The Role of Context in Understanding and Predicting Pedestrian Behavior in Urban Traffic Scenes

Amir Rasouli

A Dissertation Submitted to
the Faculty of Graduate Studies
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

Graduate Program in
Electrical Engineering and Computer Science
York University
Toronto, Ontario

May 2020

Abstract

Today, one of the major challenges faced by autonomous vehicles (AVs) is the ability to drive in urban environments. Such a task requires interactions between AVs and other road users, in particular pedestrians, to resolve various traffic ambiguities. To interact with pedestrians, AVs must be able to understand the objectives of pedestrians and predict their forthcoming actions.

In this dissertation, we investigate the role of context on understanding and predicting pedestrian behavior in urban traffic scenes. Towards this goal, we begin by explaining why behavior prediction is necessary for social interactions. Next, we conduct a meta-analysis of a large body of behavioral literature and identify the factors that potentially impact pedestrian behavior and how these factors are interconnected. We extend the past findings by conducting two behavioral studies of pedestrians. The first study shows that pedestrians often engage in different forms of communication, mainly implicit, with changes in their movement patterns and the frequency of communication varying depending on road structure, social factors, and scene dynamics. The second study identifies the diversity of pedestrian behavioral patterns at the time of crossing and how it is influenced by factors such as the road width, demographics, crosswalk delineation, and driver behavior.

As part of the behavioral studies, we collected two novel large-scale datasets of pedestrian crossing behaviors. Using the data, we empirically evaluate various state-of-the-art and classical pedestrian detection algorithms and show how diversifying training data in terms of visual properties, such as lighting conditions and pedestrian attributes, enhance the generalizability of such algorithms. Furthermore, we propose a novel pedestrian trajectory prediction algorithm that achieves state-of-the-art performance. We show that incorporating pedestrian intention to cross helps improve reasoning about future motion trajectories. In addition, we propose a novel pedestrian crossing action prediction algorithm and illustrate that by including contextual information, such as pedestrian appearance, pedestrian pose, and velocity, we can enhance the accuracy of crossing action prediction. We also show that by combining different modalities of contextual data in a hierarchical fashion better performance can be achieved compared to alternative approaches.

Acknowledgments

I want to thank my supervisor John Tsotsos for his continuous support and mentorship throughout my years of research at his lab. I also want to thank my supervisory committee members Michael Brown and Burton Ma as well as my examining committee members Andrew Blake, Regina Lee and Michael Jenkin for their time and valuable feedbacks.

I want to give my special thanks to my colleague, friend, and partner Iuliia Kotseruba who patiently supported me both emotionally and professionally through my years of graduate studies. None of what I have achieved today would be possible without Iuliia's unconditional love and friendship.

Contents

Abstract ii

Acknowledgments iii

Table of Contents vii

List of Tables viii

List of Figures x

1 Introduction 1

1.1 Autonomous Vehicles (AVs): A Brief History . . . 3

1.1.1 The Beginning . . . 4

1.1.2 Hitting the Road . . . 6

1.1.3 Achieving Autonomy . . . 8

1.1.4 Today's Autonomous Vehicles . . . 10

1.2 Where AVs Should Be and Where They Are Now . . . 11

1.3 The Importance of Prediction for Autonomous Systems . . . 12

1.4 Alternatives to Behavior Understanding and Prediction . . . 13

1.5 The Dissertation Outline . . . 14

1.6 Terminology . . . 15

1.7 Evaluation Methodology . . . 15

2 Social Interaction, Coordination and Behavior Prediction 17

2.1 Joint Attention in Human Interaction . . . 17

2.1.1 Early Studies of Joint Attention . . . 18

2.1.2 Joint Attention in Development of Social Cognition . . . 19

2.1.3 From Imitation to Coordination . . . 20

2.1.4 How Do Humans Coordinate? . . . 21

2.2 Why do We Predict Behavior? . . . 22
2.2.1 A Biological Perspective . . . 22
2.2.2 A Philosophical Perspective . . . 23
2.3 Summary . . . 23

3 Traffic Context and its Influence on Pedestrian Behavior 25
3.1 What do We Mean by Context? . . . 25
3.2 Methods of Studying Pedestrian Behavior . . . 26
3.3 Factors Influencing Pedestrian Behavior . . . 28
3.3.1 Classical Studies . . . 30
3.3.2 Studies in the Context of Autonomous Driving . . . 39
3.4 What Should be Done Next . . . 43

4 Pedestrian Communication in Traffic 46
4.1 Why is Communication Important in Traffic Context? . . . 46
4.1.1 Communication in Traditional Traffic . . . 46
4.1.2 Communicating with AVs . . . 46
4.2 Nonverbal Communication: How the Human Body Speaks to Us . . . 48
4.2.1 Studies of Nonverbal Communication . . . 49
4.2.2 Methods of Studying Nonverbal Communication . . . 50
4.2.3 Eye Contact: Establishing Connection . . . 51
4.2.4 Understanding Motives Through Bodily Movements . . . 52
4.3 Pedestrian Nonverbal Communication: An Empirical Study . . . 52
4.3.1 Joint Attention in Autonomous Driving (JAAD) Dataset . . . 53
4.3.2 Method . . . 54
4.3.3 Pedestrian Forms of Communication and Meaning . . . 55
4.3.4 How Often Pedestrians Communicate with Traffic . . . 56
4.4 Communication and Environmental Factors . . . 56
4.4.1 When and Where do Pedestrians Look . . . 56
4.4.2 Factors that Influence Communication . . . 58
4.5 Summary . . . 59

5 Understanding Pedestrian Crossing Behavior: An Empirical Study 60
5.1 Pedestrian Behavioral Data . . . 60
5.2 Pedestrian Behavior at the Time of Crossing . . . 61
5.3 Analyzing Pedestrian Crossing Behavior . . . 63
5.3.1 Attention Occurrence Prior to Crossing . . . 63

5.3.2 Crossing Action Post Attention Occurrence . . . 66
5.4 What Makes Understanding Pedestrian Actions Difficult . . . 67
5.4.1 Identifying Actions . . . 67
5.4.2 Identifying Relevant Elements . . . 68
5.4.3 Interpreting Behavior . . . 68
The Role of Context . . . 68
Identifying Relevant Actions . . . 69
5.4.4 Other Road Users . . . 70
5.5 Summary . . . 70

6 Detecting Pedestrians in Cluttered Traffic Scenes 72
6.1 A Literature Review on Pedestrian Detection . . . 73
6.1.1 Pedestrian Detection Algorithms . . . 73
6.1.2 Pedestrian Detection Datasets . . . 74
6.1.3 Data Properties and Pedestrian Detection . . . 74
6.2 A Pedestrian Detection and Attribute Dataset . . . 75
6.3 Experiment Setup . . . 77
6.3.1 Pedestrian Detection Algorithms . . . 77
6.3.2 Data . . . 77
6.3.3 Metrics . . . 78
6.4 Data Properties and Detection Accuracy . . . 78
6.4.1 Weather . . . 78
6.4.2 Pedestrian Attributes . . . 80
6.4.3 Generalizability Across Different Datasets . . . 82
6.5 Summary . . . 84

7 Understanding Pedestrians' Intentions and Their Role in Predicting Trajectories 86
7.1 A Literature Review of State-of-the-Art . . . 88
7.1.1 Vision-Based Trajectory Prediction . . . 88
7.1.2 Intention Estimation . . . 90
7.1.3 Datasets for Pedestrian Trajectory Prediction . . . 91
7.2 Pedestrian Intention Estimation (PIE) Dataset . . . 91
7.2.1 Annotations . . . 92
7.3 A Human Study on Predicting Pedestrian Crossing Intention . . . 93
7.3.1 Experiment Description . . . 93
7.3.2 Procedure . . . 94

7.3.3 Results . . . 95
7.4 Methods for Intention Estimation and Trajectory Prediction . . . 96
7.4.1 Pedestrian Intention Estimation . . . 97
7.4.2 Pedestrian Trajectory Prediction . . . 97
7.4.3 Implementation . . . 98
7.5 Experimental Evaluations . . . 99
7.5.1 Datasets . . . 99
7.5.2 Metrics . . . 99
7.5.3 Pedestrian Intention Estimation . . . 99
7.5.4 Trajectory Prediction . . . 100
7.5.5 Ego-Vehicle Speed Prediction . . . 102
7.5.6 Intention in Trajectory Prediction . . . 102
7.6 Summary . . . 103

8 Anticipating Pedestrian Crossing Action Using Contextual Cues 105
8.1 A Review of Action Prediction Algorithms . . . 106
8.2 Anticipating Crossing Using Multi-Modal Data Fusion . . . 108
8.2.1 Context for crossing prediction . . . 108
8.2.2 Architecture . . . 110
8.2.3 Implementation . . . 111
8.3 Experimental Evaluations . . . 112
8.3.1 Dataset . . . 112
8.3.2 Metrics . . . 112
8.3.3 Predicting Crossing Events . . . 113
8.3.4 When to Predict Crossing Events . . . 114
8.3.5 The Effect of Observation Length on Prediction . . . 115
8.3.6 Feature Types and Prediction Accuracy . . . 116
8.3.7 The Order of Fusion and Performance . . . 117
8.4 Does Intention Help Action Prediction? . . . 117
8.5 Summary . . . 118

9 Final Remarks 119
9.1 Dissertation Summary . . . 119
9.2 Study Limitations . . . 120
9.3 Future Work . . . 121
9.4 Appendix A: Chapters, Corresponding Publications and Contribution . . . 159

List of Tables

5.1 Patterns of behavior in JAAD 63

6.1 Pedestrian detection algorithms in the presence of different pedestrian attributes 80
6.2 Performance of pedestrian detection algorithms on JAAD and Caltech datasets 82
6.3 Generalizability of pedestrian detection algorithms on JAAD and Caltech datasets 83

7.1 A comparison between PIE and JAAD dataset 92
7.2 Performance of pedestrian intention estimation method 99
7.3 Performance of trajectory estimation method using only bounding boxes . . 100
7.4 Performance of vehicle speed prediction method 102
7.5 Performance of trajectory prediction method using different contextual information 103

8.1 Evaluation of alternative approaches for crossing prediction 113
8.2 Role of context on crossing prediction 116
8.3 Impact of feature fusion strategies on crossing prediction 117
8.4 Impact of pedestrian intention estimation on crossing prediction 118

List of Figures

1.1 A futuristic vision for autonomous driving . . . 2
1.2 Levels of autonomous driving . . . 3
1.3 A visual history of advancements in autonomous driving . . . 4
1.4 Walter's Tortoises and Standford Cart . . . 5
1.5 Early remote-controlled vehicles . . . 6
1.6 GM's automatically guided automobile . . . 6
1.7 VaMoRs, the autonomous car . . . 7
1.8 Stanley and BOSS at DARPA challenge . . . 9
1.9 Modern autonomous cars . . . 11

2.1 A monkey is imitating human gestures. . . . 17
2.2 Coordination between humans to accomplish a task . . . 20
2.3 From joint attention to crossing . . . 21

3.1 Wizard of Oz behavior study technique . . . 28
3.2 Classical data collection techniques . . . 29
3.3 Data collection methods involving AVs . . . 30
3.4 Factors involved in pedestrian crossing behavior . . . 31
3.5 Classical factors and corresponding papers . . . 39
3.6 Factors involved in pedestrian crossing behavior interacting with AVs . . . 40
3.7 Impact of AVs appearance on pedestrian behavior . . . 42
3.8 Factors identified in AV studies and corresponding papers . . . 44

4.1 Forms of distraction while driving an AV . . . 47
4.2 Different forms of nonverbal communication . . . 49
4.3 The function of hand gesture depending on its lexical meaning . . . 52
4.4 The ways pedestrians communicate in traffic . . . 53
4.5 Frequency, types and methods of communication . . . 55
4.6 The impact of crosswalk type and group size on looking frequency . . . 56

4.7 Pedestrian looking patterns in groups . . . 57

5.1 A behavioral annotation timeline from JAAD . . . 61
5.2 Pedestrians motifs at the time of crossing . . . 62
5.3 Relationship between TTC and attention . . . 64
5.4 Impact of traffic signal on pedestrian attention. . . . 64
5.5 Impact of road width on pedestrian attention . . . 65
5.6 Impact of age on duration of attention . . . 65
5.7 Crosswalk property and crossing behavior . . . 66
5.8 Variability in pedestrian action . . . 67
5.9 Identifying relevant pedestrians . . . 68
5.10 Role of context in understanding behavior . . . 69
5.11 Gestures with no symbolic meaning . . . 69
5.12 Role of other road users on pedestrian crossing behavior . . . 70

6.1 Sources of error in pedestrian detection . . . 73
6.2 Pedestrian attributes in JAAD . . . 76
6.3 ROC curves for pedestrian detection algorithms trained on JAAD . . . 78
6.4 Background and localization error in pedestrian detection algorithms . . . 79
6.5 Performance of MS-CNN and SDS-RCNN on JAAD . . . 81
6.6 Sample results of pedestrian detection algorithms on different subsets of JAAD 84

7.1 Stages of predicting pedestrian behavior . . . 87
7.2 Examples of freeze-frames from the human experiment . . . 93
7.3 Responses of human subjects on pedestrians' intentions . . . 94
7.4 Proposed system architecture for pedestrian intention estimation and trajectory prediction . . . 96
7.5 Examples of pedestrian intention estimations . . . 101
7.6 Examples of trajectory predictions . . . 104

8.1 Examples of pedestrians prior to making crossing decisions . . . 105
8.2 Architecture of SF-GRU for pedestrian crossing prediction . . . 111
8.3 Examples of crossing prediction made by SF-GRU . . . 114
8.4 Changes in crossing prediction performance with respect to varying TTE . . . 115
8.5 Impacts of TTE and observation length on crossing prediction . . . 115

Chapter 1

Introduction

Ever since the introduction of early commercial automobiles, engineers and scientists have been striving to achieve autonomy, which is removing the need for human involvement from controlling the vehicles. The fascination with autonomous driving technology is not new and goes back to the 1950s. In that era, articles appeared in the press featuring the autonomous vehicles in Utopian cities of the future (Figure 1.1) where drivers, instead of spending time controlling the vehicles, could interact with their family members or undertake other activities while enjoying the ride to their destinations [1].

Apart from the increased level of comfort for drivers, autonomous vehicles can positively impact society both at the micro and macro levels. One important aspect of autonomous driving is the elimination of driver involvement, which reduces human errors (e.g. fatigue, misperception or inattention), and consequently, lowers the number of accidents (up to 93.5%) [2]. The reduction in human error can improve both the safety of the driver or the passengers of the vehicle and other traffic participants such as pedestrians. At the macro level, fleets of autonomous vehicles can improve the efficiency of driving, better the flow of traffic and reduce car ownership (by up to 43%) through car-sharing, all of which can minimize energy consumption, and as a result, lower the environmental impact such as air pollution and road degradation [3].

Over the past few decades, the automotive industry has witnessed many significant breakthroughs in the field of autonomous driving, ranging from simple lane following [4] to complex maneuvers and interaction with traffic in complex urban environments [5]. Today, autonomous driving has become one of the major topics of interest in technology. This field has not only attracted the attention of the major automotive manufacturers, such as BMW, Toyota, and Tesla but also enticed a number of technology giants such as Google, Apple, and Intel.

