphasis is put on recent studies (mostly after 2008). In Fig. 1.2, the size of the node is proportional to the number of studies it contains. Fig. 1.2 can be used to draw several conclusions.

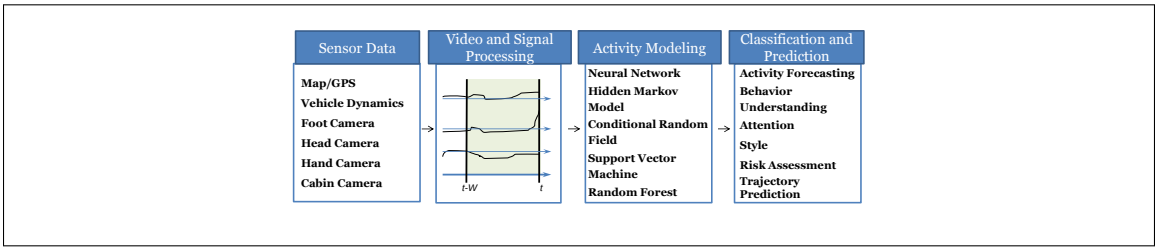


Figure 1.3: Overview of the sensing and learning pipeline commonly used to study humans in the cabin.

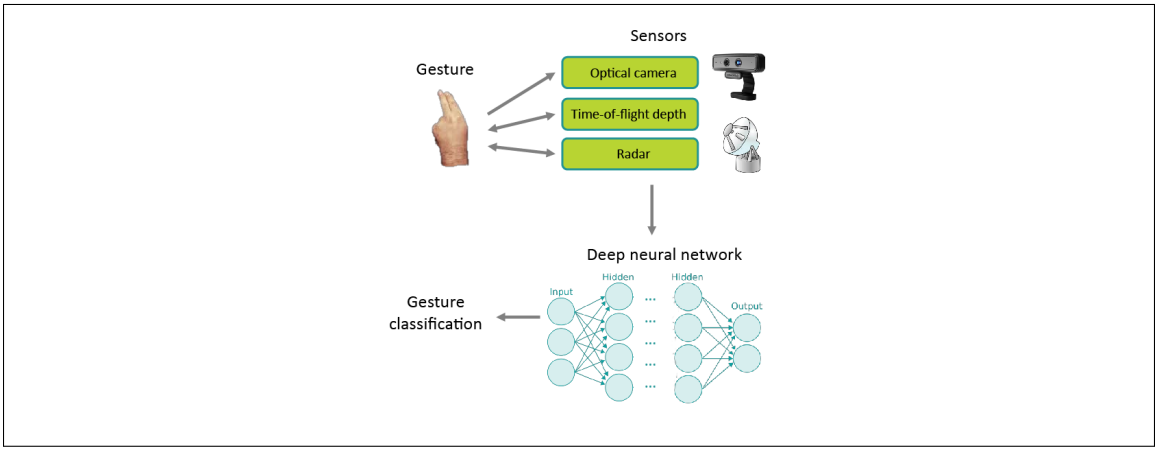
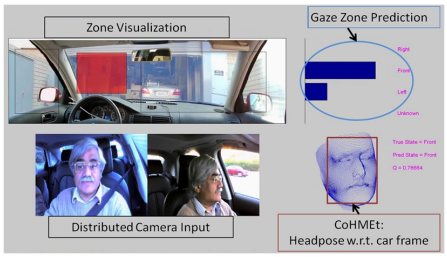


Figure 1.4: A multi-sensor driver gesture recognition system with a deep neural network [1].

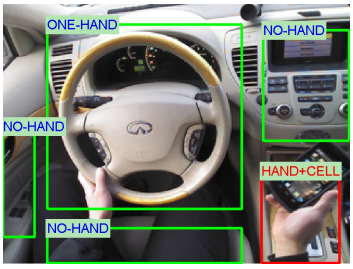
We first identify trends, and then discuss further detail of the studies in each domain in the following sections (Section 1.1.1, 1.1.2, 1.1.3).

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 1.1. Certain research tasks, such as skill and style of humans, in-vehicle occupant interaction, and activity analysis of passengers, has seen less attention.

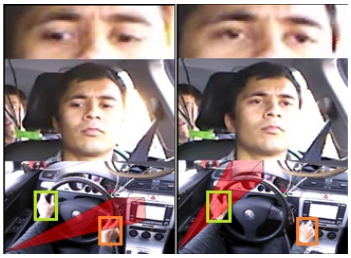
Fig. 1.2 allows for a high-level comparison between the domain of looking at humans inside the vehicle and the other two domains. Although human drivers can analyze fine-grained pose, style, and activity cues for identification of agent intent in all three domains (see Fig. 1.1), fine-grained semantic 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 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



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



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

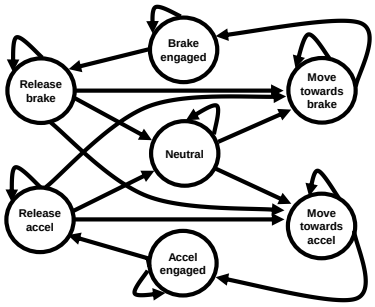


(c) Head, hand, and eye cue integration
for secondary task activity analysis [118, 18].

(d) Cabin occupant activity and interaction analysis.

Figure 1.5: Emerging research topics for studying humans inside the vehicle.

(a) Foot motion tracking.

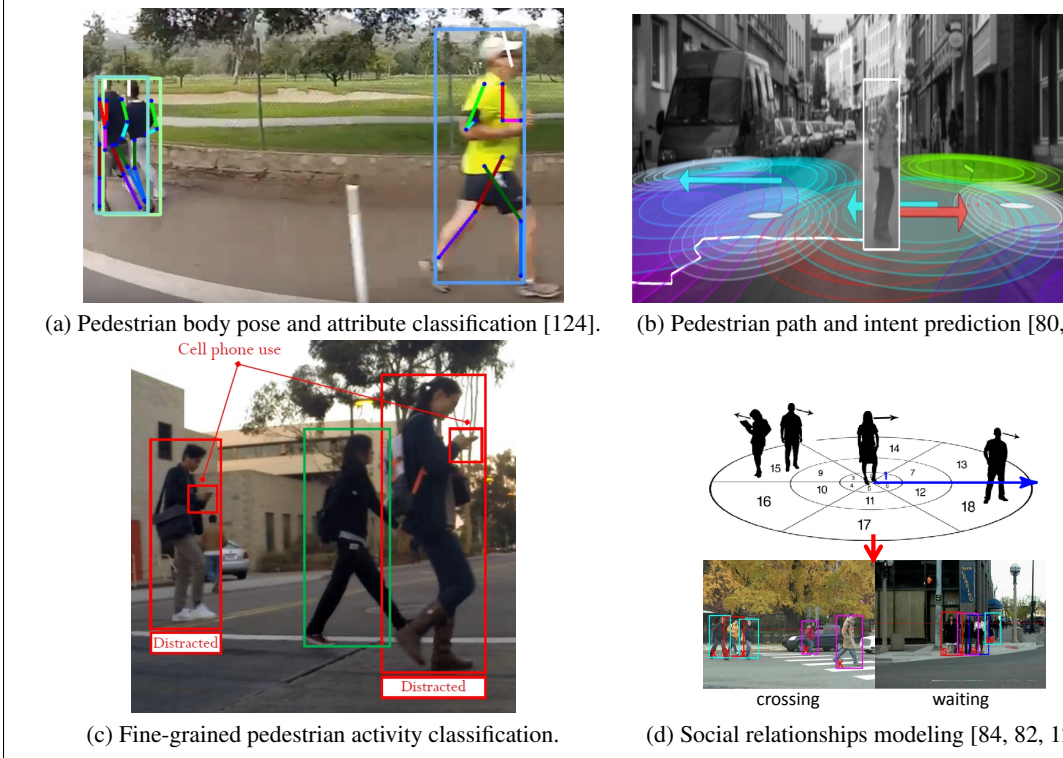


(b) Activity state model.

Figure 1.6: Foot gesture recognition and prediction using a motion tracker and a temporal state model, such as a Hidden Markov Model [2].

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



(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 1.7: Emerging research topics for studying people around the vehicle.

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 1.1.4.

Fig. 1.2 and Table 2.1 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 1.1 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.

1.1.1 Looking at Humans in the Cabin

The surveyed papers in Fig. 1.2 show large diversity in terms of the research tasks for studying humans inside the vehicle. Further detail is provided in Table 1.1 in terms of study details and cue analyzed. A highlight of the research tasks is shown in Fig. 1.5, with an example research pipeline in Figs. 1.3 and 1.4. Dynamics of driver body pose, such as head [32], hand [23], eye [28], and foot [2] (Fig. 1.6) can be

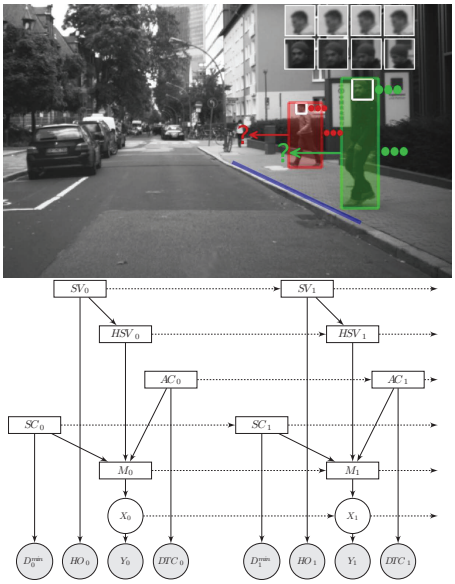


Figure 1.8: Pedestrian path prediction using a Dynamic Bayesian Network for incorporating contextual cues of pedestrian head orientation and situational awareness, situation criticality, and spatial layout cues [3].

employed for in-cabin analysis of secondary tasks [11, 18, 31, 22, 20, 126, 127] and 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. 1.5 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 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.