Despite the significant amount of interest in the field, there is still much to be done

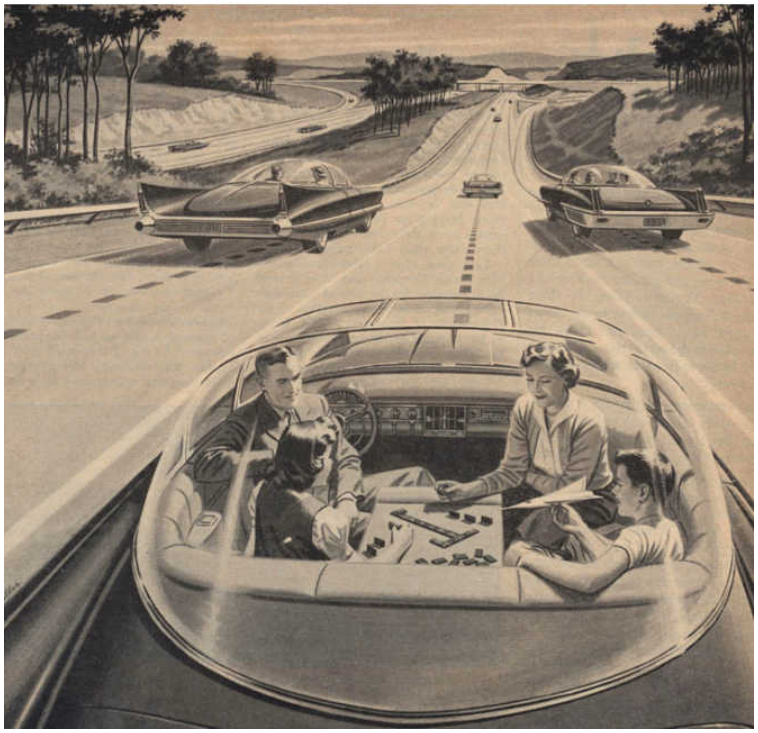


Figure 1.1: A view of a futuristic autonomous vehicle in which a family of four are playing a board game while enjoying a ride to their destination, 1956. Source: [1].

to achieve fully autonomous driving behavior in the sense of designing a vehicle capable of handling all dynamic driving tasks without any human involvement. One of the major challenges, besides developing efficient and robust algorithms for tasks such as visual perception and control, is interaction with other road users in chaotic traffic scenes. Interaction is a vital component in resolving various traffic ambiguities such as yielding to others or asking for the right of way. In order for the interaction to be effective, the parties are required to understand each others' behavior, to have the ability to predict each others' actions and to communicate their intentions.

The objective of this chapter is to discuss why pedestrian behavior understanding and prediction is necessary for autonomous driving systems. Before doing so, however, it is important to define what autonomy means in the context of driving, what is the progress in developing autonomous driving systems and what remains unsolved. To achieve these objectives, we present a brief history of autonomous driving systems and discuss some of the major milestones as well as unresolved challenges. We argue why pedestrian behavior understanding and prediction are important in the context of autonomous driving and how they can impact the flow of traffic and the safety of road users.

1.1 Autonomous Vehicles (AVs): A Brief History

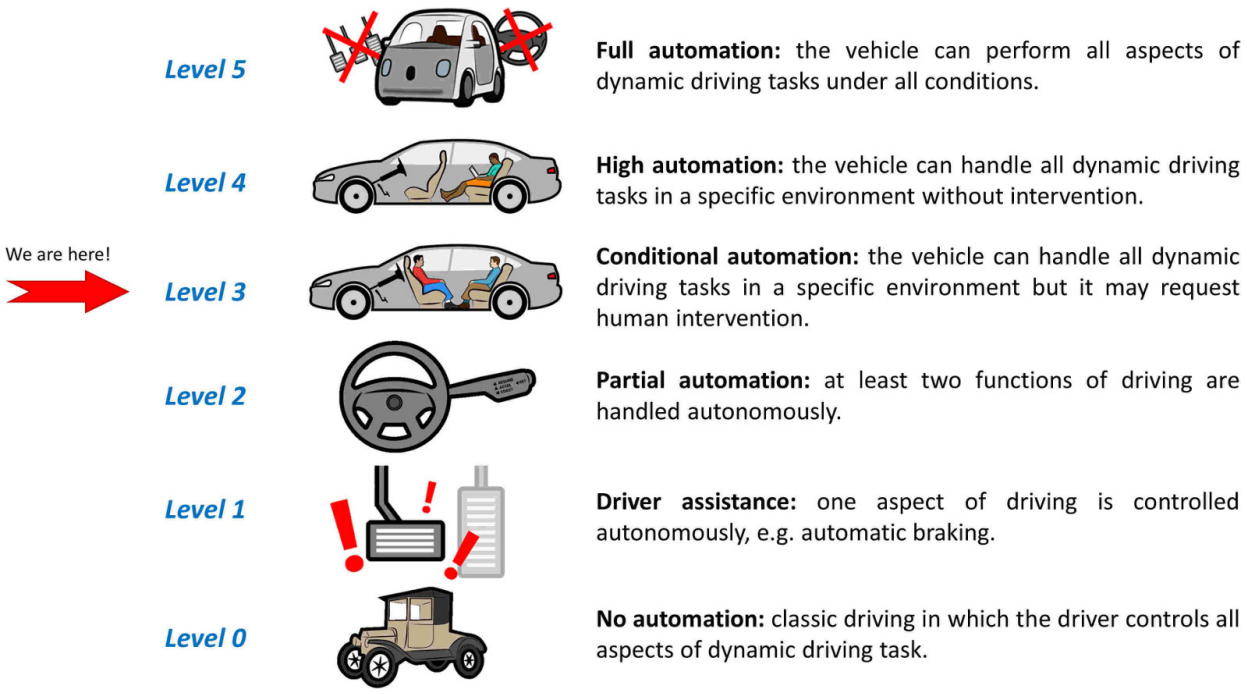


Figure 1.2: Six levels of driving automation. Today we have achieved level 3 autonomy. Source: [6].

Before reviewing the development of autonomous driving technologies, it is necessary to define what we mean by autonomy in the context of driving. Traditionally, there have been four levels of autonomy including no autonomy (the driver is in the control of all driving aspects), advisory autonomy (such as warning systems in the vehicle which partially aid the driver), partial control (such as auto-braking or lane adjustment) and full control (all aspects of the dynamic driving tasks are handled autonomously)[7].

Today, the automotive industry further breaks down the levels of autonomy into six categories: (see Figure 1.2)[8]:

Level 0: *No Automation*, where the human driver controls all aspects of the dynamic driving tasks. This level may include an enhanced warning system but no automatic control is taking place.

Level 1: *Driver Assistance*, where only one function of driving such as steering or acceleration/deceleration, using information about the driving environment, is handled autonomously. The driver is expected to control all other aspects of driving.

Level 2: *Partial Automation*. In this mode, at least two functionalities of the dynamic driving task, both steering and acceleration/deceleration, are controlled autonomously.

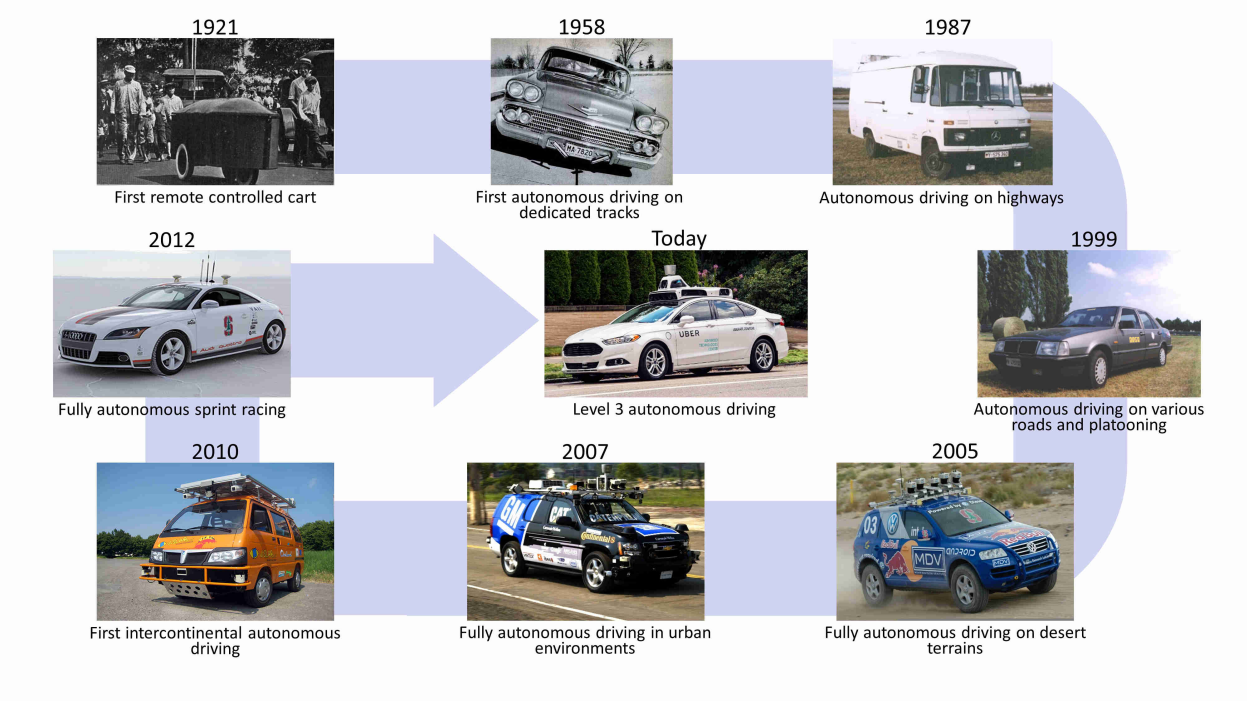


Figure 1.3: A century of developments in driving automation. This timeline highlights some major milestones in autonomous driving technologies from the first attempts in the 1920s to today's modern autonomous vehicles. Sources (in chronological order): [9, 1, 10, 11, 12, 13, 14, 15, 16]

Level 3: *Conditional Automation*, where the autonomous system can handle all aspects of the dynamic driving tasks in a specific environment but may require human intervention in the cases of failure.

Level 4: *High Automation*. This mode is similar to level 3 with the exception that no human intervention is required at any time during the environment-specific driving task.

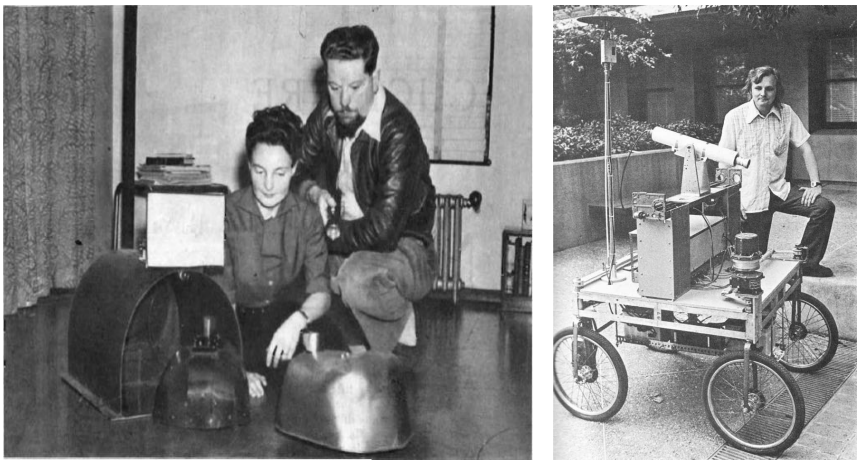
Level 5: *Full Automation*. As the name implies, in this mode all aspects of the dynamic driving tasks under any environmental conditions are fully handled by an automated system.

The current level of autonomy available on the market, such as the one offered by Tesla, is level 3.

The following subsections will review the developments in the field of autonomous driving during the past century. A summary of some of the major milestones are illustrated in Figure 1.3.

1.1.1 The Beginning

Much of today's autonomous driving technology is owing to the pioneering works of roboticists such as Sir William Grey Walter, a British neurophysiologist who invented the robots



(a) (b)

Figure 1.4: a) W. G. Walter and his Tortoises [17], and b) Hans Moravec and Stanford Cart [18].

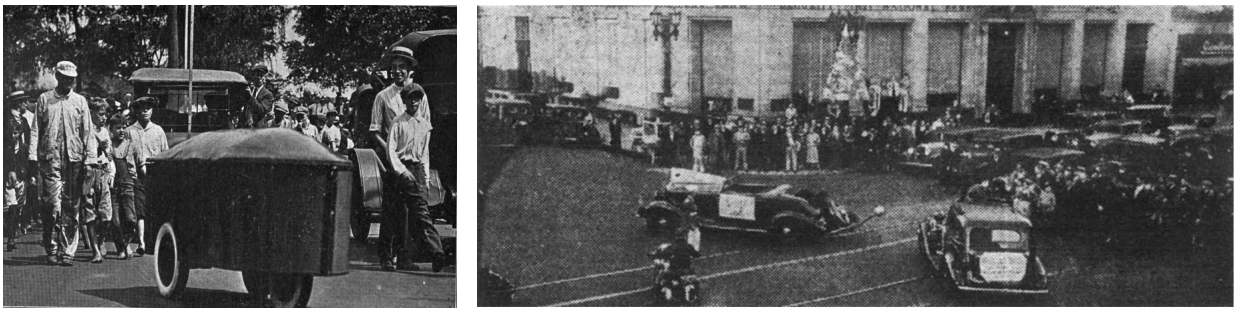
Elsie and Elmer (also known as Tortoises)(Figure 1.4a), in the year 1948 [19]. These simple robotic agents were equipped with light and pressure sensors and are capable of phototaxis by which they could navigate their way through the environment to their charging station. The robots were also sensitive to touch which allows them to detect simple obstacles on their path.

A more modern robotic platform capable of autonomous behaviors is the Stanford Cart (Figure 1.4b) [20, 21]. This mobile platform was equipped with an active stereo camera and could perceive the environment, build an occupancy map and navigate its way around obstacles. In terms of performance, the robot successfully navigated a 20 m course in a room filled with chairs in just under 5 hours.

Autonomous vehicles rely on similar techniques as in robotics to perform various perception and control tasks. However, since vehicles are used on roads, they generally require different and often stricter performance evaluations, in terms of robustness, safety, and real-time reactions. In the remainder of this chapter we particularly focus on robotic applications that are used in the context of autonomous driving.

Early attempts at developing autonomous driving technology go as far back as the first commercial vehicles. In this era, autonomous driving was realized in the form of remote-controlled vehicles removing the need for the driver to be physically present in the car.

In 1921, the first driverless car (Figure 1.5a) was developed by the McCook Air Force Base in Ohio [1]. This 2.5 meter-long cart was controlled via radio signals transmitted from the distance of up to 30 m. In the 1930s, this technology was implemented on actual vehicles some of which were exhibited in various parades (Figure 1.5b) to promote the future of

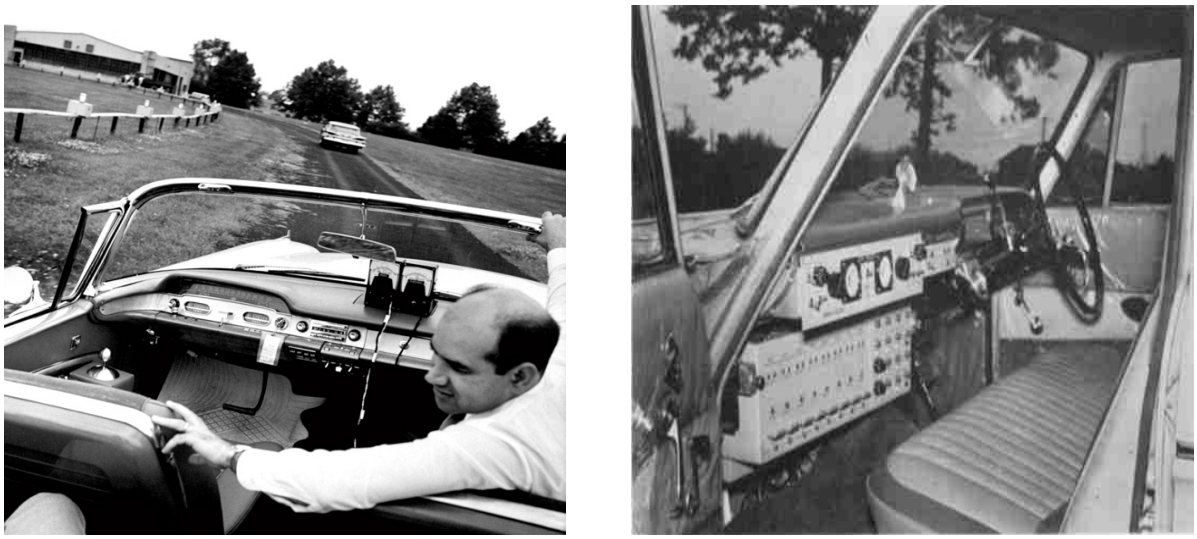


(a) (b)

Figure 1.5: a) The first remote-controlled vehicle, 1921 [9], and b) a more modern version of a commercial vehicle at Safety Parade 1936 [1].

driverless cars and to show how they could increase driving safety [1].

1.1.2 Hitting the Road



(a) (b)

Figure 1.6: a) GM's automatically guided automobile [22] and b) Tsukaba Lab's autonomous car [23].

The earliest instance of driving an automobile without human involvement was introduced in 1958 by General Motors (GM). The autonomous vehicle called "automatically guided automobile" (see Figure 1.6a) was capable of autonomous driving on a test track with electric wires laid on the surface which were used to automatically guide the vehicle steering mechanism [1].

Perhaps, the first truly autonomous vehicle was introduced in 1977 by the research team

Figure 1.7: VaMoRs and a view of its interior [24].

at Tsukuba Mechanical Engineering Lab in Japan [25]. This vehicle could drive with the speed of 33 km/h by recognizing and following lane markings detected by a computer vision algorithm (see Figure 1.6b). The car, however, still relied on inputs received from an elevated rail installed on the road. In the late 1980s, one of the pioneers of modern autonomous driving, E. D. Dickmanns [4, 26], alongside his team of researchers at Daimler, developed a visual algorithm for road detection in real-time. They employed two active cameras to scan the road, detect its boundaries, and then measure its curvature. To reduce computation time, a Kalman filter was used to estimate the changes in the curvature as the car was traversing the road.

In the early 1990s, the team at Daimler enhanced the algorithm by adding an obstacle detection capability. This algorithm identified parts of the road as obstacles if their height was more than a certain elevation threshold above the 2D road surface [27]. In the same year, the visual perception algorithm was tested on an actual Mercedes van, VaMoRs (Figure 1.7). Using the algorithm in conjunction with an automatic steering mechanism, VaMoRs was able to drive up to the speed of 100 km/h on highways, and up to 50 km/h on regional roads. The vehicle could also perform basic lane-changing maneuvers and safely stop before obstacles when driving up to 40 km/h.

Throughout the same decade, we witnessed the emergence of learning algorithms such as neural networks which were designed to handle various driving tasks. ALVINN is one such example developed that was developed as part of the NAVLAB autonomous car project by Carnegie Melon University (CMU). The system used a neural network to learn and detect different types of roads (e.g. dirt or asphalt) and obstacles [28, 29, 30, 31]. The algorithm, besides passive camera sensors, relied on laser range finders and laser reflectance sensors (for surface material assessment) to achieve a more robust detection.

To guide the vehicle, a similar learning technique was used by the NAVLAB team to learn

driving controls from recordings collected from an expert driver [32]. An extension of this project used an online supervised learning method to deal with illumination changes, and a neural network to identify more complex road structures such as intersections [33]. The NAVLAB project was implemented on a U.S. Army High Mobility Multipurpose Wheeled Vehicle (HMMWV) and was capable of obstacle avoidance and autonomous driving up to 28 km/h on rugged terrains and 88 km/h on regular roads.

Despite the fact that learning algorithms achieved promising performance in various visual perception tasks, in the late 90s, the traditional vision algorithms still remained popular. Methods such as color thresholding [11] and various edge detection filters such as Sobel filters [34] or model-based algorithms for road boundary estimation and prediction [35] were widely used.

In the mid-90s, autonomous assistive technologies had become standard features in a number of commercial vehicles. For instance, an extension of the lane detection and following algorithms developed by Dickmanns' team [36, 37] was used in the new lines of Mercedes-Benz vehicles [38]. This new extension, in addition to road detection, could detect cars by identifying symmetric patterns in their rear views. Using the knowledge of the road, the automatic system adjusted the position of the vehicle within the lanes and performed emergency braking if the vehicle came too close to an obstacle. An interesting feature of this system was the ability to track objects, allowing the vehicle to autonomously follow a car ahead of it, i.e. the ability to platoon.

The new millennium was the time during which autonomous vehicles started to enjoy the technological advancements in both the design of sensors and increase in computing power. At this time, we observed an increase in the use of high power sensors such as GPS, LIDAR, high-resolution stereo cameras [39] and IMU [40]. The information from various sources of sensors was commonly used by autonomous vehicles, thanks to the availability of high computing power, which allowed them to achieve better performance in tasks such as assessment of the environment, localization, and navigation of the vehicle and mapping. The emergence of such features brought the automotive industry one step closer to achieving full autonomy.

1.1.3 Achieving Autonomy

In 2004 the Defense Advanced Research Projects Agency (DARPA) organized one of the first autonomous driving challenges in which the vehicles were tasked to traverse a distance of 240 km between Las Vegas and Los Angeles [40]. In this competition, none of the 15 finalists were able to complete the course, and the longest distance traveled was only 11.78

(a) (b)

Figure 1.8: a) Stanley from Stanford in DARPA 2005 [12], and b) BOSS from CMU in DARPA 2007 [13].

km by the team Red from CMU.

The following year a similar challenge was held over the course of 212 km on a desert terrain between California and Nevada [41]. This year, however, 5 cars finished the entire course (one of them over the 10 hours limit), out of 23 teams that participated in the final event. Stanley (Figure 1.8a), the winning car from Stanford, finished the race in 6 hours and 53 minutes while maintaining an average speed of 30 km/h throughout the race [42].

Stanley benefited from various sources of sensory input including a mono color camera for road detection and assessment, GPS for global positioning and localization, and RADAR and laser sensors for long and short-range detection of the road respectively. The Stanley project produced a number of state-of-the-art algorithms for autonomous driving such as the probabilistic traversable terrain assessment method [43], a supervised learning algorithm for driving on different surfaces [44] and a dynamic path planning technique to deal with challenging rugged roads [45].

In the year 2007, DARPA hosted another challenge, and this time it took place in an urban environment. The goal of this competition was to test vehicles' ability to drive a course of 96 kilometers under 6 hours on urban streets while obeying traffic laws. The cars had to be able to negotiate with other traffic participants (vehicles), avoid obstacles, merge into traffic and park in a dedicated spot. In addition to robot cars, some professional drivers were hired to drive on the course.

Among the 11 finalists, BOSS (Figure 1.8b) from CMU [46] won the race. Similar to Stanley, BOSS benefited from a wide range of sensors and was able to demonstrate safe driving in traffic at the speed of up to 48 km/h.

Ever since the DARPA challenges, continuous improvements have been made in various tasks that contribute to achieving full autonomy, such as high-resolution and accurate mapping [47, 48], and complex control algorithms capable of estimating traffic behavior and responding to it [49, 50, 51].

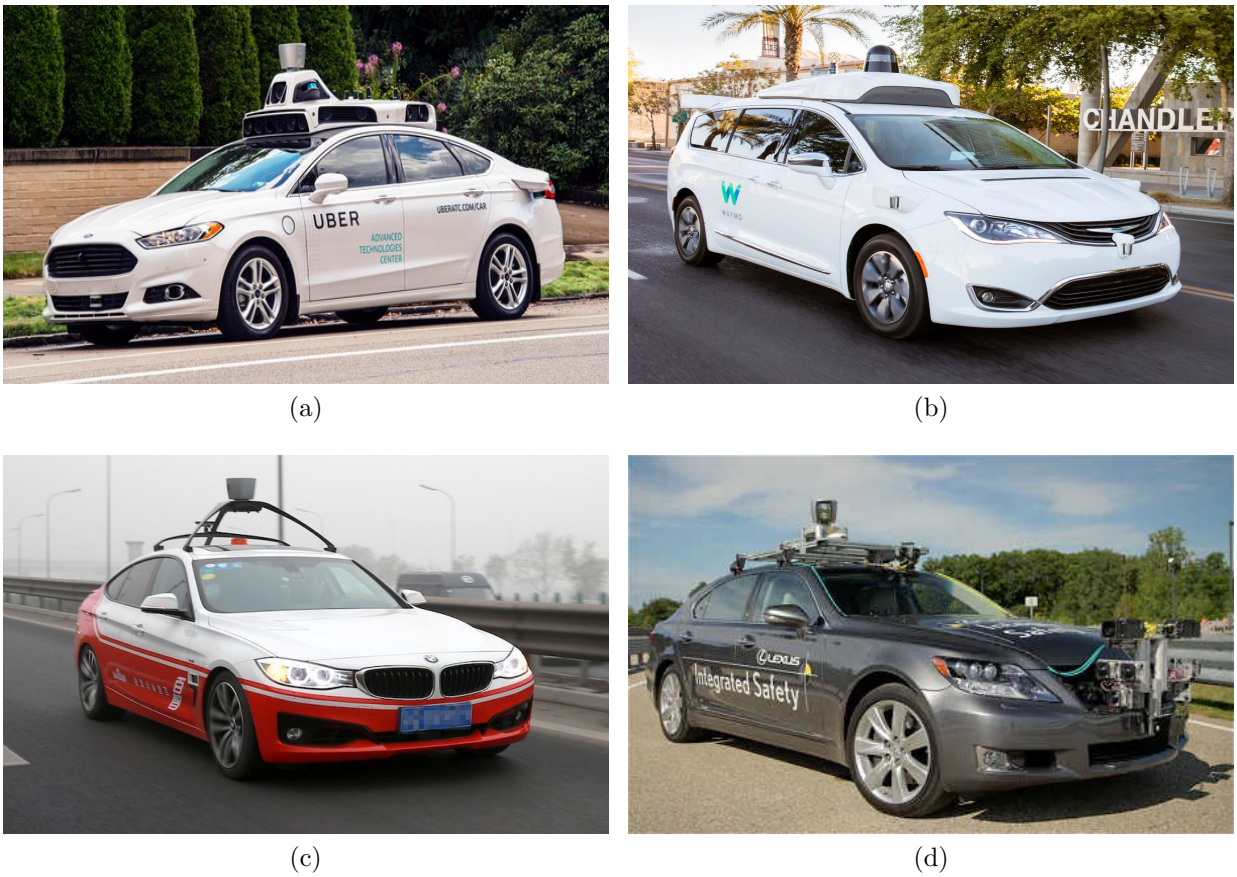
Autonomous vehicles have also been put to the test on larger scales. In the year 2010, the autonomous driving company, VisLab, held an intercontinental challenge by setting the goal of driving the distance of over 13000 km from Parma in Italy to Shanghai in China [52]. Four autonomous vans each with 5 engineers onboard participated in the challenge. One unique feature of this challenge was that the autonomous vehicles, for most of the course, performed platooning in which one vehicle led the way and assessed the road while the others followed it. The driving challenge was concluded after three month of driving. During this time, the vehicles collected 50 TB of data to be used in for future developments.

Furthermore, autonomous cars have found their way into racing. Shelley from Stanford [53] was one of the first vehicles that autonomously drove the 20 km world-famous Pikes Peak International Hill Climb in only 27 minutes while reaching a maximum speed of 193 km/h.

1.1.4 Today's Autonomous Vehicles

Today, more than 40 companies are actively working on autonomous vehicles [57] including Tesla [58], BMW [59], Mercedes [60], and Google [61]. Although most of the projects run by these companies are in the research stage, some are currently being tested on actual roads (see Figure 1.9). A few companies such as Tesla already sell their newest models with autonomous driving capability and claim that these vehicles have all the hardware needed for fully autonomous driving.

Autonomous driving research is not limited to passenger vehicles. Recently, Uber has successfully tested its autonomous truck system, Otto, to deliver 50,000 cans of beer by driving the distance of over 190 km [62]. The Reno lab, at the University of Nevada, also announced that they are working on an autonomous bus technology and are planning to put it to test on the road in the near future [63]. Autonomous driving technology is even coming to ships. In a recent news release, Rolls-Royce has disclosed its plans on starting a joint industry project in Finland, called Advanced Autonomous Waterborne Applications (AAWA), to develop a fully autonomous ship technology by no later than 2020 [64].



(a) (b)

(c) (d)

Figure 1.9: Modern autonomous cars: a) Uber [16], b) Waymo (from Google) [54], c) Baidu [55], and d) Toyota [56].

1.2 Where AVs Should Be and Where They Are Now

How far we are from achieving fully autonomous driving technology is a subject of controversy. Companies such as Tesla [65] and BMW [59] are optimistic and claim that their first fully autonomous vehicles will enter the market by 2020 and 2022 respectively. Other companies such as Toyota are more skeptical and believe that we are nowhere close to achieving level 5 autonomy [66].

So the question is, what are the challenges that we need to overcome in order to achieve autonomy? Besides challenges associated with developing suitable infrastructures [67] and regulating autonomous cars [68], technologies currently used in autonomous vehicles are not robust enough to handle all traffic scenarios such as different weather or lighting conditions (e.g. snowy weather), road types (e.g. driving on roads without clear marking or bridges) or environments (e.g. cities with large buildings) [69]. Relying on active sensors for navigation significantly constrains these vehicles, especially in crowded areas. For instance, LIDAR, which is commonly used as a range finder, has a high chance of interference if similar sensors

are present in the environment [70].

Some of the consequences of these technological limitations are evident in recent accidents reports involving autonomous vehicles. Cases that have been reported include minor rear-end collisions [71, 72], car flipping over [73], and even fatal accidents [74, 75, 76].

Moreover, autonomous cars face another major challenge, namely interaction with other road users in traffic scenes [77]. The interaction involves understanding the intention of other traffic participants, communicating with them and predicting what they are going to do next.

1.3 The Importance of Prediction for Autonomous Systems

Prediction as part of interaction between road users is important because:

1. **It ensures the flow of traffic**. We as humans, in addition to official traffic laws, often rely on informal laws (or social norms) to interact with other road users. Such norms influence the way we perceive others and how we interpret their actions [78]. Lack of understanding of such behaviors can potentially slow down the flow of traffic. For instance, a recent report shows that AVs get confused by bicyclists and their "fixies" [79]. In this scenario, instead of putting a foot on the ground, the bicyclist was pedaling forward and backward while trying to maintain his balance and moving a minimal distance. Observing this action, the AV thought the bicyclists is about to cross the intersection and therefore it stopped despite having the right of way to cross first. Although this not a dangerous decision per se, it does contribute to confusion in traffic flow and is thus undesirable.