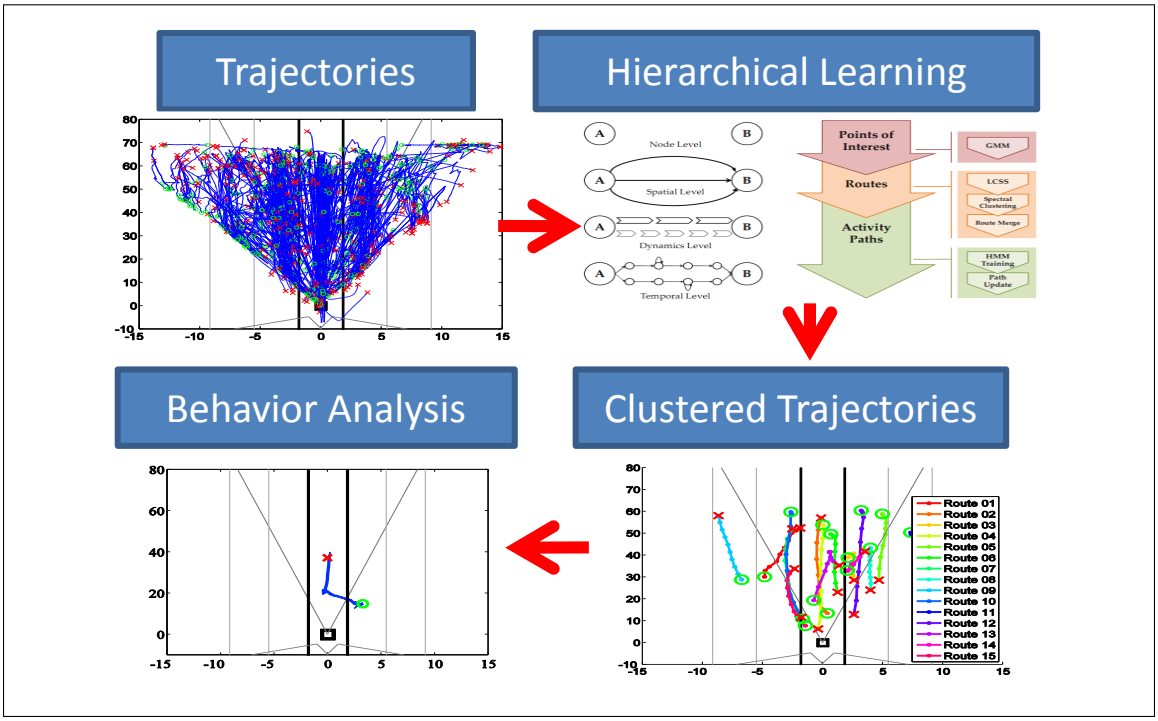


Figure 1.9: Activity analysis of people in surrounding vehicles. In [4], a hierarchical representation of the trajectory dynamics is used to perform behavior analysis of vehicle motion patterns. A Hidden Markov Model is used to perform trajectory classification and detect abnormal trajectory events.

1.1.2 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 1.1 demonstrates a variety of research goals and cues employed to study pedestrians, with a highlight of research tasks shown in Fig. 1.7. 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. 1.8). Gait analysis using body pose for walking activity classification has been studied in [83, 85]. Spatio-temporal relationships between people have been incorporated in [84] for activity classification. As shown in Table 2.1, 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

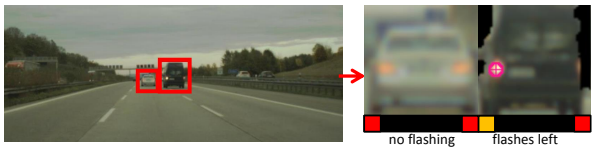
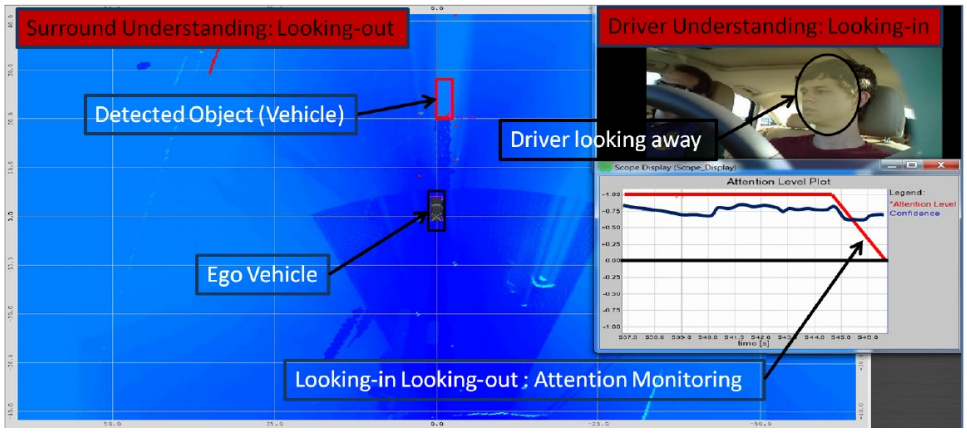
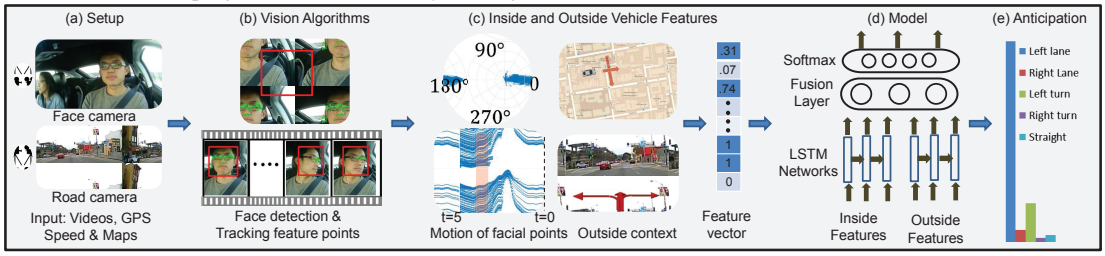


Figure 1.10: Intent detection using turn signal analysis [5]. First, vehicles are detected and tracked using a Mixture-of-Experts model and a Kanade-Lucas-Tomasi tracker. Consequently, light spots are detected, and classification of events is performed with an AdaBoost classifier over frequency-domain 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 1.11: Emerging research topics in integrative frameworks for on-road activity 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. 1.7 depicts some of the relevant research tasks), this domain of human analysis for intelligent vehicles is expected to have high research and commercial activity.

1.1.3 Looking at Humans in Surround Vehicles

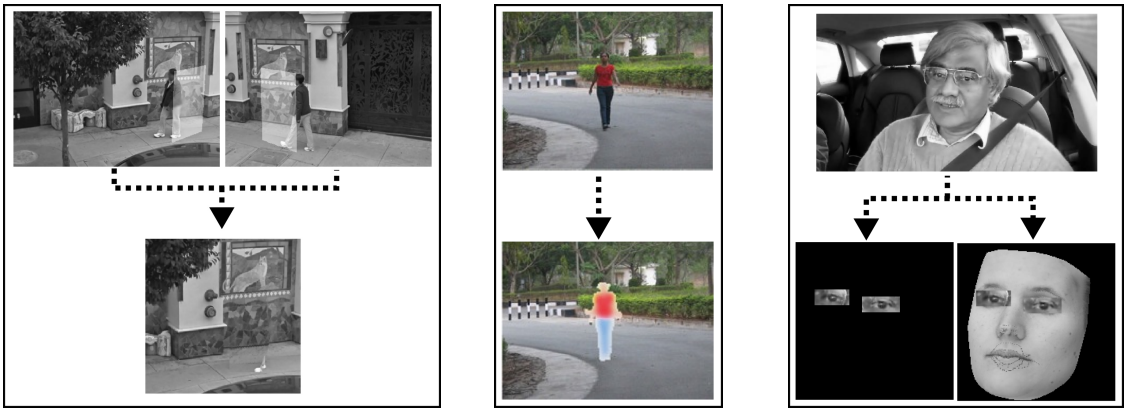
Understanding intent of drivers in surround vehicles, a task continuously performed by human drivers, is also useful for machine drivers. The research tasks are therefore shared across the three domains of humans in intelligent vehicles. When looking at humans in surround vehicles, vision-based algorithms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

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. 1.10). 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.

1.1.4 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. 1.9.

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. 1.11). Jain *et al.* [116] employs multi-modal Long Short-Term Memory



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

Figure 1.12: 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.

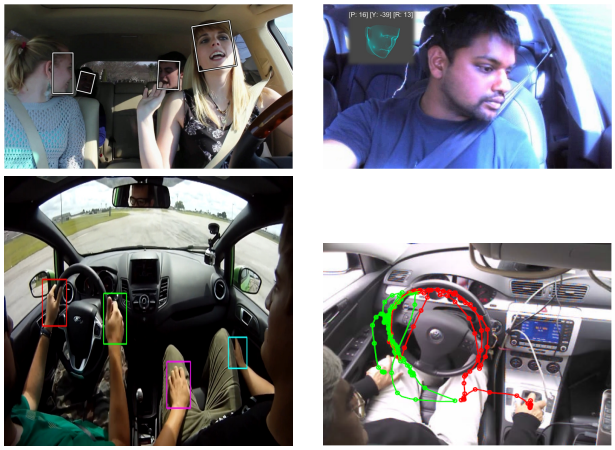
Table 1.3: Overview of selected publicly available naturalistic datasets from a mobile vehicle platform.