   In addition, communication, as part of the interaction, is necessary for managing traffic flow. Road users may communicate to disambiguate certain situations, e.g. if a car wants to turn at a non-signalized intersection to a street with heavy traffic, it might wait for another driver's signal indicating the right of way. In some cases, traffic officials are responsible for directing traffic by transmitting various nonverbal signals. The inability to understand these signals may lead to interruption of traffic flow. For instance, a recent article reports that an autonomous driving vehicle got stuck and halted the traffic at a school zone because it could not understand the instructions of a crossing guard [80].

2. **It improves safety**. Interaction can guarantee the safety of road users, particularly

pedestrians who are the most vulnerable traffic participants. For instance, at the point of crossing, pedestrians often establish eye contact with drivers or wait for an explicit signal from them. This assures the pedestrians that they have been seen, therefore if they commence crossing, the drivers will slow down or stop before them [81]. Failure to understand the intention of others, in an autonomous driving context, may result in accidents of the kind reported in recent years [82, 83].

3. **It helps identify malicious behaviors**. Given that autonomous cars may potentially commute without any passengers on board, they are subject to being disrupted or bullied [78]. For example, people may step in front of the car to force it to stop or change its route. Such instances of bullying have been reported involving some autonomous robots currently being used in malls. Some of these robots were defaced, kicked or pushed over by drunk pedestrians [84].

1.4 Alternatives to Behavior Understanding and Prediction

Designing practical systems for understanding road users' behaviors is quite challenging (as will be discussed later in this dissertation). To avoid this issue and ensure safe driving, many scientists turn to making autonomous vehicles behave conservatively, i.e. driving slower than usual, selecting paths that minimize the need for complex interactions, e.g. avoiding left turns, and coming to a stop at the smallest possibility of road conflicts. Although employing such an approach can potentially enhance the safety of road users, it can negatively impact the flow of the traffic. For example, recent reports mention that some autonomous cars behave hesitantly and slow down and stop often [85]. Some are pulled over by law enforcement officials for driving under the minimum speed limit [86].

Conservative driving can also negatively influence the riding experience of passengers, and, consequently, the adoption of these vehicles as a means of transportation. A recent report on Waymo's autonomous cars shows that these vehicles often select very easy and less crowded routes which can add up to 100% to overall travel time to the destinations [87].

Another approach that has been suggested by some roboticists is to retrain road users, in particular pedestrians, to behave appropriately around autonomous vehicles. Following this idea, pedestrians are expected to behave less erratically and cross only at designated crosswalks. Some scientists question this approach and believe that this is simply redefining the problem and if we are going to completely segregate autonomous vehicles, there are already existing technologies, such as trains, that do that [88]. Others state that the main

premise of autonomous vehicles was to eliminate traffic deaths. Now if this is to be achieved by asking humans to change their behaviors around these vehicles, this is simply shifting the responsibility to someone else [89].

As the arguments above suggest, the alternative approaches to behavior understanding move autonomous vehicles away from their main objectives, that are to ensure the safety of the other road users and improve traffic flow.

1.5 The Dissertation Outline

As discussed earlier, interaction with traffic participants is fundamental for safe driving. As part of an effective interaction, one requires a proper understanding of other road users' behaviors and various contextual factors that impact such behaviors and the ability to predict what will happen next.

The focus of this dissertation is on understanding road users behavior in the context of traffic interaction. In particular, we emphasize understanding and predicting pedestrian behavior , not only as a task of tracking and predicting the trajectory by extrapolating from current movement patterns, but importantly, doing so within the physical and social context in which that movement is occurring. In other words, we are trying to understand the contextual factors that impact such predictions and understanding. To this end, we structure the remainder of this dissertation as follows: Chapter 2 presents an overview of the psychology literature on social interaction and argues why humans need to predict behavior when interacting with one another. In Chapter 3, we perform a meta-analysis of a large body of literature on pedestrian behavior understanding in traffic context and identify the set of contextual factors that impact the behavior, the interconnections between these factors, and the implications of these findings for developing practical systems. In Chapter 4, an overview of studies of nonverbal communication is presented, followed by an empirical study on pedestrian communication based on a novel large-scale naturalistic driving dataset collected as part of this dissertation. The study details the methods of communication and their meanings as well as factors that impact the way pedestrians communicate. Chapter 5 looks at pedestrian crossing actions by analyzing the patterns of behaviors that pedestrians exhibit at the time of crossing and the factors that impact the ways pedestrians make crossing decisions.

The remaining chapters focus on the practical aspects of pedestrian behavior understanding. Chapter 6 investigates the limitations of pedestrian detection algorithms with respect to the properties of traffic datasets. In Chapter 7, we discuss pedestrian intention estimation and present findings from a human experiment on a novel large-scale traffic dataset. We

use the outcome of the experiment to develop a practical algorithm for pedestrian intention estimation and show how it can be incorporated into a pedestrian trajectory prediction framework for improved results. Chapter 8 is dedicated to pedestrian crossing prediction. In this chapter, we analyze the impact of various sources of contextual information on crossing prediction and also evaluate different architectural designs for incorporating such multi-modal data in a learning framework. In Chapter 9, we summarize the findings of our studies, discuss their limitations and recommend future research directions. It should be noted that parts of this dissertation were collaborative work. These contributions are detailed in Appendix A.

1.6 Terminology

Throughout this dissertation we use a number of terms to discuss the problem of pedestrian behavior understanding. For better understanding of the content, below we briefly explain some of these terms. These terms and others often have varying and inconsistent uses in the literature, so it is important to specify their meaning for this dissertation. Note that some of these terms will be explained in more detail later in this dissertation.

Behavior. We use the most general definition of behavior as in the series of actions one may perform. This includes the way one conducts oneself and acts in response to a situation or stimulus.

Action refers to the process of doing something, in our case, walking, gesturing, crossing the road, etc.

Context includes everything that potentially has an effect on one's behavior or the occurrence of an event. In traffic scenarios, context may consist of the road structure, weather conditions, norms, social factors, traffic volume, etc.

Intention refers to the underlying motives of someone to do something. For example, a pedestrian standing at the curb might have the intention of crossing the road or might have the intention of hailing a taxi. Whether they will act on it depends on surrounding conditions, e.g. whether the signal for crossing is green or a taxi is available in the vicinity.

1.7 Evaluation Methodology

In an autonomous driving system, the pedestrian prediction module is only one component among many other modules that are responsible for performing various tasks such as perception (e.g. observing the scene, detecting and tracking objects), planning (e.g. route planning) and control (e.g. steering, braking). Being a part of a bigger system, the performance of the prediction module highly depends on the performance of other components.

For example, if the perception is faulty (e.g. it produces wrong detections or poorly localizes objects), the prediction module would receive erroneous inputs, and consequently, produce inaccurate outputs. Control and planning modules can also impact prediction. For instance, if the vehicle does not maintain the speed or driving direction that the prediction algorithm has expected, the behavior of pedestrians would be different from what was predicted.

Moreover, in a driving framework, the full observations of pedestrians' current behaviors are not always possible. The observations are often gradual (one or more frames), partial (full sequence of observation is not always available), and disconnected (pedestrians might be occluded or out of view of the sensors for a period of time). All of these factors may affect the accuracy of a prediction algorithm.

In this dissertation, we do acknowledge that in order to report the true performance of a prediction algorithm, all of the conditions mentioned above should be present. However, to evaluate novel approaches or effect of new sources of contextual information, it is not feasible to test the algorithms in the entire driving platform as a whole. Apart from enormous time and resources required for such evaluations, it is difficult to measure the performance of the individual component within a complex system and separate sources of noise introduced by other modules.

To better reflect the performance of the proposed algorithms in this dissertation, we evaluate our methods using conventional approaches accepted in the computer vision and robotics communities. For the methods presented in Chapters 7 and 8, we use ground-truth observation tracks of pedestrians as input to the systems. If the methods expect a certain length of observations (e.g. 15 frames), we only consider tracks that are at least that long. In the datasets used in our work, care was taken to diversify the data as much as possible by collecting samples on different roads, under different weather and lighting conditions, capturing pedestrians with different characteristics, etc. In our experiments, we did not exclude challenging cases, such as those created by occlusion, shadow, reflection, etc., that might affect the visibility of pedestrians, and as a result, impact the prediction results.

Chapter 2

Social Interaction, Coordination and Behavior Prediction

In the previous chapter, we argued that prediction as part of social interaction is an essential component in safe driving and is necessary for resolving various traffic ambiguities, such as yielding to others or asking for the right of way, between drivers and pedestrians. For drivers to safely interact with pedestrians, they should be able to coordinate their actions. This requires an understanding of pedestrian behavior and the context in which the behavior is being observed. Before studying the role of context on pedestrian behavior, it is important to understand the nature of human interaction, how humans coordinate with one another and why behavior prediction is necessary in the course of an interaction. In the following section, we begin by explaining the joint attention phenomenon which is the first step for making interaction possible.

2.1 Joint Attention in Human Interaction

Figure 2.1: The monkey is imitating the human experimenter's gestures. Source: [90]

The precursor to any form of social interaction between humans (or primates [90], see Figure 2.1) is the ability to coordinate attention [91], which means the interacting parties at

very least should be able to pay attention to one another, discern the relevant objects and events of each other's attentional focus, and implement their own lines of action by taking into account where and toward what others may be attending. This dissertation asserts that this same ability to jointly attend to the same task must play a role in interactions between autonomous vehicle and human road users, as well as between autonomous vehicles.

In developmental psychology, the ability to share attention and to coordinate behavior is defined under the *joint attention* framework, also referred to as, *shared attention* [91] or *visual co-orientation* [92]. Traditionally, joint attention has been studied as a visual perception mechanism in which two or more interacting parties establish eye contact and/or follow one another's gaze towards an object or an event of interest [93, 94]. More recently, joint attention has also been investigated in different sensory modalities such as touch [95] or even remotely via the web [96]. Since the objective of this dissertation is reasoning based on visual perception in autonomous driving, in the following chapters we only focus on the problem of joint visual attention and simply address it as joint attention.

What does joint attention really mean? Joint attention is often defined as a triadic relationship between two interacting parties and a shared object or event [97, 98, 99, 91]. Intuitively, joint attention means the simultaneous engagement of two or more individuals with the same external thing [100].

In the traditional definitions, an important part of joint attention is the ability to reorient and follow the gaze of another subject to an object or an event [97, 99]. However, in more recent interpretations of joint attention, the gaze following requirement is relaxed and replaced by terms such as "mental focus" [100] or "shared intentionality" [101]. This means joint attention constitutes the ability of a person to engage with another for the sake of a common goal or task, which may not involve explicit gaze following action.

2.1.1 Early Studies of Joint Attention

Joint attention first was discovered in the context of early childhood development. In 1975, Scaife and Bruner [93] were the first to describe the joint visual attention mechanism and its role in early developmental processes in infants. They observed that infants below the age of 4 months were able to respond to the gaze changes of the adults in interactive sessions about one-third of the time. In comparison, the older infants, above the age of 11 months, almost always responded to the changes and could follow the gaze of the adults while interacting with them. In addition, at this age, infants could follow the eye movements of the adults as well as their head movements.

Butterworth and Cochran [92] further investigated joint attentional behavior and revealed

that infants between the age of 6 to 18 months adjust their line of gaze with those of the adult's focus of attention, however, they act only if the adult is referring to loci within the infant's visual space. For example, if the adult looked behind the infant, the infant only scans the space in front of them. The authors add that at early stages infants do not follow the gaze to the intended object, instead they turn their head to the corresponding side but focus their own gaze on the first object that comes in their field of view. The authors conclude that it is only in the second year when infants are able to focus on the same object that is intended by the adult.

In a subsequent study by Moore *et al.* [97] it is shown that while sharing attention, the actual movement is critical in gaze following. Through experimental evaluations, the authors illustrate that if only the final focus of the adult is presented to the infant they would not necessarily focus on the correct object.

2.1.2 Joint Attention in Development of Social Cognition

Joint attention has been linked to the development of social cognitive abilities such as learning of artifacts and environments [102, 103]. More specifically, joint attention is a fundamental component in language development through which infants learn to describe their surroundings [92, 94, 104]. In a study by Tomasello and Todd [94], it is argued that the lexical development of children depends on the way the joint attention activity is administered between the adult and the infant. It is shown that when mothers initiated interaction by directing their child's attention, rather than following it, their child learned fewer object labels and more personal-social words, i.e. their lexical development suffered but they were more expressive.

In short, Tomasello and Carpenter [101] summarize the social cognitive skills that are acquired through joint attention into four groups:

1. Gaze following and joint attention
2. Social manipulation and cooperative communication
3. Group activity and collaboration
4. Social learning and instructed learning

The importance of joint attention is not limited to early childhood and is believed to be vital for social competence at all ages. Adolescents and adults who cannot follow, initiate, or share attention in social interactions may be impaired in their capacity for forming relationships [91]. There are a large number of studies on the effects of joint attention weaknesses and

social disorders, in particular in people with autism [98, 105, 99, 106]. For instance, autistic children are found to have minor deficits in responding to joint attention while struggling to initiate joint attention. The effect of aging on joint attention has also been investigated [106]. It is shown that adults tend to get slower in gaze-cuing as they age.

2.1.3 From Imitation to Coordination

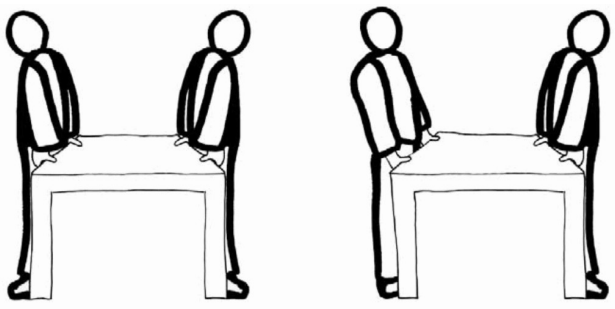


Figure 2.2: Coordination been humans for carrying a table. Rather than imitating each other's actions, (left image), people must sometimes perform complementary actions to reach a common goal (right image). Source: [107]

The traditional view of joint attention, as discussed earlier, focuses on the role of joint attention as a means whereby infants interact with adults and imitate their behavior to learn about their surroundings.

However, as adults, we often engage in more complex interactions, which can take many forms such as competition, conflict, coercion, accommodation, and cooperation [108]. Co-operation, as in the context of traffic interaction, refers to a social process in which two or more individuals or groups intentionally combine their activities towards a common goal [109], e.g. crossing an intersection.

What makes a complex coordination possible? Of course, the immediate answer is that a form of joint attention has to take place so the parties involved focus on a common objective. However, joint attention in its classical definition does not fully satisfy the requirements for cooperation. First, although certain cooperative tasks can be resolved by imitation (e.g. make the same movements to balance a table while carrying it), in some scenarios complementary actions are required to accomplish the task (e.g. the person at the front watches for obstacles while the one behind carries the table) [107], as shown in Figure 2.2. Second, even though involved parties focus on a common object or event, this does not mean that they also share the same *intention* (we will talk about intention more in Chapter 7). In this regard, in the context of cooperation, some scholars use the term *intentional joint attention* [110] indicating that the agents are not only mutually attending to the same entity,

but they also intend to do so.

This type of cooperation that involves a form of attention sharing is often referred to as joint action [110, 107]. In some literature, joint attention is considered as the simplest form of joint action [110]. However, for simplicity, throughout the rest of this report, we address the whole phenomenon as joint attention.

2.1.4 How Do Humans Coordinate?

Figure 2.3: From joint attention to crossing. Are these pedestrians going to cross?[1]

As described earlier, joint attention provides a mechanism for sharing the same perceptual input and directing attention to the same event or object [107]. Here, perhaps, the most crucial components to trigger this attentional shift are eyes, because they naturally attract the observer's attention even if they are irrelevant to the task. Of course, other means of communication such as hand gesture or body posture changes can be used for seeking attention [111]. This is particularly true in the case of traffic interaction where nonverbal communication between road users is the main means of establishing joint attention (see Chapter 4 for more details).

Next, the interacting parties have to understand each other's intentions in order to cooperate [110]. In some scenarios establishing joint attention might convey a message indicating

[1] a) no, b) no, c) no, d) yes, e) yes, and f) no.

the intention of the parties. For instance, at crosswalks, pedestrians often establish eye contact with the drivers indicating their intention of crossing [112]. However, joint attention on its own is not sufficient for understanding the intention of others (see Figure 2.3 for an example) or what they are going to do next. A more direct mechanism is action observation [107].

2.2 Why do We Predict Behavior?

2.2.1 A Biological Perspective

When it comes to understanding the observed actions of others, humans do not rely entirely on vision [113, 114]. In a study by Umilta *et al.* [114], the authors found that there is a set of neurons (referred to as mirror neurons) in the ventral premotor cortex (part of the motor cortex involved in the execution of voluntary movements) of macaque monkeys that fire both during the execution of the hand action and the observation of the same action in others. The authors show that a subset of these neurons become active during action presentation, even though the part of the action that is crucial in triggering the behavioral response is hidden and can therefore only be inferred. This implies that the neurons in the observer action system are the basis of action recognition [115].

Such anticipatory behaviors are also observed in humans. Humans, in general, have limited visual processing capability (due, in part, to foveation and saccadic eye movements), especially when it comes to observing multiple moving objects. Therefore, they actively anticipate the future poses of objects when interpreting a perceived activity [116]. In an experiment by Flanagan and Johansson [117], the authors showed a video of a person stacking blocks to a number of human subjects and measured their eye movements. They noticed that the gaze of the observers constantly preceded the action of the person stacking blocks and predicted a forthcoming grip in the same way they would perform the same task themselves. The authors then concluded that when observing an action, the human behavior is *predictive* rather than *reactive*.

It is necessary to note that such predictive behaviors in action observation have biological advantages for humans (and likely for machines too). In addition to dealing with visual processing limitations, anticipatory behaviors can help to deal with visual interruptions due to occlusion in the scene.

2.2.2 A Philosophical Perspective

From a philosophical perspective, it can also be shown that behavior (or action) prediction is the only way to engage in social interaction. For this purpose, we refer to the arguments presented by Dennet [118] and the comments on the topic by Baron-Cohen [119].

The ability to find "an explanation of the complex system's behavior and predicting what it will do next", which may include beliefs, thoughts, and intentions [118], or as Baron-Cohen terms it "mindreading", is crucial in both making sense of one's behavior and communication. Dennet argues that mindreading, or as he calls it adopting "intentional stance", is the only way to engage in social interaction. He further elaborates that the two alternatives to intentional stance, namely physical stance and design stance, are not sufficient for interpretation of one's intentions or actions.

According to Dennet, physical stance refers to our understanding of systems whose physical properties we know about, for instance, we know that cutting skin results in bleeding. In terms of understanding complex behavior, however, in order to rely on physical properties, we need to know millions of different physiological (brain) states that give rise to different behaviors. As a result, mindreading is an infinitely simpler and more powerful solution.

Design stance, on the other hand, tries to understand the system in terms of the functions of its observable parts. For example, one does not need to know anything about the microprocessor's internal design to understand the result of pressing the Delete key on the keyboard. Similarly, the design stance can explain some aspects of human behavior, such as blinking reflex in response to blowing on an eye surface, but it does not suffice to make sense of complex behaviors. This is primarily due to the fact that people have very few external operational parts for which one could work out a functional or design description.

In addition to behavioral understanding, Baron-Cohen [119] argues that mindreading is a key element in communication. Apart from decoding communication cues or words, we try to understand the underlying communicative intention. In this sense, we try to find the "relevance" of the communication by asking questions such as what that person "means" or "intends me to understand". For instance, if someone gestures towards a doorway with an outstretched arm and an open palm, we immediately assume that they mean (i.e. intend us to understand) that we should go through the door.

2.3 Summary

In this chapter, we elaborated on social interaction with a particular focus on the joint attention phenomenon, or the ability to share attention regarding a common object or an event.

We showed that, traditionally, joint attention was explored in the context of early human development to explain how children engage in social interactions and imitate the behaviors of adults to learn various behavioral and lexical skills. We argued that joint attention is also important in various cooperative tasks where, in addition to sharing attentional focus, parties involved are required to adjust their behavior accordingly in order to accomplish the task in hand.

Coordination between humans is only possible if they are aware of each other's intentions (or underlying motives). In this context, sharing the focus of attention can convey a message indicating the intention, e.g. looking at the traffic may indicate the intention of a pedestrian to cross the road.

However, from both biological and philosophical perspectives, we argued that intention is not always easily observable due to the complex factors underlying human behavior and decision-making process. The alternative to direct observation of intention is prediction based on various behavioral cues such as knowledge of the task, communication, and understanding of the context in which the interaction is taking place.

Chapter 3

Traffic Context and its Influence on Pedestrian Behavior

3.1 What do We Mean by Context?

If humans rely on predicting forthcoming behaviors of one another then how is a prediction generated? We answer this question in two parts: making sense of one's actions and interpreting communication cues. According to Humphrey [120], when observing someone's action, we first need to perceive the current state of being by relying on our sensory inputs. Next, we need to understand the meaning of the action by relating it to the knowledge of the task (e.g. crossing the street). This knowledge is either biologically encoded in our brain, for instance, people have a very accurate knowledge of the human body and how it moves [116], or, in more complex scenarios, it requires knowing the stimulus conditions (context) under which an individual performs an action. The context may include various physical or behavioral attributes present in the scene. Humphrey also emphasizes that in order to understand others, we need to predict the consequences of our actions and realize how they can influence their behavior [120].

The role of context is also highlighted in communication and how it can influence the way we convey communication cues. Sperber and Wilson [121], in their theory of relevance, argue that communication is achieved either by encoding and decoding messages via a code system that pairs internal messages with external signals or by using the evidence from the context to infer the communicator's intention. Although this theory was originally developed for verbal communication, it has implications that can certainly be relevant to nonverbal communication as well.

The scholars behind the theory of relevance claim that *code model* does not explain

the transmission of semantic representations and thoughts that are actually communicated. They believe that there is a need for an alternative model of communication, what they call *inferential model*. In an inferential process, there is a set of premises as input and a set of conclusions as output which follow logically from the premises. The communicator intentionally modifies the environment of his audience by providing a stimulus that takes two forms: the informative intention that informs the audience of something, and the communicative intention that informs the audience of the communicator's informative intention. On the other hand, the communicatee makes an inference using his background knowledge that he is sharing with the communicator, i.e. their knowledge of the context in which the communication is taking place [121].

To characterize the shared knowledge involved in communication, the authors use the term *cognitive environment*, which refers to a set of facts that are manifested to an individual. Intuitively speaking, the total cognitive environment of an individual consists of all the facts that he is aware of as well as all the facts that he is capable of becoming aware of at that time and place [121].

Sperber and Wilson use the term "relevance" to connect context to communication. They argue that any assumptions or phenomena (as part of cognitive environment) are relevant in communication if and only if they have some effect in that context. They add that the word "relevance" signifies that the contextual effect has to be large in the given context and at the same time requires small effort to be processed [121]. The amount of processing required to understand the context, however, is a subject of controversy.

Pedestrian behavior understanding and prediction in traffic scenes is a very complex task. Here, context can be quite broad involving various elements such as dynamic factors, (e.g. speed of the cars, the distance of the pedestrians), social factors (e.g. demographics), social norms, and environment configuration (e.g. street structure), and traffic signals. Given the complexity of the problem, in the past literature, numerous methodologies have been proposed for studying pedestrian behavior and identifying contextual factors that impact pedestrian decision-making process. In the following section, we review some of these methods.

3.2 Methods of Studying Pedestrian Behavior

The methods of studying human behavior in traffic scenes have transformed during past decades as new technological advancements have emerged. Traditionally, written questionnaires [122, 123] or direct interviews [124] were widely used to collect information from traffic participants or authorities monitoring the traffic. Some modern studies still rely on ques-

tionnaires especially in cases where there is a need to measure the general attitudes of people towards various aspects of driving, e.g. crossing in front of autonomous vehicles [125]. These forms of studies, however, have been criticized for potentially biased answers, the questionable honesty of responses or even how well the interviewees could recall a particular traffic situation.

Traffic reports are mainly composed by professionals, such as police forces after accidents [126]. The advantage of traffic reports is that they provide thorough description of the elements involved in a traffic accident, albeit not being able to substantiate the underlying reasons.

In addition, behavior can be analyzed via on-site observations by the researcher either present in the vehicle [127] or standing outside [128] while recording the behavior of the road users. Observations can be both naturalistic and scripted. In a naturalistic format, normal activities of road users are monitored without notifying them of such recording [129]. In a scripted setting, the participants, e.g. drivers or pedestrians, are instructed to perform certain actions, and then the reactions of other parties are observed [130, 131]. A major drawback of observation is the strong observer bias, which can be caused by both the observers' misperception of the traffic scenes or their subjective judgments.

New technological developments in the design of sensors and cameras have given rise to different modalities of recording traffic events. Eye-tracking devices are one such system that can record participants' eye movements during driving [132] or gaze of pedestrians who are crossing a street [133]. Computer simulations [134] and video recordings of traffic scenes [123] are also widely used to study the behavior of drivers in laboratory environments. These methods, however, are criticized for not providing realistic driving conditions, therefore the observed behaviors may not necessarily reflect the ones exhibited by road users in a real traffic scenario.

Naturalistic recording of traffic scenes (both videos [135] and photos [136]), is, perhaps, one of the most effective methods for studying traffic behavior. Although the first instances of such studies date back to almost half a century ago [137], they are still widely being used in recent years. In this method of study, a camera (or a network of cameras) is placed either inside the vehicle [135, 138] or outside on sidewalks [139, 140]. Since the objective is to record the natural behavior of the road users, the cameras are located in inconspicuous places not visible to the observees. In the context of recording driving habits, although the presence of the camera might be known to the driver, it does not alter the driver's behavior in the long run. In fact, studies show that the presence of cameras may only influence the first 10-15 minutes of the driving, hence the beginning of each recording is usually discarded at the time of analysis [127]. An added advantage of recording compared to on-site observation

(a) (b)

Figure 3.1: Examples of Wizard of Oz technique. a) The driver is disguised as a car seat [131] and b) the driver is driving the car from a right-hand steering wheel while a dummy driver is sitting in the actual driver's seat [141].

is the possibility of revisiting the observation and using multiple observers to minimize bias [137].

Naturalistic recording, similar to on-site observation, may also be affected by observer bias. Moreover, in some cases, it is hard to recognize certain behaviors or underlying motives, e.g. whether a pedestrian notices the presence of the car or looks at the traffic signal in the scene and why. To remedy this issue, it is common to employ a hybrid approach where recordings or observations are combined with on-site interviews [112]. Using this method, after recording a behavior, the researcher approaches the corresponding road user and asks questions regarding their actions, for example, whether they looked at the signal prior to crossing. Overall, the hybrid approach can help resolve the ambiguities observed in certain behaviors.

In the context of autonomous driving research, the Wizard of Oz technique [141] is common in which the experimenters simulate the behavior of an intelligent system to observe the reaction of subjects. Using this technique, experimenters may disguise themselves as a car seat [131] or control the vehicle from a hidden place inside the vehicle [141] that is not observable by the participants (see Figure 3.1).

Figures 3.2 and 3.3 summarize the works presented in this chapter and their methods of study. Note that in this figure, literature survey, refers to expert studies that generate new findings based on past works.

3.3 Factors Influencing Pedestrian Behavior

We divide pedestrian behavior studies into two categories, classical studies investigating the interactions between pedestrians and human drivers and studies involving interactions with

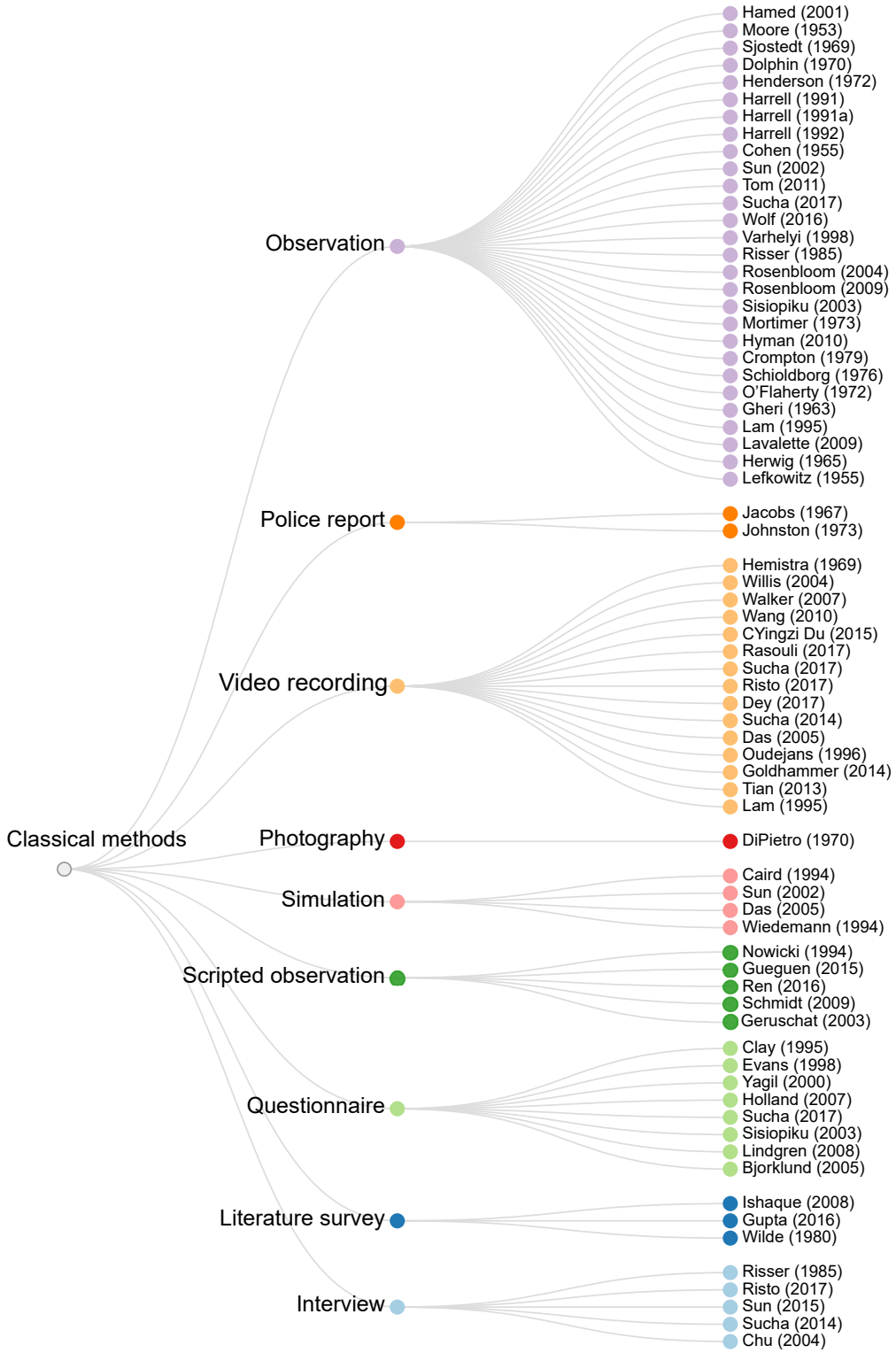


Figure 3.2: Data collection methods used in the classical pedestrian behavior studies.