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.

networks for maneuver anticipation.

1.2 Naturalistic Datasets and Analysis Tools

The survey of related research studies in Section 2.2 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) 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 1.13: Example images from publicly available datasets (Table 1.3) for analysis of humans inside and outside of the vehicle.

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.

1.2.1 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.

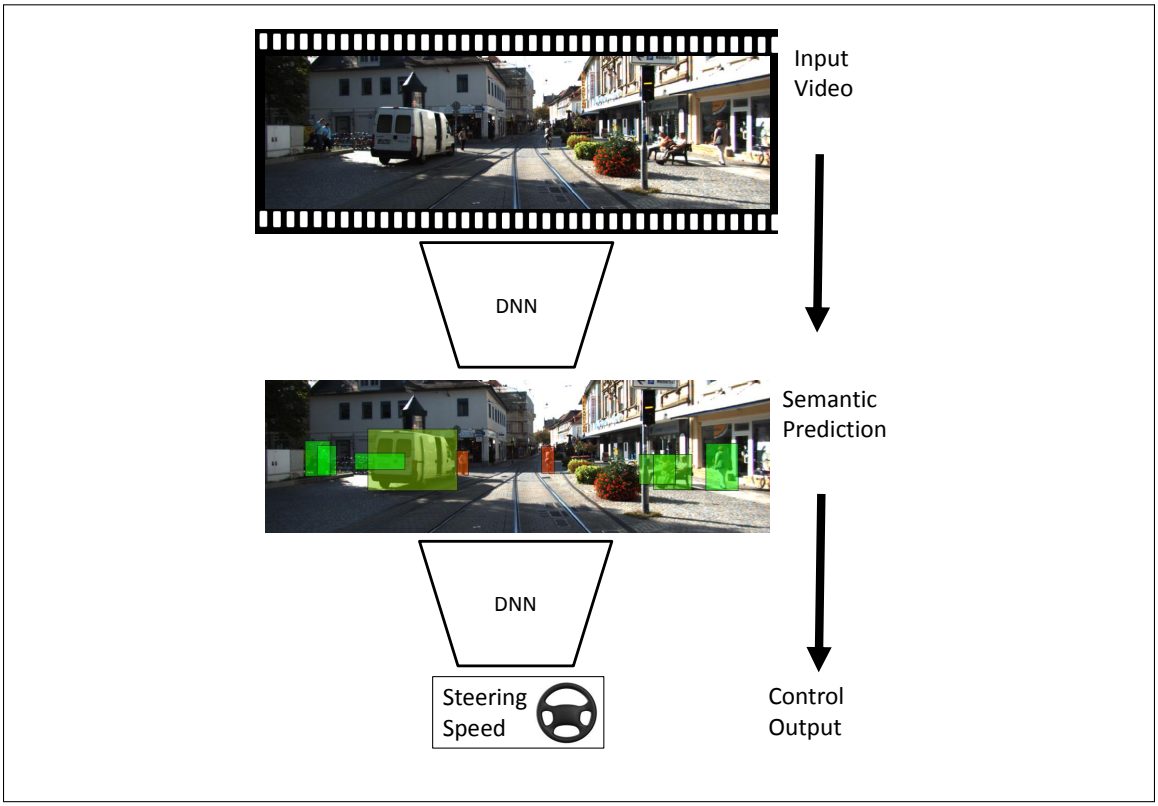


Figure 1.14: Example video-to-control policy pipeline (mediated-semantic perception [6, 7]) with deep networks (DNN), where initial prediction of semantic scene elements is followed by a control policy algorithm.

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

1.2.2 Naturalistic Driving Datasets

Table 1.3 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. 1.13, 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. 1.13, 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.

1.3 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. 1.14) 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.

Chapter 2

Multi-cue Driver Behavior Modeling

We study techniques for monitoring and understanding real-world human activities, in particular of drivers, from distributed vision sensors. Real-time and early prediction of maneuvers is emphasized, specifically overtake and brake events. Study 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.

2.1 Introduction

Distributed camera and sensor networks are needed for studying and monitoring agent activities in many domains of application [176]. Algorithms that reason over the multiple perspectives and fuse information have been developed with applications to outdoor or indoor surveillance [177]. 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 [178]. 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

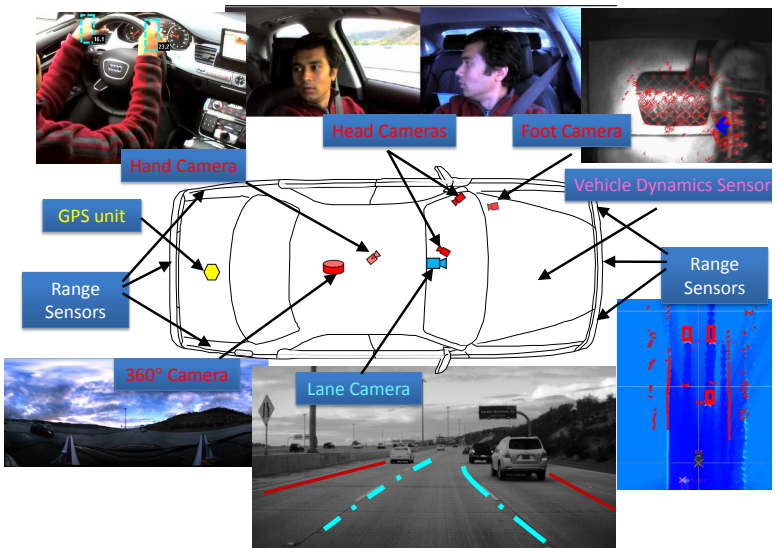


Figure 2.1: Distributed, synchronized network of sensors used in this study. A holistic representation of the scene allows for prediction of driver maneuvers. Knowledge of events a few seconds before occurrence and the development of effective driver assistance systems could make roads safer and save lives.

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 [179]. Most of these 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 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.

Table 2.1: 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 (front looking camera).
†: Defined lane-change at lane crossing. ‡: Explicitly models pre-intent cues.

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 [180–182] 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. 2.1). 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 2.2, 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 2.8 and 2.4, respectively. The different signals and feature extraction modules are detailed in Section 2.5. Two temporal modeling approaches for maneuver representation and fusion will be discussed in Section 2.6, and the experimental evaluation (Section 2.8) demonstrates analysis of different cues and modeling techniques in terms of their predictive power.

2.2 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 [183] and maps [184], vehicle dynamics [59], and surround sensors [185] 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. A thorough recent review can be found in [186].

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.

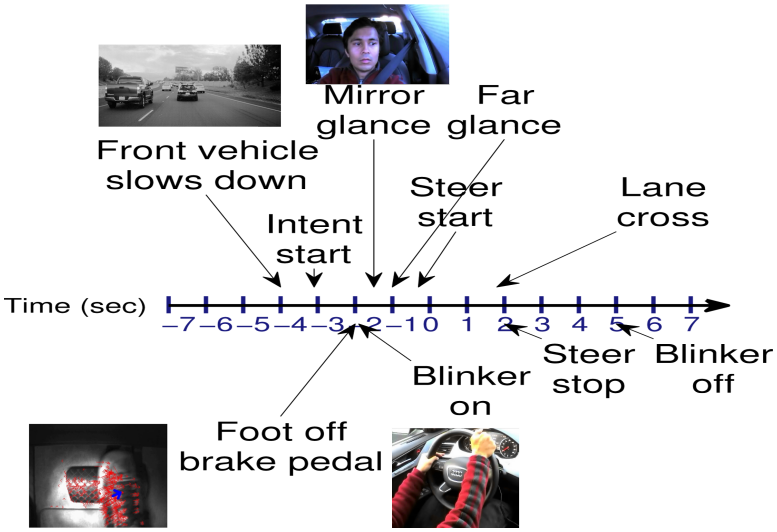


Figure 2.2: Timeline of an example overtake maneuver. Our algorithm analyzes cues for intent prediction, intent inference, and trajectory estimation towards the end of the maneuver.

Several recent simulator studies have been performed using a variety of cues for intent inference. In [187], 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 [188] 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 [189] for emergency brake application prediction. Table 2.1 lists related research based on the maneuver studied, the learning approach, and the cues used for comparison with this work. Table 2.1 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 2.1. 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 2.1). 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. 2.2).

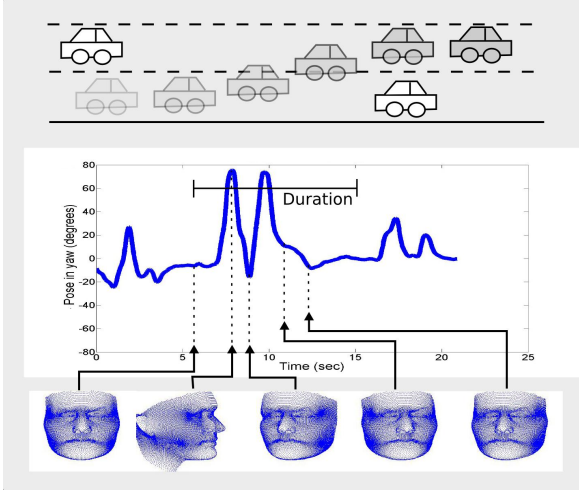


Figure 2.3: An example overtake maneuver with head dynamics. An overtake event may be defined in multiple ways, discussed in Section 2.3. Head-cues are important for detecting intent. See also Fig. 2.4.

2.3 Event Definition

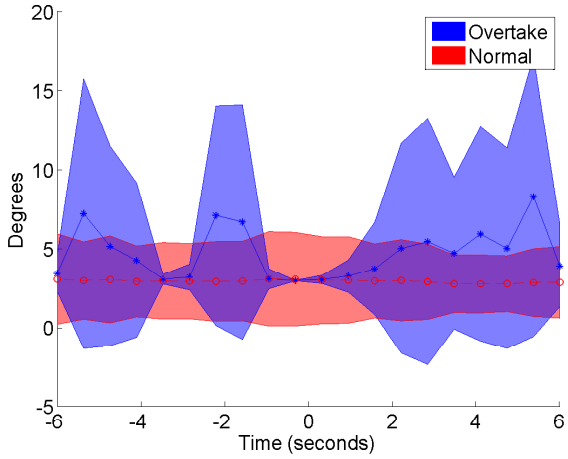
Commonly, a lane change event or an overtake event (which includes a lane-change) are defined to begin at the lane marker crossing. On the contrary, in this work the beginning of an overtake event is defined earlier when the lateral motion started. We note that there are additional ways to define a maneuver such as an overtake or a lane-change (in [182], and the our definition is significantly earlier than those in several related studies in Table 2.1. For instance, techniques focusing on trajectory-based prediction define lane-change at the lane marker crossing.

Nonetheless, as shown in (Fig. 2.2), 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.

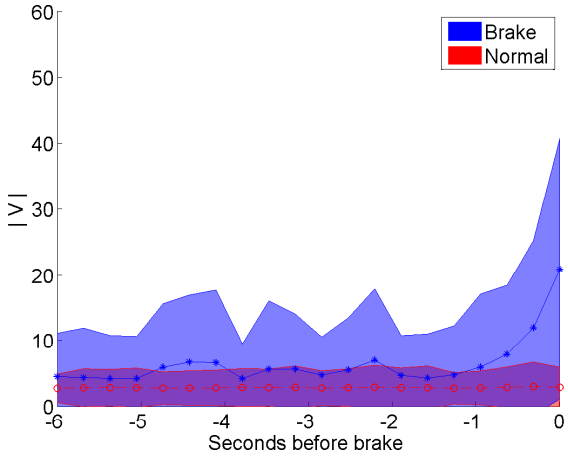
In addition to studying overtake maneuvers, which involve lateral control of the vehicle, we study a longitudinal control maneuver which is also essential in preventing accidents and monitoring for driver assistance. These are events where the driver chose to brake due to a situational need. While brakes 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.

2.4 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



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



(b) Foot velocity during a braking event.

Figure 2.4: Mean and standard deviation of signals from the head pose and foot motion tracking modules during the two maneuvers studied in this work. Time 0 for overtake is the moment when the ego-vehicle crossed the lane marking. The brake pedal is used to define a braking event.

automotive testbed is outfitted with extensive auxiliary sensing for the research and development of advanced driver assistance technologies. Fig. 2.1 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 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.

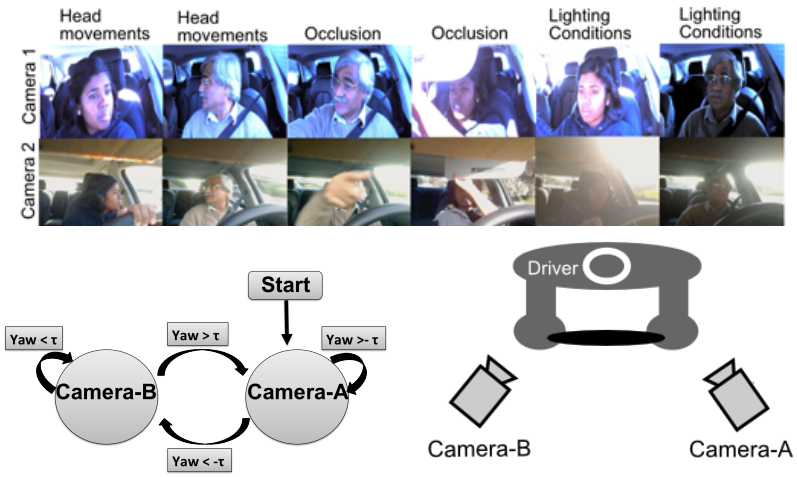


Figure 2.5: A two camera system overcomes challenges in head pose estimation and allow for continuous tracking even under large head movements, varying illumination conditions, and occlusion.

- 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 measurements 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.

2.5 Maneuver Representation

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

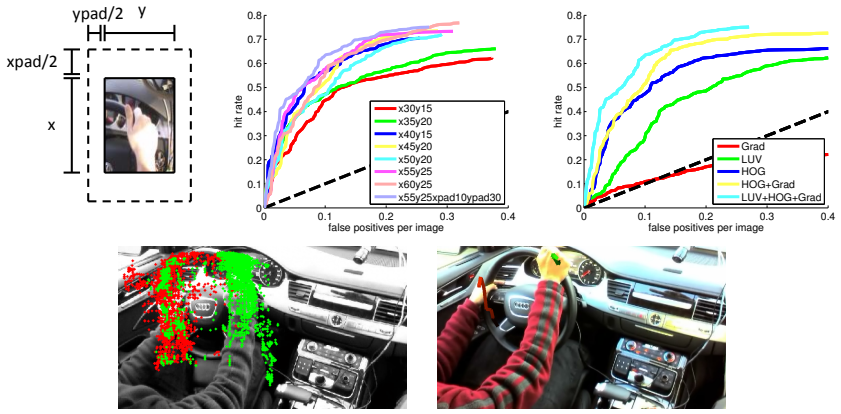


Figure 2.6: Top: Hand detection results with varying patch size and features. Bottom: Scatter plot of left (in red) and right (in green) hand detection for the entire drive. A hand trajectory of reaching towards the signal before an overtake is shown (brighter is later in time).

2.5.1 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. On the other hand, the foot and hand signals occur with the driver intention to operate a controller in the vehicle.

Multiple cameras for human activity analysis [190] and face analysis [191] have been shown to reduce occlusion-related failures. In [192], 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 [193], and their corresponding points on a 3D mean face model [194]. 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. 2.5. In order to handle camera selection and hand-off, multiple techniques have been proposed in literature (a survey of different methods can be found at [176]). 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 $-\tau$. In our current implementation we let $\tau = 10^\circ$. 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

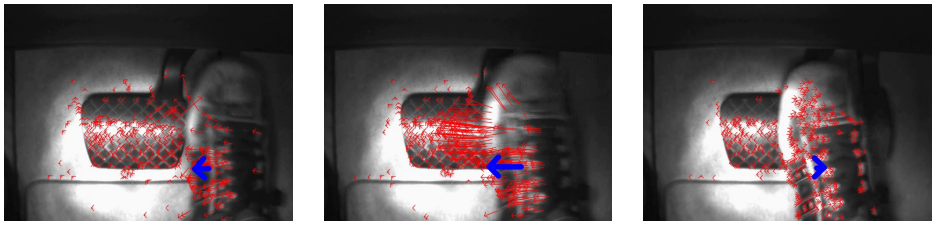


Figure 2.7: Foot tracking using iterative pyramidal Lucas-Kanade optical flow. Majority vote produces location and velocity.

is employed.

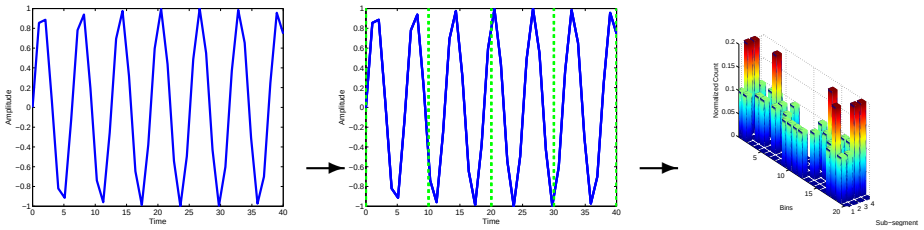
Hand: The hand signal will be used to study preparatory motions before a maneuver is performed. Below, we specify the hand detection and tracking module. Hand detection is a difficult problem in computer vision, due to the hand's tendency to occlude itself, deform, and rotate, producing a large variability in its appearance [127]. We use integral channel features [195] which are fast to extract. Specifically, for each patch extracted from a color image, gradient channels (normalized gradient channels at six orientations and three gradient magnitude channels) and color channels (CIE-LUV color channels were experimentally validated to work best compared to RGB and HSV) are extracted. 2438 instances of hands were annotated, and an AdaBoost classifier with decision trees as the weak classifiers is used for learning [196]. 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. 2.6). 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 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, running at 30Hz) based motion cue is employed to determine the location and magnitude of relatively significant motions in the pedal region. 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 [197] 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



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

Figure 2.8: Two features used in this work: raw trajectory features outputted by the detection and tracking, and histograms of sub-segments.

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. 2.1) 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 a offline 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 minimal.

2.5.2 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,

$$F_t = (f_{t-L+1}, \ldots, f_t) \tag{2.1}$$

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. 2.8). 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.