autonomous vehicles. Compared to studies with autonomous vehicles, the classical studies focus on pedestrian behavior while interacting with human drivers instead of vehicles. All

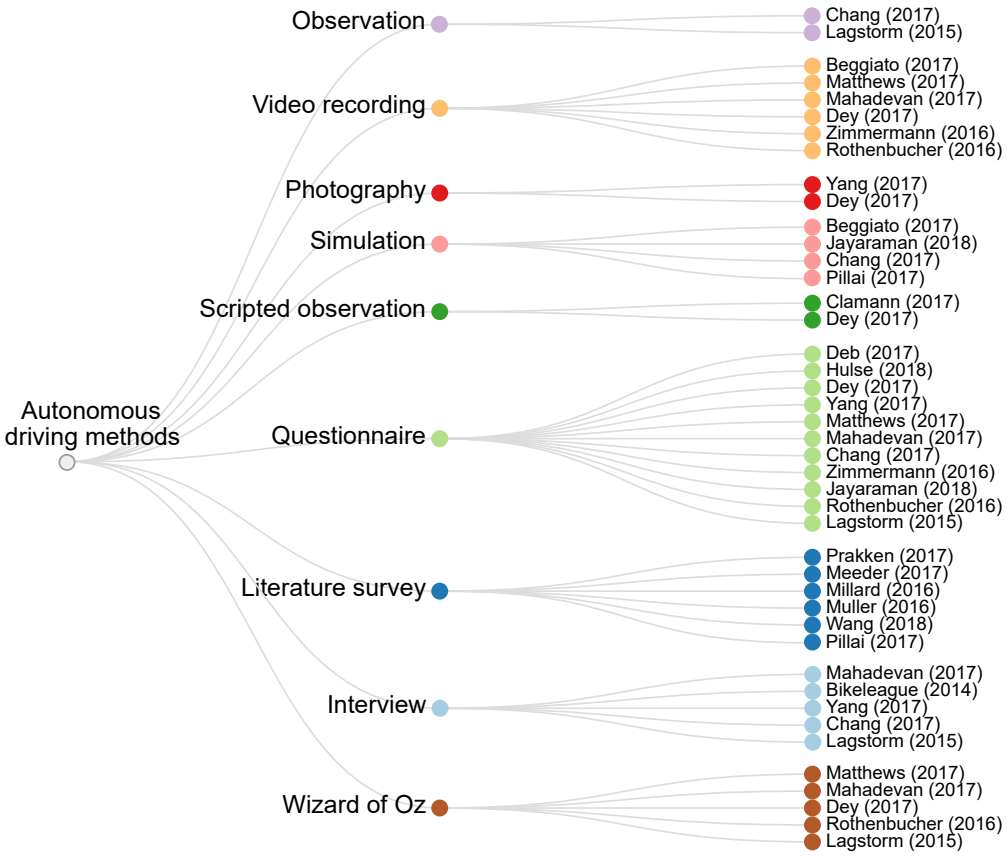


Figure 3.3: Data collection methods used in the pedestrian behavior studies involving autonomous vehicles.

the factors identified in the literature are italicized in the text.

3.3.1 Classical Studies

The early works in pedestrian behavior studies come from early 1950s, and since then there has been a tremendous amount of research done on various factors that impact pedestrian behavior. Given the magnitude of the work in this area, an exhaustive survey of all the literature would be prohibitive. As a result, only a subset of major works will be presented.

We divide the factors that influence pedestrian behavior into two groups, the ones that directly relate to pedestrians and environmental ones. For a summary of these factors and how they are interrelated refer to Figure 3.4.

Pedestrian Factors

Social Factors. Among the social factors, perhaps, *group size* is one of the most influential ones. Heimstra *et al.* [137] conducted a naturalistic study to examine the crossing behavior

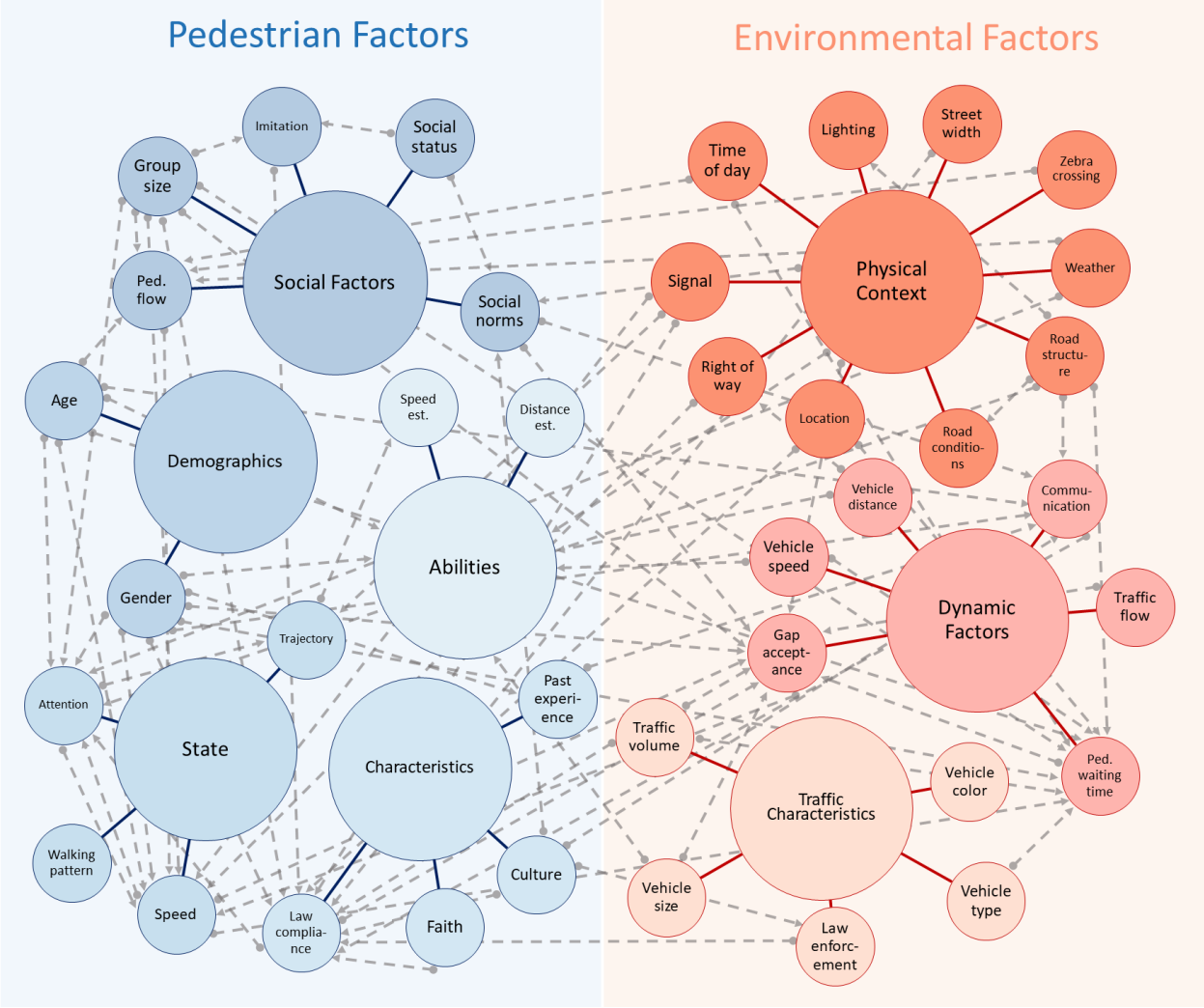


Figure 3.4: Factors involved in pedestrian decision-making process at the time of crossing. The diagram is based on a meta-analysis of the past literature. The large circles refer to the major factors and small circles connected with solid lines are sub-factors. The dashed lines show the interconnection between different factors and arrows show the direction of influence.

of children and found that they commonly (in more than 80% of the cases) tend to cross as a group rather than individually. *Group size* changes both the behavior of the drivers with respect to the pedestrians and the way the pedestrians act at crosswalks. For instance, it is shown that drivers more likely yield to groups of pedestrians (3 or more) than individuals [139, 142].

When crossing as a group, pedestrians tend to be more careless, pay less *attention* at crosswalks and often accept shorter gaps between the vehicles to cross [140, 143, 144] or do not look for approaching traffic [112]. *Group size* is also found to impact the way pedestrians comply with the traffic laws, i.e. *group size* exerts some form of social control over individual

pedestrians [145]. It is observed that individuals in a group are less likely to follow a person who is breaking the law, e.g. crossing on the red light [129].

In addition, *group size*, for obvious reasons, influences *pedestrian flow* which determines how fast pedestrians cross the street. Wiedemann [146] indicates that if there is no interaction between the pedestrians, there is a linear relationship between *pedestrian flow* and *pedestrian speed*. This means, in general, pedestrians walk slower in denser groups.

Social norms, or as some experts refer to as "informal rules" [78], play a significant role in how traffic participants behave and how they predict each other's intention [122]. *Social norms* also influence how acceptable a particular action is in a given traffic situation [147]. The difference between *social norms* and legal norms (or formal rules) can be illustrated using the following example: formal rules define the speed limit of a street, however, if the majority of drivers exceed this limit, the *social norm* is then quite different [122].

The influence of *social norms* is so significant that merely relying on formal rules does not guarantee safe interaction between traffic participants. To highlight this fact, Johnston [148] describes the case of a 34-year old married woman who was extremely cautious (and often hesitant) when facing yield and stop signs. In a period of four years, this driver was involved in 4 accidents, none of which she was legally at fault. In three out of four cases this driver was hit from behind, once by a police car. This example illustrates how disobeying *social norms*, even if it is legal, can disrupt traffic flow.

Social norms even influence the way people interpret the law. For example, the concept of "psychological right of way" or "natural right of way" has been studied [122]. This concept describes the situation in which drivers want to cross a non-signalized intersection. The law requires the drivers to yield to the traffic from the right. However, in practice drivers may do quite the opposite depending on the *social status* (or configuration) of the street. It is found that factors such as *street width*, *lighting* conditions or the presence of shops may determine how the drivers would behave [149].

Imitation is another social factor that defines the way pedestrians (as well as drivers [150]) would behave. A study by Yagil [151] shows that the presence of a law-adhering (or law-violating) pedestrian increases the likelihood of other pedestrians to obey (or disobey) the law. This study shows that the impact is more significant when law violation is involved.

The probability of *imitation* occurrence may depend on the *social status* of the person who is being imitated. In the study by Leftkowitz *et al.* [129] a confederate was asked by the experimenter to cross or stand on the sidewalk. The authors observed that when the research confederate was wearing a fancy outfit, there was a higher chance that other pedestrians would imitate his actions (either breaking the law or complying). This idea, however, is challenged by Dolphin *et al.* [152] whose findings indicate that *social status* and

gender have no effect on *imitation*. The authors claim that *group size* is a better predictor of *imitation*, which means the larger the size of the group, the lower the chance of pedestrians imitating others.

Demographics. Arguably, *gender* is one of the factors that influences pedestrian behavior the most [137, 153, 154]. Studies show that women in general are more cautious than men [137, 153, 151] and demonstrate a higher degree of *law compliance* [128, 155].

Furthermore, *gender* differences affect the motives of pedestrians when complying with the law. Yagil [151] argues that crossing behavior in men is mainly predicted by normative motives (the sense of obligation to the law) whereas in women it is better predicted by instrumental motives (the perceived danger or risk). He adds that women are influenced by social values, e.g. what people think about them, while men mainly care about physical conditions, e.g. *road structure*.

Men and women differ in the way they pay *attention* to the environment before or during the crossing. For instance, Tom and Granie [128] show that prior to and during a crossing event, men more frequently look at vehicles whereas women look at traffic lights and other pedestrians, i.e. they have different *attention* patterns. Women also change their gaze pattern according to *road structure*, show a higher behavior variability [153], and cross with a lower *speed* compared to men [156].

Age impacts pedestrian behavior in obvious ways. Generally, elderly pedestrians are less physically capable compared to adults, and as a result, they walk slower [156], have a more varied *walking pattern* (e.g. do not have steady velocity) [157] and are more cautious in terms of *gap acceptance* [139, 158]. Being more cautious means that older pedestrians, compared to adults and children, spend longer time paying *attention* to the traffic prior to crossing. Furthermore, the elderly and children are found less able to correctly assess the speed of vehicles, hence they are more vulnerable [132]. It is also interesting to note that there is a higher variability observed in younger pedestrians' behavior, making them less predictable [153].

State. The *speed* of pedestrians is thought to influence their visual perception of dynamic objects. Oudejans *et al.* [159] argue that while walking, pedestrians have better optical flow information, and consequently, a better sense of *speed and distance estimation*. Thus walking pedestrians are less conservative to crossing compared to standing ones.

Pedestrian *speed* may vary depending on the conditions such as *road structure*. For instance, pedestrians tend to walk faster during crossing compared to when they walk on sidewalks [160] and walk faster on wider sidewalks where the density of pedestrians can be lower [154]. When vehicles have *the right of way* or pedestrians' *trajectory* is towards the vehicles, they tend to cross faster [160]. In addition, *road structure* impacts crossing

speed. For example, Crompton [161] reports pedestrian mean speed at different crosswalks as follows: 1.49 m/s at *zebra crossings*, 1.71 m/s as crossing with pedestrian refuge island and 1.74 m/s at pelican crossings.

Other factors that have been shown to affect pedestrian *speed* include *group size*, generally slower in larger groups, [136, 162, 163], *age*, pedestrians tend to get slower as they age, [164, 163], *time of day*, generally walk faster in early morning rush, and *road structure*, if there is more space for pedestrians, they tend to walk faster [163].

The effect of *attention* on traffic safety has been extensively studied in the context of driving [165, 166, 167, 168]. As for pedestrians, the majority of them tend to pay *attention* prior to crossing, the frequency of which may vary depending on the presence of traffic *signals* or *zebra crossing* lines at the crosswalk. A study by Geruschat *et al.* [133] shows that the type of objects pedestrians pay *attention* to may vary depending on their *speed*, *law compliance*, *age* and *road structure*. For example, while moving, pedestrian subjects primarily fixated on crossing elements, and when standing at the curb, on cars. In addition, pedestrians who were crossing early against the light were looking at the cars whereas others were focusing on the traffic light. Some findings suggest that when pedestrians make eye contact with drivers, the drivers are more likely to slow down and yield [169].