2.6 Temporal Modeling

A model for the signals extracted by the modules in Section 2.5 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) [198] and the other using Multiple Kernel Learning (MKL) [199]. 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. 2.1, $\mathbf{x} = \{F_t^{(1)}, \ldots, F_t^{(s)}\}$, 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) [198]. 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) [200] 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) [198], 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 $\mathbf{h}$ gives

$$P(\mathbf{y}|\mathbf{x};\Lambda) = \sum_{\mathbf{h}} P(\mathbf{y}|\mathbf{h},\mathbf{x},\Lambda)P(\mathbf{h}|\mathbf{x},\Lambda) = \sum_{\mathbf{h}:\forall h_i \in H_{y_i}} P(\mathbf{h}|\mathbf{x};\Lambda) \tag{2.2}$$

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

$$P(\mathbf{h}|\mathbf{x};\Lambda) = \frac{\exp\left(\sum_k \Lambda_k \cdot \mathbf{T}_k(\mathbf{h},\mathbf{x})\right)}{\sum_{\mathbf{h}} \exp\left(\sum_k \Lambda_k \cdot \mathbf{T}_k(\mathbf{h},\mathbf{x})\right)} \tag{2.3}$$

We follow [198], where the function $\mathbf{T}_k$ is defined as a sum of state (vertex) or binary transition (edge) feature functions,

$$\mathbf{T}_k(\mathbf{h},\mathbf{x}) = \sum_{i=1}^{m} l_k(h_{i-1}, h_i, \mathbf{x}, i) \tag{2.4}$$

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

function,

$$L(\Lambda) = \sum_{i}^{n} \log P(\mathbf{y}_i|\mathbf{x}_i, \Lambda) - \frac{1}{2\sigma^2}||\Lambda||^2 \tag{2.5}$$

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. 2.2). 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 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. 2.8). 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 2.5. 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.

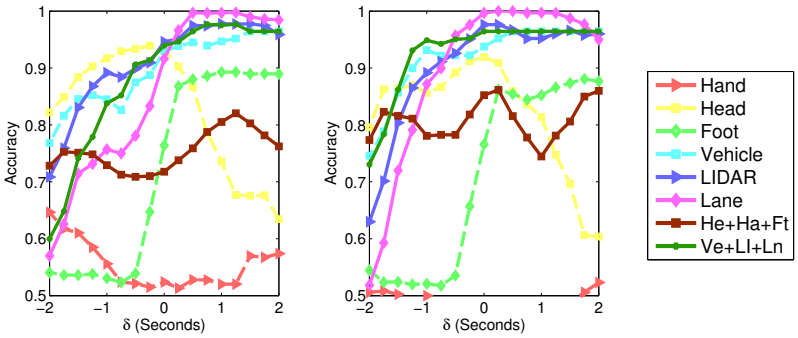
The kernels are combined by learning a probability distribution $\mathbf{p} = (p^1, \ldots, p^s)$, with $p \in \mathbb{R}_+$ and $\mathbf{p}^T\mathbf{1} = 1$, such that the combination of kernel matrices,

$$\mathbf{K}(\mathbf{p}) = \sum_{l=1}^{s} p^l \mathbf{K}^{c_l} \tag{2.6}$$

is optimal. In this work, the weights are learned using stochastic approximation [199]. LIBSVM [201] is used as the final classifier. The histogram features were shown to work well with the MKL features, performing better than simply using the raw temporal signal features.

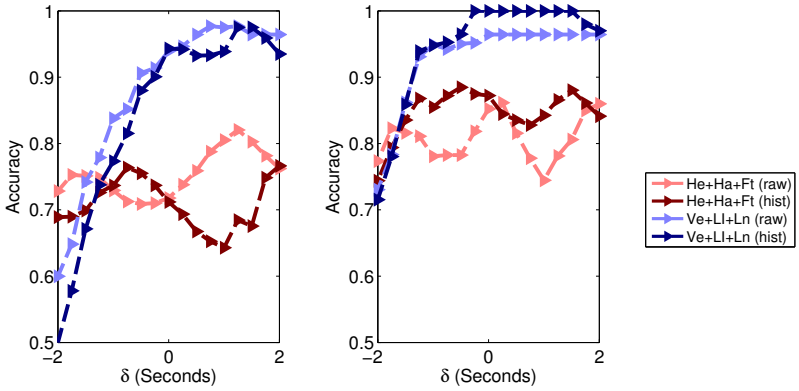
2.7 Experimental Setup

Several experiments are conducted in order to test the proposed framework for recognition of intent and prediction of maneuvers. As mentioned in Section 2.3, 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.



(a) LDCRF (b) MKL

Figure 2.9: Classification and prediction of overtake-late/brake (Experiment 1a) maneuvers using **raw trajectory features**. He+Ha+F 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.



(a) LDCRF (b) MKL

Figure 2.10: Comparison of the two temporal features (see Section 2.5.2) 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, while no benefit is shown to the LDCRF.

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 2.8).

- **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 attempt to answer the question, 'how and why did a driver decide to perform a brake?',

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

2.8 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. 2.9 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. 2.9 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. 2.10, 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. 2.11 for experiments 2 and 3, where the entire set of signals described in Section 2.5 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. 2.11(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 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

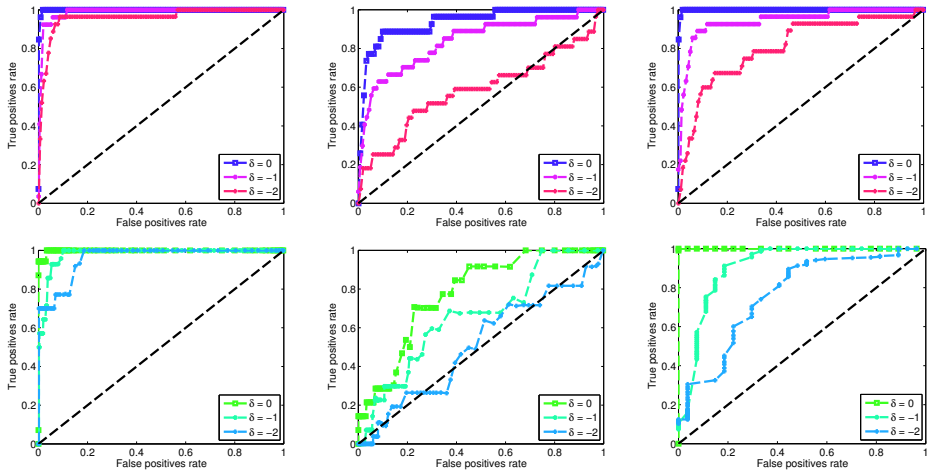


Figure 2.11: Measuring prediction by varying the time in seconds before an event, δ. **Top**: MKL results. **Bottom**: LDCRF results. (a) Experiment 2a: Overtake-late vs. normal (b) Experiment 2b: Overtake-early vs. normal (c) Experiment 3: Brake vs. normal. Note how prediction of overtake-early events, which occur seconds before the beginning of an overtake-late events, is more difficult.

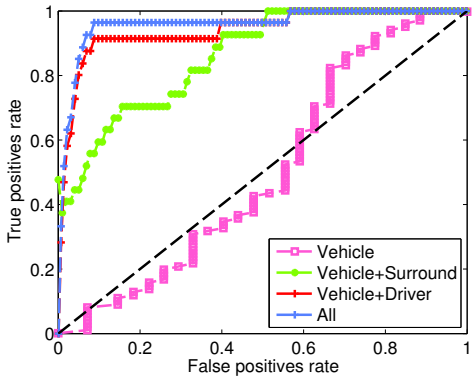
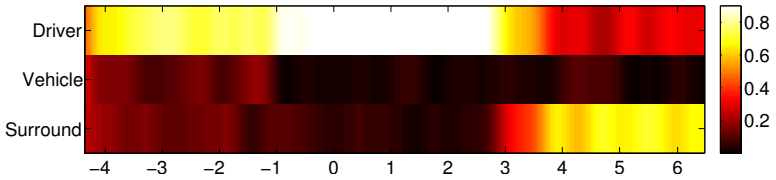


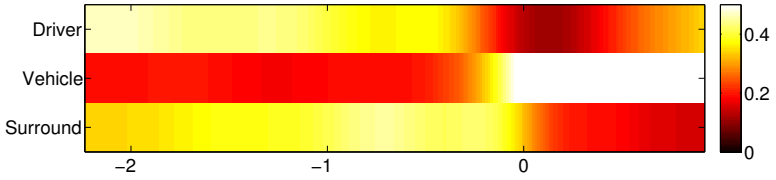
Figure 2.12: For a fixed prediction time of $\delta = -2$ seconds, we show the effects of appending cues to the vehicle dynamics under overtake-late/normal (experiment 2a). The surround cues utilize lidar, lane, and visual data. Driver cues include the hand, head, and foot signals.

depicted in Fig. 2.12. Although both systems provide an advantage, most gains for early prediction come for prediction by observing driver related cues.

Fig. 2.13 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. 2.13 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 motion begins (t=0) in Fig. 2.13(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. 2.13(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



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



(b) Brake/normal (Experiment 3)

Figure 2.13: 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.

with vehicle dynamics occurs around the time of the pedal press.

2.9 Chapter Concluding Remarks

In this work, a surveillance application of driver assistance was studied. Automotive driver assistance systems 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.

Chapter 3

Spatio-Temporal, Human-Centric Scene Understanding

This chapter provides a unified analysis of object recognition, behavior modeling, and human perception in a combining 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.

3.1 Introduction

There is a great need for smarter and safer vehicles [202, 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 agents surrounding a vehicle, is an essential component in the development of such systems [203–206]. 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-

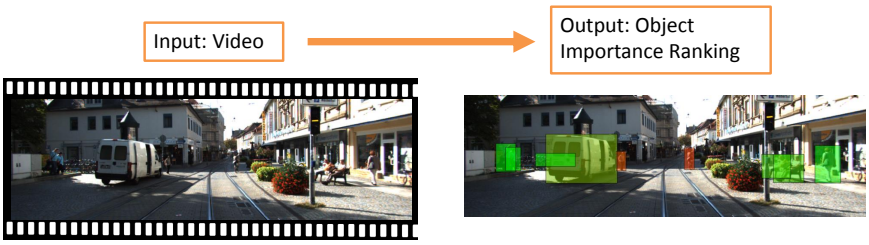


Figure 3.1: What makes an object salient in the spatio-temporal context of driving? Given a video, this work aims to rank agents in the surrounding scene by relevance to the driving task. Furthermore, the notion of importance defined in this work allows a novel evaluation of vision algorithms and their error types. The importance score (averaged over subjects' annotations) for each object are shown, colored from high to moderate to low.

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. 3.1. 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. 3.1, 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.

3.1.1 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 [207], but may also be useful to other application domains in computer vision requiring spatio-temporal analysis and human perception modeling, including saliency modeling [208, 209], robotics [210], and ego-centric vision [211].

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.

3.2 Motivation and Related Research Studies

Importance analysis: Importance ranking essentially involves modeling context. Capturing spatial image context has been heavily studied [212, 213]. Berg *et al.* [214] 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 [215], understanding of what an agent can do, intends to do, or how multiple agents may interact [216]. Lee *et al.* [217] 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.* [218] 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 [208, 219], 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 [220, 221]). 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 tremendous progress in recent years, most notably on the ILSVRC [222–224], PASCAL [225, 161], Caltech [195], 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.

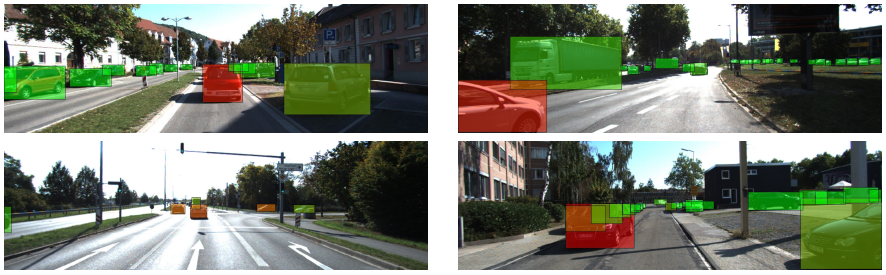


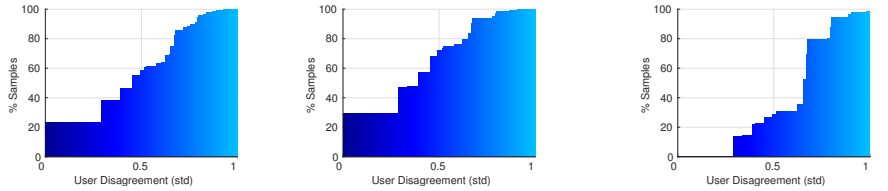
Figure 3.2: This study is motivated by the fact that not all objects are equally relevant to the driving task. As shown in example frames from the dataset with overlaid object-level importance score (averaged over subjects), drivers' attention to road occupants varies based on task-related, scene-specific, and object-level cues.

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 [226–229] where importance is measured differently from in the driving domain. We experimentally demonstrate the impact of such bias in evaluation on KITTI (Section 3.5). 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. 3.2) 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 3.5 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 drivers do not generally pay attention to all objects in the scene (Fig. 3.2), 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.

3.3 Importance Annotation Dataset

The KITTI dataset [141, 207] 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,

Figure 3.3: 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.



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

Figure 3.4: 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. 3.3 for discussion).

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. 3.3. 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. 3.5. Consistency analysis (Fig. 3.4) of the annotators output demonstrates that many instances in the 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

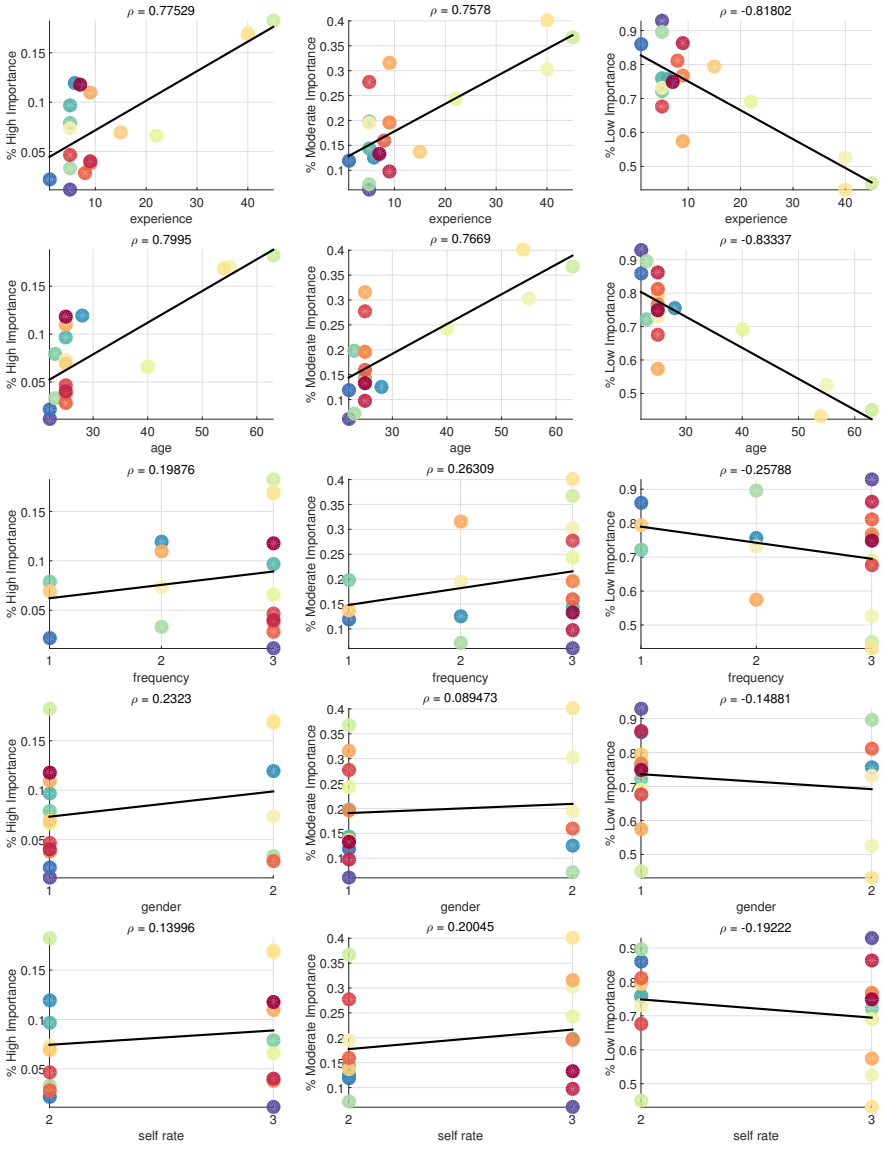


Figure 3.5: Relationship between importance level (grouped by columns) and subject personal information (grouped by rows). Each subject has been assigned a unique color, and is represented in each figure by a dot. From top row: (1) driving experience in years, (2) age in years, (3) frequency of driving, 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.

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

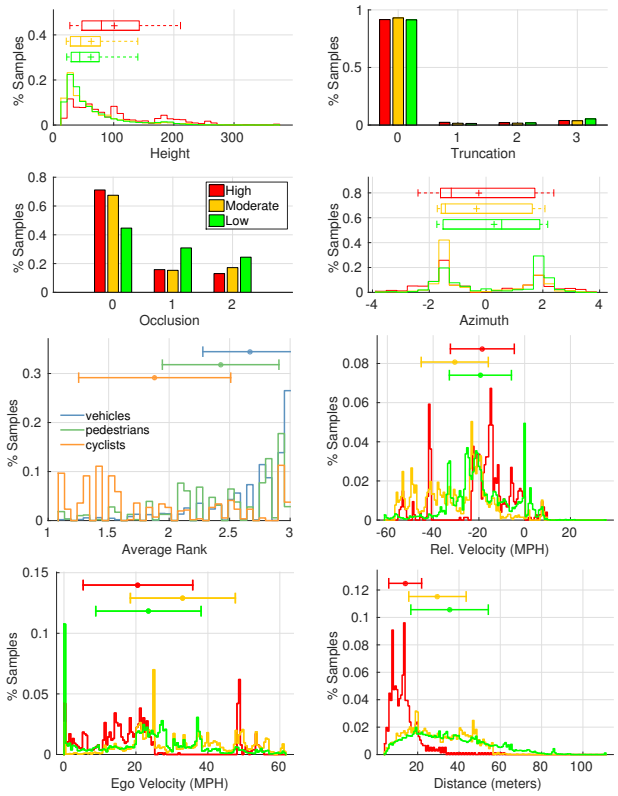


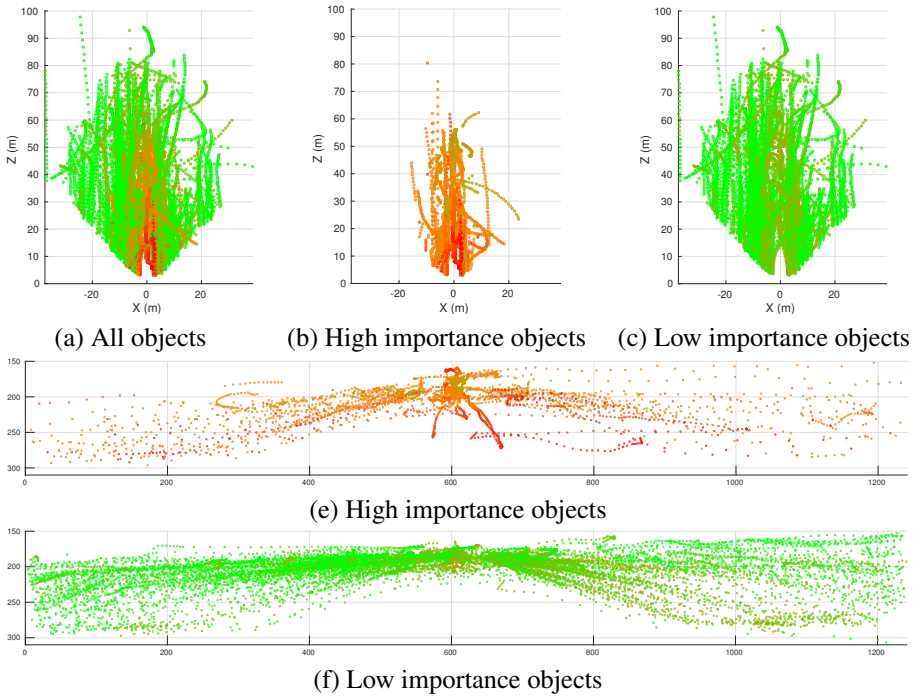
Figure 3.6: Object statistics corresponding to three classes of object importance in the dataset.

of a barrier in traffic, head orientation cues for pedestrians (also studied in [230–232]), 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. 3.6. 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. 3.7(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. 3.7(e) demonstrates the distribution of the position in the image plane for high importance objects.