Hymann *et al.* [170] investigate the effect of *attention* on pedestrian walking *trajectory*. They show that pedestrians who are distracted by the use of electronics, such as mobile phones, are 75% more likely to display inattentional blindness (not noticing the elements in the scene). Distracted pedestrians often change their walking direction and, on average, walk slower than undistracted pedestrians.

Trajectory or pedestrian walking direction is another factor that plays a role in the way pedestrians make a crossing decision. Schmidt and Farber [130] argue that when pedestrians are walking in the same direction as the vehicles, they tend to make riskier decisions regarding whether to cross. According to the authors, walking direction can alter the ability of pedestrians to estimate speed. In fact, pedestrians have a more accurate *speed estimation* when the approaching cars are coming from the opposite direction.

Characteristics. Among different pedestrian characteristics, *culture* plays an important role. It defines the way people think and behave and forms a common set of *social norms* they obey [171]. Variations in traffic *culture* exist not only between different countries but also within the same country, e.g. between towns and countrysides or cities [172].

A number of studies connect *culture* to the types of behavior that road users exhibit. Lindgren *et al.* [171] compare the behaviors of Swedish and Chinese drivers and show that they assign different levels of importance to various traffic problems such as speeding or jaywalking. Schmidt and Farber [130] point out the differences in *gap acceptance* of Indians

who on average cross between 2 to 8s whereas Germans cross between 2 to 7s time to collision. Clay [132] indicates the way people from different cultures perceive and analyze a situation. She notes that Americans judge traffic behavior based on characteristics of the pedestrians whereas Indians rely more on contextual factors such as traffic conditions, road structure, etc.

Some researchers go beyond *culture* and study the effect of *faith* or religious beliefs on pedestrian behavior. Rosenbloom *et al.* [173] gather that ultra-orthodox (in a religious sense) pedestrians in an ultra-orthodox setting are three times more likely to violate traffic laws than secular pedestrians.

Generally speaking, pedestrian level of *law compliance* defines how likely they would break the law (e.g. crossing at a red light). In addition to demographics, *law compliance* can be influenced by physical factors, for instance, the *location* of a designated crosswalk influences the decision of pedestrians whether to jaywalk [174].

Another factor that characterizes a pedestrian is his/her *past experience*. For example, non-driver female pedestrians generally tend to be more cautious when making crossing decisions [153].

Abilities. The ability to *estimate speed and distance*, can influence the way pedestrians perceive the environment and consequently the way they react to it. In general, pedestrians are better at judging *vehicle distance* than *vehicle speed* [175]. For instance, they can correctly estimate *vehicle speed* when the vehicle is moving below the speed of 45 km/h, whereas *vehicle distance* can be correctly estimated when the vehicle is moving up to a speed of 65 km/h.

Environmental Factors

Physical context. The presence of traffic *signals* or *zebra crossings*, has a major effect on the way traffic participants behave [154], or on their degree of *law compliance* [176]. Some scholars distinguish between the way traffic *signals* and *zebra crossings* influence yielding behavior. For example, traffic signals (e.g. traffic lights) prohibit vehicles to go further and force them to yield to crossing pedestrians. At non-signalized *zebra crossings*, however, drivers usually yield if there are pedestrians present at the curb who either clearly communicate their intention of crossing (often by eye contact) or start crossing (by stepping on the road) [112].

Signals can alter pedestrians' level of cautiousness. In [128], the authors show that pedestrians look at vehicles 69.5% of the time at signalized and 86% of the time at unsignalized intersections. In addition, the authors point out that pedestrians' *trajectory* differs at unsignalized crossings, i.e. they tend to cross diagonally when no signal is present.

Some studies discuss the likelihood of pedestrians to use dedicated *zebra crossing*. In general, women and children use dedicated zebra crossings more often [154, 155]. *Traffic volume* and the presence of *law enforcement* personnel near crossing lines are also shown to induce pedestrians to use designated crossing lines. The effect of *law enforcement*, however, is much stronger on men than women [154].

In terms of crossing *speed*, pedestrians tend to walk faster at signalized crosswalks [177, 176]. The presence of signals also induces pedestrians to comply with the law, although this effect seems to be opposite for one-way streets [178].

Road structure (e.g. crossing type and road geometry) and *street width* impact the level of crossing risk (or affordance) [159]. For example, pedestrians pay more *attention* prior to crossing in wide streets and accept a smaller gap in narrow streets [130]. *Road structure* is also believed to alter the way drivers behave, which subsequently can influence pedestrians' expectations [172].

With respect to *law compliance*, contradictory findings have been reported. While some researchers claim that larger *street width* can increase the chance of compliance [179], others report the opposite and show that it can increase crossing violation [178].

Weather or *lighting* conditions affect pedestrian behavior in many ways [144]. For instance, in bad *weather* conditions pedestrians' *speed estimation* is poor, therefore they become conservative while crossing [175]. Pedestrians (especially the elderly and women) are found to be more cautious in warm weather than cold [144]. Moreover, lower illumination level (e.g. nighttime) reduces pedestrians' major visual functions (e.g. resolution acuity, contrast sensitivity, and depth perception), causing them to make riskier decisions. Another direct effect of *weather* would be on *road conditions*, such as slippery roads due to rain, that can impact movements of both drivers and pedestrians [180, 154].

Dynamic factors. One of the key dynamic factors is *gap acceptance* or how much gap in traffic (typically in time) pedestrians consider safe to cross. *Gap acceptance* depends on two dynamic factors, *vehicle speed* and *vehicle distance* from the pedestrian. The combination of these two factors defines Time To Collision (or Contact) (TTC), or how far the approaching vehicle is from the point of impact [181, 182]. The average pedestrian *gap acceptance* is between 3 and 7 seconds, i.e. usually pedestrians do not cross when TTC is below 3s [136] and very likely cross when it is higher than 7s [130]. As mentioned earlier, *gap acceptance* may highly vary depending on social factors (e.g. *demographics* [140, 183], *group size* [136], *culture* [130]), level of *law compliance* [156], and the *street width*. For instance, women and the elderly generally accept longer gaps [184] and people in groups accept a shorter time gap [183].

The effects of *vehicle speed* and *vehicle distance* are also studied in isolation. It is shown

that increase in *vehicle speed* deteriorates pedestrians' ability to estimate speed [132] and distance [175]. In addition, pedestrians are found to rely more on distance when crossing, i.e. within the same TTC, and they cross more often when the speed of the approaching vehicle is higher [130].

Some scholars look at the relationship between pedestrian *waiting time* prior to crossing and *gap acceptance*. Sun *et al.* [139] argue that the longer pedestrians wait, the more frustrated they become and, as a result, their *gap acceptance* lowers. The impact of *waiting time* on crossing behavior, however, is controversial. Wang *et al.* [140] dispute the role of *waiting time* and claim that in isolation *waiting time* does not explain the changes in *gap acceptance*. They add that to be considered effective, *waiting time* should be studied in conjunction with other factors such as pedestrians' personal characteristics.

Pedestrian *waiting time* can be influenced by a number of factors such as *age*, *gender*, *road structure*, *location* (e.g. how close to one's destination) and pedestrian *walking speed*. Females generally have longer *waiting time* compared to men [136, 185]. Pedestrians who can walk faster (which is affected also by *age*) tend to spend less time waiting prior to crossing [185]. As for *road structure*, studies show that, when crossing a road with a refuge island, pedestrians cross faster from one side to the island than the island to the other side.

Although *traffic flow* is a byproduct of *vehicle speed and distance*, on its own it can also be a predictor of pedestrian crossing behavior [130]. By observing the overall pattern of traffic, pedestrians might form an expectation about what approaching vehicles might do next.

Communication In road traffic, any kinds of signals transmitted between road users constitute communication. In this context, communication is particularly precarious because, firstly, there exists no official set of signals and most of them are ambiguous, and secondly, the type of communication may change depending on the atmosphere of the traffic situation, e.g. city or country [127]. For example, in a case study by Varhelyi [186] it is shown that drivers maintain their speed or accelerate to communicate their intention of not yielding to pedestrians. This means pedestrian reaction (or intention of crossing) may vary depending on the behavior of drivers. The stopping behavior of vehicles may also contain a communicational cue. Studies show when drivers stop their cars far shorter than where they legally must stop, they are signaling their intention of giving the *right of way* to others [187].

Drivers also often make eye contact and gaze at the face of other road users to assess their intentions [188]. It is found that the presence of eye contact between road users increases compliance with instructions and rules. For instance, drivers who make eye contact with pedestrians will more likely yield right of way at crosswalks [189].

When speaking of *communication*, two additional factors should be considered, namely

culture and *social norms* which determine the type and the meaning of communication signals used by road users [77]. For example, Gupta *et al.* [190] show how in Germany police officers raise one hand to convey attention command, whereas in India this would be communicated by raising both hands.

Traffic characteristics. *Traffic volume* or density affects pedestrian [150] and driver behavior [130] significantly. Essentially, the higher the density of traffic, the lower the chance of pedestrians to cross [156]. This is particularly true when it comes to *law compliance*, i.e. pedestrians are less likely to cross against the *signal* (e.g. red light) if the traffic volume is high. The effect of *traffic volume*, however, is stronger on male pedestrians than women [151].

The effects of vehicle characteristics such as *vehicle size* and *vehicle color* on pedestrian behavior have been investigated. Although *vehicle color* has not shown to have a measurable effect, *vehicle size* can influence crossing behavior in two ways. First, pedestrians tend to be more cautious when facing a larger vehicle [182]. Second, the size of the vehicle impacts pedestrian *speed and distance estimation* abilities. In an experiment involving 48 men and women, Caird and Hancock [191] reveal that as the size of the vehicle increases, there is a higher chance that people will underestimate its arrival time.

When making a crossing decision, the *vehicle type* matters and can influence different *genders* differently. For example, compared to women, men are generally better in judging the type of vehicles and are more accurate at estimating the arrival time of vans and motorcycles [191]. In addition, pedestrians exhibit different *waiting time* when facing different types of vehicles, e.g. they tend to cross faster in front of passenger vehicles [185].

A summary of the factors from the classical literature is illustrated in Figure 3.5. Here we can see that more studies have been conducted on factors such as *gender*, *group size*, *age* and *gap acceptance*, compared to *culture*, *vehicle size*, *right of way*, and *faith*. Due to the emergence of intelligent transportation systems and the availability of technology for collecting data, studies on factors such as *communication*, *attention*, *pedestrian trajectory* and *culture* have gained popularity in the past few years. However, a number of factors such as *lighting*, *road conditions*, *vehicle type*, *past experience*, *social status*, and *pedestrian flow* are left unaddressed in recent works.

It should be noted that understanding the factors that influence pedestrian behavior has two important applications: First, factors such as *lighting* conditions, *road structure*, *signals*, etc. can potentially lead to the design of better roads and intersections, resulting in safer crossing conditions for both drivers and pedestrians. Second, understanding these factors can shape drivers' expectations and their abilities to predict pedestrian behavior under various conditions. Consequently, the same understanding of pedestrian behavior can directly be

Figure 3.5: A circular dendrogram of the factors influencing pedestrian behavior and the classical studies that identified them. Leaf nodes represent the individual studies (identified by the first author and year of publication) and internal nodes represent minor and major factors.

used in the design of autonomous driving systems.

3.3.2 Studies in the Context of Autonomous Driving

Similar to classical studies, we divide behavioral studies involving autonomous vehicles into two groups of pedestrian and environmental factors. A summary of these factors and their

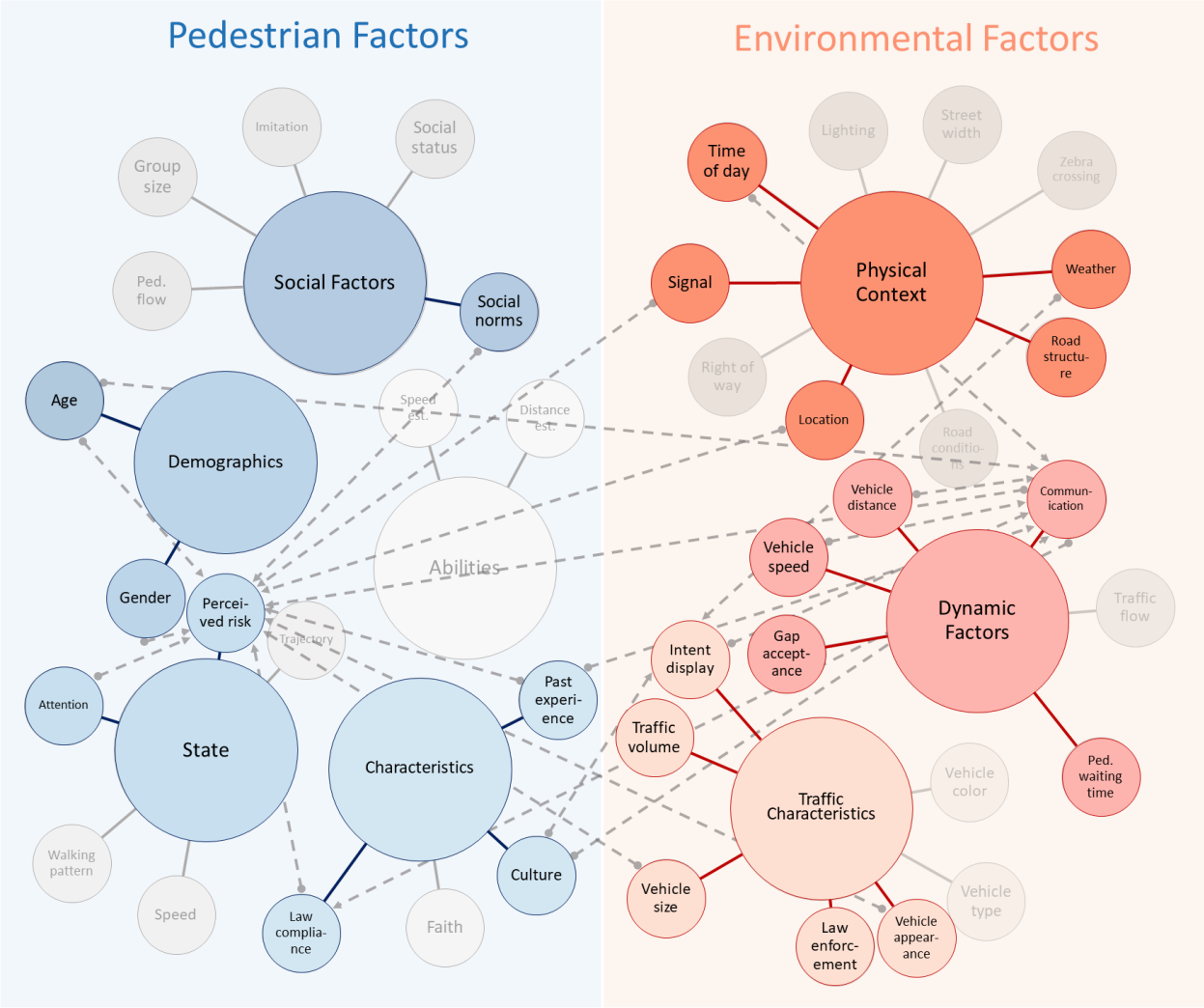


Figure 3.6: Factors involved in pedestrian decision-making process when facing autonomous vehicles. The diagram is based on the meta-analysis of the past literature. The large circles refer to the major factors and small circles connected with solid lines are sub-factors. The dashed lines show the interconnection between different factors and arrows show the direction of influence. The grey faded diagram at the background shows the factors from classical studies.

connections can be found in Figure 3.6.