3.4 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 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.



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

(e) High importance objects

(f) Low importance objects

Figure 3.7: 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 high to moderate to low importance.

3.4.1 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,

$$M_{attributes}(s) = \boldsymbol{w}_{c,2D-obj}^{T}\phi_{2D-obj}(s) + \boldsymbol{w}_{c,3D-obj}^{T}\phi_{3D-obj}(s) + \boldsymbol{w}_{c,ego}^{T}\phi_{ego}(s) + \boldsymbol{w}_{c,temporal}^{T}\phi_{temporal}(s) \tag{3.1}$$

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. 3.7, 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 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. 3.3. 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 [215].

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.

3.4.2 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

$$M_{visual}(s) = \boldsymbol{w}_{c,obj}^T \phi_{obj}(s) + \boldsymbol{w}_{c,spatial}^T \phi_{spatial}(s) + \boldsymbol{w}_{c,temporal}^T \phi_{temporal}(s) \tag{3.2}$$

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) [233] convolutional network. The network was pre-trained on the ImageNet dataset [223] and fine-tuned on KITTI using Caffe [234].

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.

3.5 Importance Metrics for Object Detection

As described in Section 3.2, 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 3.4.1 and 3.4.2, 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) [235], 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 [236], 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

$$L(p, u, \gamma, t^u, v) = L_{cls}^{IG}(p, u, \gamma) + \lambda_{loc}[u \geq 1]L_{loc}(t^u, v) \tag{3.3}$$

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

$$\alpha_\gamma = \begin{cases} \lambda & \gamma \leq 2.25 \\ 1/\lambda & \text{otherwise} \end{cases} \tag{3.4}$$

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 [236]. 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.

3.6 Experimental Evaluation

3.6.1 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 3.1. 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

Table 3.1: 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**

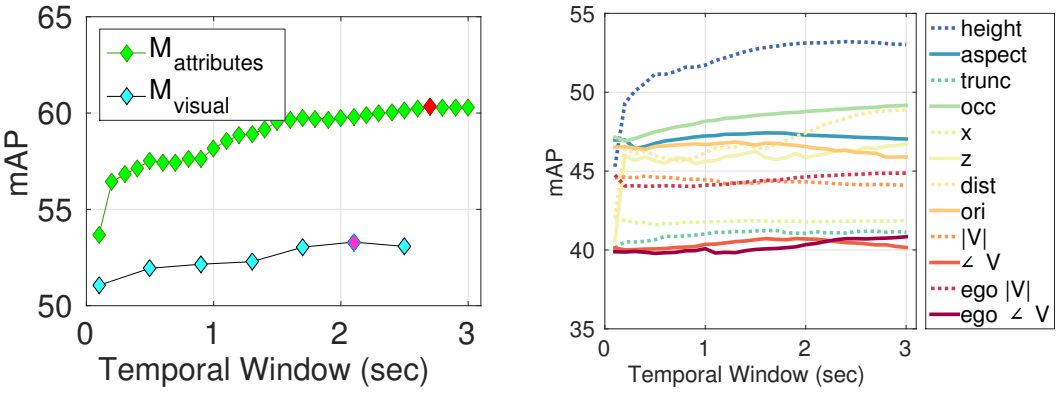


Figure 3.8: Cue analysis with the importance models. (a) Classification accuracy when varying the time window used for computing $\phi_{temporal}$ in both models. (b) Classification accuracy with each of the attributes in $M_{attributes}$ with an increasing temporal window used for a temporal feature extraction.

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 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 3.1 shows the performance of the attributes-based model. We note that for the experiments in Table 3.1, 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. 3.8, 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. 3.9 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 category

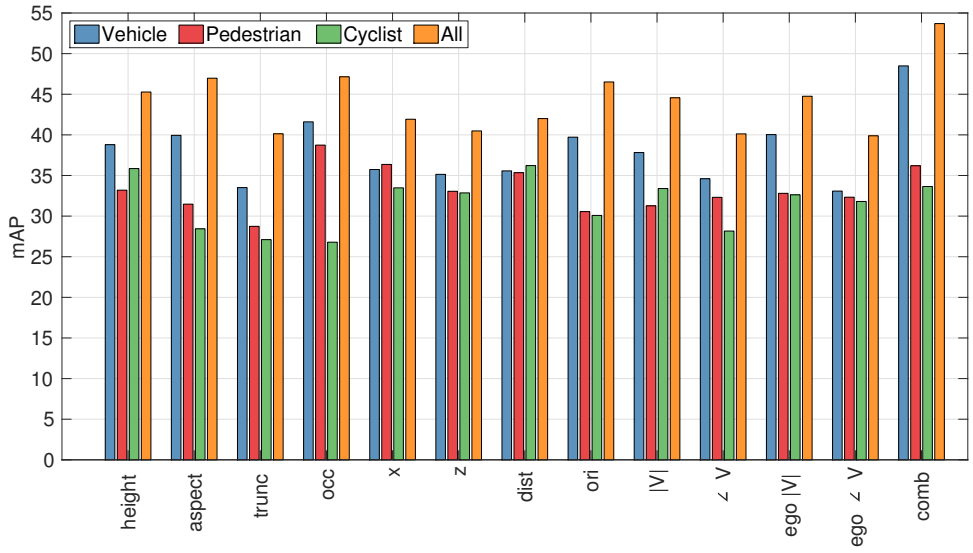


Figure 3.9: Object importance classification results using each attribute in $M_{attributes}$ separately, as well as with a combination of all attributes ('comb'). Results are shown for training and evaluation on each object class separately, as well as in an object class agnostic manner ('All'). No temporal feature extraction is used in these experiments.

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. 3.8(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. 3.8(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. 3.10 shows the PR curves used to compute the final performance summary in Table 3.1. Fig. 3.10 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 dataset could resolve such issues. Furthermore, additional insights may be gained by subject-specific modeling and evaluation, which is left for future work.

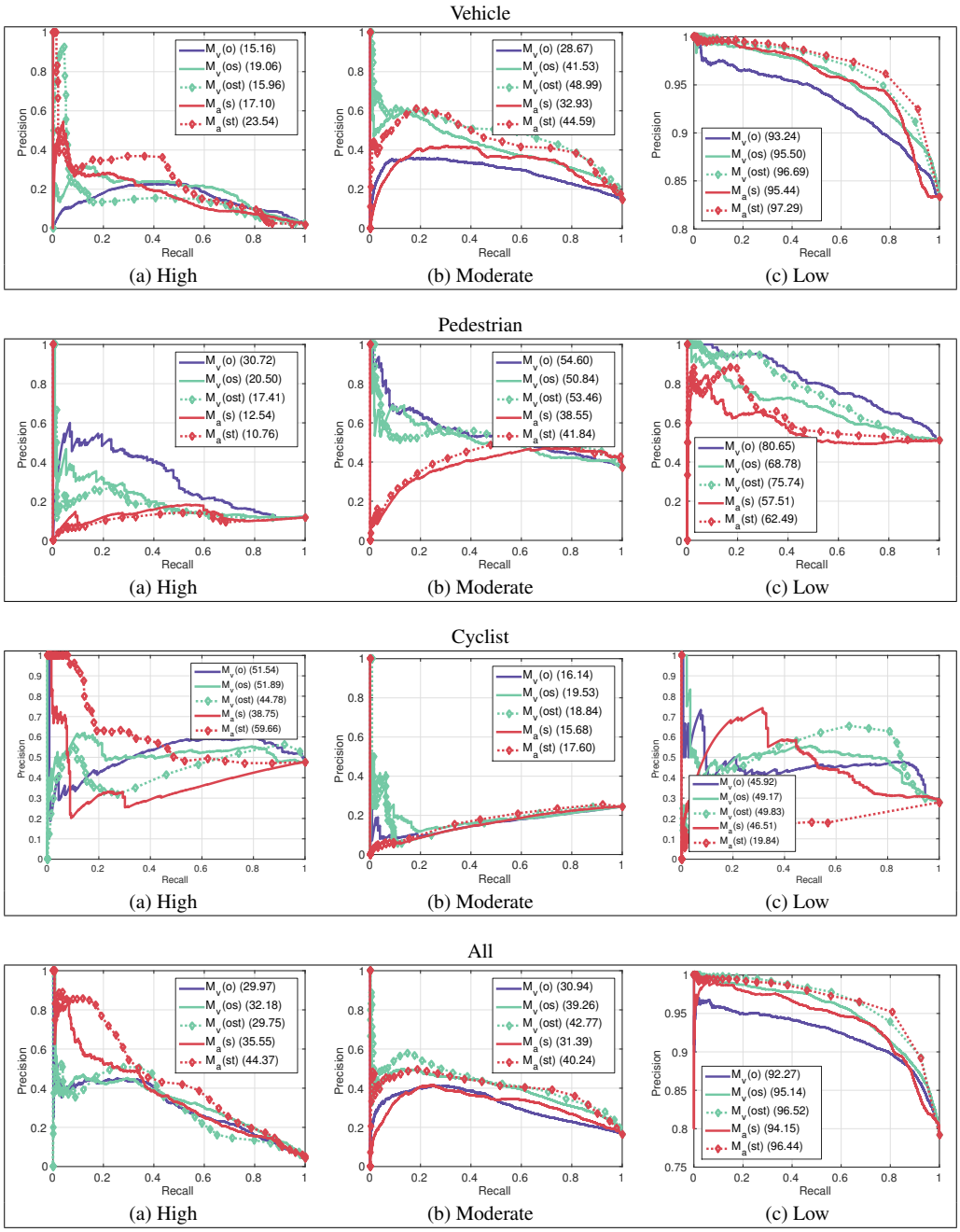


Figure 3.10: For each object class (rows) and object importance level (columns), we show performance precision-recall curves when employing different models and cue types. For the attributes model (M_a), performance without and with temporal features is shown as 's' and 'st', respectively. Similarly, 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.

Evaluation of M_{visual}: Table 3.1 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

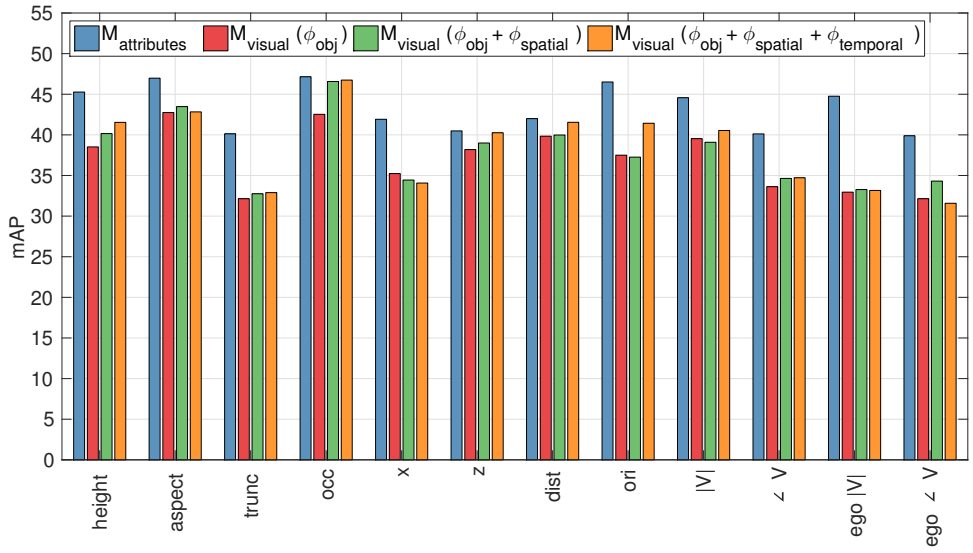


Figure 3.11: Regressing each attribute using various feature combinations in M_{visual} and consequently using the attribute for importance class classification allows for explicit analysis of the limitations of M_{visual}.

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. 3.8, 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 3.1 motivate further study of models suitable for capturing spatio-temporal visual cues [237–239, 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. 3.11. 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 experiment 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. 3.11 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.

Table 3.2: 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

3.6.2 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 3.5. Results are shown in Table 3.2 for both the ZF [240] and the VGG [233] network architectures. Table 3.2 depicts the complementary relationship between the proposed set of importance metrics and traditional test settings (defined in Section 3.3). 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 3.2 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 [241], 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 3.2, this is not the case with importance classes.

3.7 Chapter Concluding Remarks

This work studies object recognition 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 [238, 58, 242, 202]. Further investigation of the cost-sensitive training procedure [243, 244, 239] 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 [245, 246, 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.

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," *Under Review*, 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 Intelligent Vehicles Symposium*, 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, pp. 1–10, 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. Bhanu, C. V. Ravishankar, A. K. Roy-Chowdhury, H. Aghajan, and D. Terzopoulos, Eds., *Distributed Video Sensor Networks*. Springer, 2011.

[177] 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.

[178] "2012 motor vehicle crashes: overview," National Highway Traffic Safety Administration, Washington, D.C., Tech. Rep. DOT HS 811 856, 2013.

[179] 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.

[180] 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.

[181] 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.

[182] "A comprehensive examination of naturalistic lane-changes," National Highway Traffic Safety Administration, Washington, D.C., Tech. Rep. DOT HS 809 702, 2004.

[183] R. Simmons, B. Browning, Y. Zhang, and V. Sadekar, "Learning to predict driver route and destination intent," in *IEEE Conf. Intelligent Transportation Systems*, 2006.

[184] S. Lefèvre, C. Laugier, and J. Ibañez-Guzmán, "Exploiting map information for driver intention estimation at road intersections," in *IEEE Intelligent Vehicles Symposium*, 2011.

[185] M. Ortiz, F. Kummert, and J. Schmudderich, "Prediction of driver behavior on a limited sensory setting," in *IEEE Conf. Intelligent Transportation Systems*, 2012.

[186] A. Doshi and M. M. Trivedi, "Tactical driver behavior prediction and intent inference: A review," in *IEEE Conf. Intelligent Transportation Systems*, 2011.

[187] 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.

[188] 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.

[189] 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.

[190] 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.

[191] 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.

[192] 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.

[193] 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.

[194] 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.

[195] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2014.

[196] C. Zhang and P. A. Viola, "Multiple-instance pruning for learning efficient cascade detectors," in *Advances in Neural Information Processing Systems*, 2007.

[197] S. Martin, C. Tran, and M. Trivedi, "Optical flow based head movement and gesture analyzer (OHMeGA)," in *IEEE Intl. Conf. Pattern Recognition*, 2012.

[198] 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.

[199] 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.

[200] 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.

[201] 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.

[202] 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.

[203] A. Doshi and M. M. Trivedi, "Tactical driver behavior prediction and intent inference: A review," in *IEEE Conf. Intell. Transp. Syst.*, 2011.

[204] 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.

[205] 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.

[206] E. Ohn-Bar and M. M. Trivedi, "What makes an on-road object important?" in *Intl. Conf. Pattern Recognition*, 2016.

[207] 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.

[208] A. Borji, Dicky, N. Sihite, and L. Itti, "Probabilistic learning of task-specific visual attention," in *CVPR*, 2012.

[209] A. Doshi and M. M. Trivedi, "Attention estimation by simultaneous observation of viewer and view," in *CVPRW*, 2010.

[210] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *HRI*, 2013.

[211] G. Rogez, J. S. Supancic, and D. Ramanan, "Understanding everyday hands in action from RGB-D images," in *ICCV*, 2015.

[212] 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.

[213] 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.

[214] 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.

[215] H. Pirsiavash, C. Vondrick, and A. Torralba, "Assessing the quality of actions," in *ECCV*, 2014.

[216] W. Chen, C. Xiong, R. Xu, and J. J. Corso, "Actionness ranking with lattice conditional ordinal random fields," in *CVPR*, 2014.

[217] Y. J. Lee and K. Grauman, "Predicting important objects for egocentric video summarization," *IJCV*, vol. 114, no. 1, pp. 38–55, 2015.

[218] C. S. Mathialagan, A. C. Gallagher, and D. Batra, "Vip: Finding important people in images," in *CVPR*, 2015.

[219] N. Pugeault and R. Bowden, "Learning pre-attentive driving behaviour from holistic visual features," in *ECCV*, 2010.

[220] D. M. Y. Zhu, Y. Tian and P. Dollár, "Semantic amodal segmentation," *CoRR*, vol. abs/1509.01329, 2015.

[221] 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.

[222] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[223] 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.

[224] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *PAMI*, vol. 35, no. 8, pp. 1915–1929, 2013.

[225] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *IJCV*, 2010.

[226] X. Chen and A. Gupta, "Webly supervised learning of convolutional networks," in *ICCV*, 2015.

[227] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR*, 2011.

[228] R. N. Rajaram, E. Ohn-Bar, and M. M. Trivedi, "An exploration of why and when pedestrian detection fails," in *ITSC*, 2015.

[229] ——, "Looking at pedestrians at different scales: A multiresolution approach and evaluations," *TITS*, 2016.

[230] 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.

[231] J. Kooij, N. Schneider, F. Flohr, and D. Gavrila, "Context-based pedestrian path prediction," in *ECCV*, 2014.

[232] T. Gandhi and M. M. Trivedi, "Pedestrian protection systems: Issues, survey, and challenges," *IEEE Trans. Intelligent Transportation Systems*, vol. 8, pp. 413–, 2007.

[233] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[234] 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.

[235] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *AAAI*, 2015.

[236] R. Girshick, "Fast r-cnn," in *Intl. Conf. on Computer Vision*, 2015.

[237] 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.

[238] 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.

[239] 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.

[240] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, 2014.

[241] B. Pepik, R. Benenson, T. Ritschel, and B. Schiele, "What is holding back convnets for detection?" in *GCPR*, 2015.

[242] 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.

[243] 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.

[244] O. Beijbom, M. Saberian, D. Kriegman, and N. Vasconcelos, "Guess-averse loss functions for cost-sensitive multiclass boosting," in *ICML*, 2014.

[245] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *PAMI*, vol. 31, no. 12, pp. 2179–2195, 2009.

[246] 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.