Studies concerning the social aspects of autonomous driving generally focus on two major factors, namely *communication* and *attention*. The focus of many of these studies is on designing effective interfaces to transmit the intentions of AVs to other road users. Matthews *et al.* [192] measure the importance of using an *intent display* in *communication* with pedestrians. The authors used a remotely controlled golf cart with and without an intent display mechanism. They observed that when the vehicle equipped with a display was encountering pedestrians, there was 38% improvement in resolving deadlocks. The authors show

that further improvements can be achieved based on the pedestrians' *past experience*. The group of participants who were familiarized with the communication technology prior to the experiment exhibited more trust in the vehicle.

Although *intent displays* have been shown to improve the overall experience of pedestrians during interaction [192, 193], they don't always seem to be very effective. In her studies, Yang [194] used a display to show "Safe to Cross" message to pedestrians. When interviewed by the experimenter, the participants responded that the display did not have a significant effect on their crossing decisions. In another study, Clamann *et al.* [195] found that pedestrians still focus on legacy factors such as *vehicle speed and distance* when making crossing decisions. The use of the display only influenced 12% of the participants' decisions and overall increased the time of decision-making. In this context, however, the authors show that informative displays (e.g. with information about vehicle's speed) compared to advisory displays (e.g. cross or not to cross signal) are more effective. The authors add that the traditional social and environmental factors such as *age*, *gender road structure*, *waiting time* and *traffic volume* are still very important in the context of autonomous driving. In a study by Pillai [196], the author similarly concludes that pedestrians mainly rely on implicit behavior of the vehicle to make crossing decisions, however, under certain circumstances, e.g. under *weather* conditions with poor visibility, additional *intent display* mechanisms such as audio signals can be very effective. The author adds that *culture* plays an important role in the design of communication interfaces.

Other forms of *intent display* methods have also been examined. Chang *et al.* [197] propose the use of moving eyes installed at the front of the vehicles. Using experimental data collected from 15 participants, the authors show that more than 66% of participants made street crossing decisions faster in the presence of eyes, and if the eyes were looking at the participants, this number rose to more than 86%. The empirical evaluation of this study, however, is limited to virtual reality environments without any direct risk of accident.

Mahadevan *et al.* [198] investigate various modalities of *communication* such as audio, visual, motion, etc. The authors note that in the absence of an explicit *intent display* mechanism, pedestrians rely on *vehicle speed and distance* to make crossing decisions. As for different means of *communication*, pedestrians generally prefer LED sequence signals to LCD displays and other modalities of communication such as auditory and physical cues. The authors show that the use of human-like features for communication such as animated faces on displays was not well-received by the participants. Overall, the authors recommend that a combination of modalities including visual, physical and auditory should be considered. They point out that there is no limit on where the informative cues are located and can be either on the vehicle or in the environment. It should be noted that although this study is

Figure 3.7: The vehicles used in [131], an aggressive looking BMW (*left*) and a friendly looking Renault (*right*).

very thorough in terms of evaluating different design approaches, its scope is very limited. Only 10 subjects participated in the final phase of the study (Wizard of Oz phase) and the participants were all North American. Furthermore, the authors admit that *culture* can play a very important role in the modality and type of communication preference.

Implicit forms of communication such as the vehicle's motion pattern (*speed and distance*) have also been investigated. Zimmerman *et al.* [193] show that abrupt acceleration behavior and short stopping distance by autonomous vehicles can be perceived as erratic behavior by pedestrians and negatively influence their crossing decisions. The authors suggest that to be effective, a well-balanced acceleration and deceleration with sufficient distance to other road users should be used by autonomous vehicles. In another study, Beggiato *et al.* [199] examine the effect of the vehicle's braking action whereby the vehicle can communicate its intention. The authors argue that the interpretation of the signal may vary with respect to other factors such as *time of day*, *vehicle speed*, and *age*. For instance, older pedestrians generally make more conservative crossing decisions when the *vehicle speed* is lower.

Moving away from *communication*, the authors of [125] and [200] argue that the *perceived risk* of autonomous vehicles may vary depending on pedestrians' *age*, *gender*, *past experience*, level of *law compliance*, *location*, and *social norms*. For example, younger male pedestrians, people with higher acceptance for innovation and people living in urban environments are more receptive of autonomous driving technology. People with traffic violation history also tend to be more comfortable when crossing in front of autonomous vehicles.

Dey *et al.* [131] evaluate the impact of *vehicle type* on the *perceived risk* of autonomous vehicles. The authors use two different types of vehicles, a BMW with an aggressive look and a Renault with a friendlier look (see Figure 3.7). They report that the *vehicle speed and distance* compared to *vehicle size* and *appearance* play a more dominant role in making a crossing decision. Apart from dynamic factors, roughly 30% of the participants claimed that they merely relied on the behavior of the car when making a crossing decision, whereas the

rest mentioned that *vehicle size* was important to them rationalizing that the smaller the vehicle, the higher their chance of moving out of its way. The majority of the participants agreed that the friendliness of the vehicle design did not factor in their decision-making process.

Evaluating the impact of autonomous vehicle behavior on pedestrian crossing, Jayaraman *et al.* [201], argue that the presence of traffic *signals* at crosswalks has little impact on pedestrian crossing decisions and that decisions are highly determined by autonomous vehicle's driving behavior. The implication of these findings, however, is limited because the evaluation was performed only in a virtual reality environment.

Figure 3.8 summarizes all of our findings on pedestrian behavior studies involving autonomous vehicles. At first glance, we can see that, compared to classical studies, pedestrian behavior in the context of autonomous driving is fairly understudied. The majority of research currently focuses on the role of *communication*, *intent display*, *perceived risk* and *attention*, while factors such as *signal*, *location*, *road structure*, *gap acceptance*, and *social norms* are rarely addressed. More importantly, some of the factors widely studied in classical works, namely *group size*, *pedestrian speed*, and *street width*, have not been evaluated in the context of autonomous driving. As was mentioned earlier, these factors significantly impact the way pedestrians make crossing decision. This means the lack of considerations for such factors in the design of autonomous systems can lead to misjudgment of pedestrian behavior, and consequently result in accidents or overly cautious behavior that may interrupt the flow of traffic.

3.4 What Should be Done Next

As we saw in Section 3.3.1 pedestrian behavior in the context of traffic interactions is a very well studied field. In recent years, studies of similar nature in the context of autonomous vehicles have gained momentum, however, the number of these studies is still relatively small, compared to classical studies (see Figure 3.8). Perhaps, one contributing factor is the lack of means, such as pedestrian questionnaires or validated simulators [202], that can aid the study of pedestrian behavior in the context of autonomous driving systems. Although classical studies have a number of implications for autonomous driving systems, it is reasonable to expect that pedestrians might behave differently when facing autonomous vehicles. This means that more studies of similar nature to classical studies have to be conducted involving autonomous vehicles. To achieve this, the following elements should be considered in the study of pedestrian behavior and the development of pedestrian intention estimation algorithms.

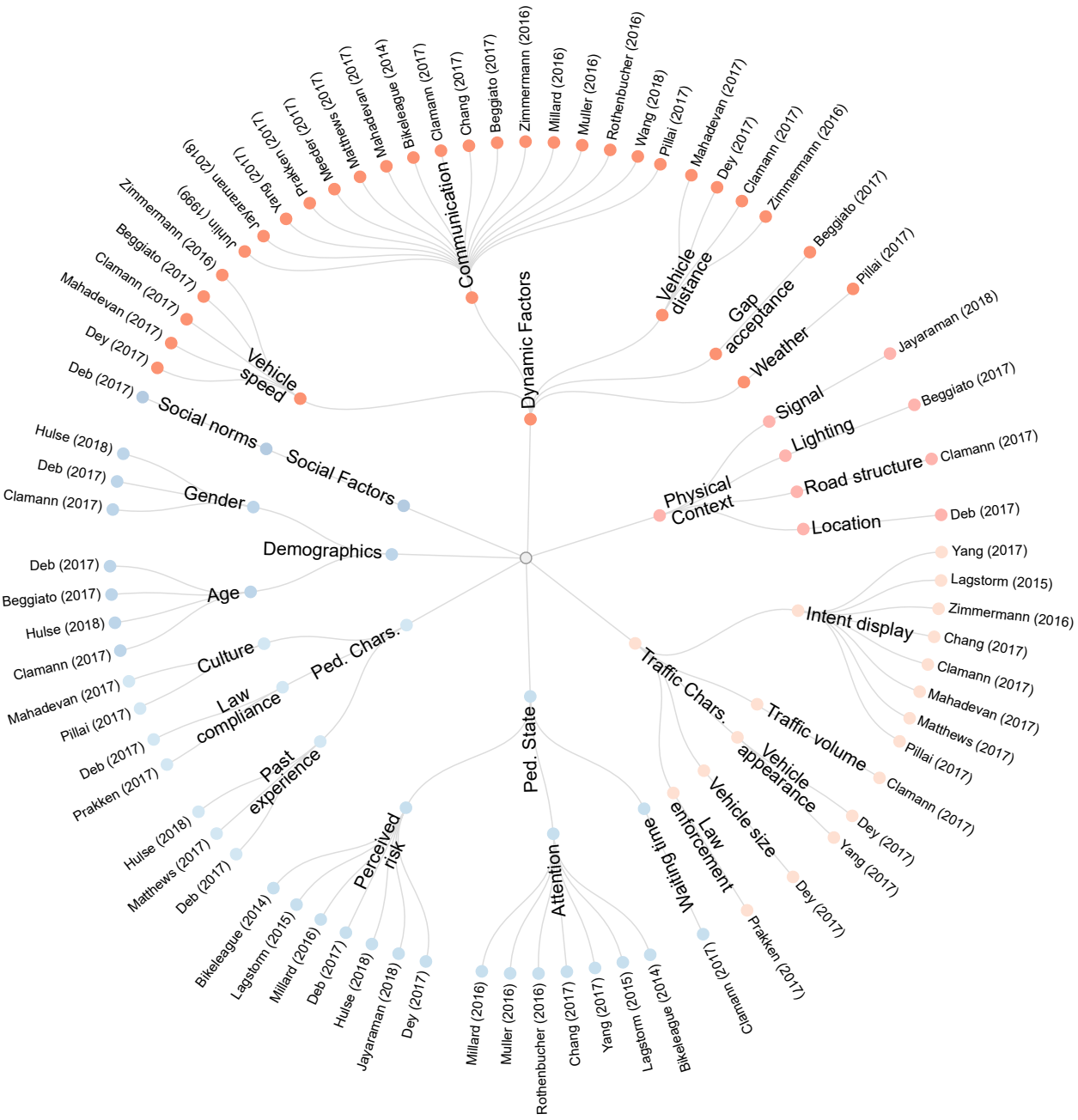


Figure 3.8: A circular dendrogram of the factors influencing pedestrian behavior and the autonomous driving studies that identified them. Leaf nodes represent the individual studies (identified by the first author and year of publication) and internal nodes represent minor and major factors.

Holistic vs focused studies. Pedestrian behavior studies often are conducted on a small subset of factors in traffic. As our meta-analysis of the literature shows that there are strong interrelationships between factors that influence pedestrian behavior (see Figure 3.4). This means that studying only a small subset of these factors may not capture the true

underlying reasons behind pedestrian crossing decisions. Therefore to avoid fallacies when reasoning about pedestrian behavior, studies have to be multi-modal and account for chain effects that factors might have on each other.

Social norms should be the focal point. We found a general consensus in the literature regarding the impacts of some of the factors that influence pedestrian behavior, e.g. how group size influences gap acceptance or how individuals behave based on their demographics. However, we noticed that the results presented by some of the studies are contradictory especially the ones on topics such as communication, the influence of imitation, the role of attention, waiting time influence on gap acceptance, etc. Although some of these contradictions can be explained by the differences in the methods of studies, we believe that the main reason is the variations in *social norms* and *culture*. These studies often are conducted in different geographical locations where culture and social norms can be quite different. This means that these studies should be reproduced in different regions to account for cultural differences.

Large scale studies are needed. Unfortunately, the scope of the majority of behavioral studies involving autonomous vehicles is relatively limited, both in terms of sample size (often less than 100) and demographics of participants (e.g. university students). As a result, some of these studies have reported very contradictory findings, for instance, regarding the need for communication or pedestrians' need to engage in eye contact. To be useful for the design of interactive autonomous vehicles, these works have to be conducted on a much larger scale and demographically diverse population, and of course, they should follow the same considerations as classical behavior studies.

Time changes everything. Changes in socioeconomic and technological factors also influence traffic behavior. For example, compared to the 1950s or 1960s when early behavioral studies had been conducted, today's vehicles are much safer, roads are built and maintained better, the number of vehicles and pedestrians have increased significantly, and traffic laws have been changed, all of which affect traffic dynamics. To account for current pedestrian behavior, some of these studies have to be repeated. The same is also true for studies involving autonomous vehicles. Today, the deployment of autonomous vehicles is very limited and the majority of pedestrians have not been exposed to them. As time goes by and more autonomous vehicles become available on roads, pedestrians' attitude towards them certainly would change.

Chapter 4

Pedestrian Communication in Traffic

4.1 Why is Communication Important in Traffic Context?

4.1.1 Communication in Traditional Traffic

In the previous section, we showed that communication is one of the contextual factors that influence pedestrian behavior. Communication is considered as one of the main factors in resolving traffic ambiguities [122, 132, 112]. In fact, the lack of communication or miscommunication is found to greatly contribute to traffic conflicts. It is shown that more than a quarter of traffic conflicts is due to the absence of effective communication between road users. In a study of traffic conflicts, it was found that out of conflicts caused by miscommunication, 47% of the cases occurred with no communication, 11% was due to the lack of necessary communication and 42% happened during communication [127]. In particular, pedestrians heavily rely on communication when making crossing decisions and report feeling uncomfortable when the communication is non-existent and certain vehicle behaviors are not observed [203].

4.1.2 Communicating with AVs

The autonomous driving community is divided about the necessity of *communication* with pedestrians. Millard [204] argues that the interaction between pedestrians and autonomous vehicles resembles, what he refers to as the game of "crosswalk chicken". In a normal situation involving a human driver, if a pedestrian chooses to cross, they accept a large risk because the norms permit not yielding to pedestrians, the driver might be distracted or assume the pedestrian would not intend to cross. According to Millard, in the case of autonomous

Figure 4.1: Driver's conditions used in the experiments conducted in [141].

driving the *perceived risk* of crossing is almost nonexistent because the pedestrian knows that the autonomous vehicle will stop, and as a result, there is no need for any form of *communication* to reach an agreement with the vehicle. Using field studies, Rothenbucher *et al.* [205] support the same argument and show that without *communication* and *attention* (the need for establishing eye contact), when facing an autonomous vehicle, pedestrians eventually adjust their behavior and cross the street. The result of this study, however, is questionable because the trials took place on a university campus where the speed limit was very low and the vehicle posed minimal threat to pedestrians. The subjects who were observed or participated in the interviews may also have heard about the experiment, or in general, had higher acceptance compared to the general population for autonomous driving technologies.

Overall, arguments in the favor of *communication* necessity in autonomous driving are stronger. A number of studies relate to existing literature and *past experience* to support the role of *communication* [206, 207, 208, 209, 210]. Muller [206] argues that identifying autonomous vehicles in traffic is not always intuitive. Road users might recognize an autonomous vehicle as a traditional vehicle and expect certain behaviors from it. As for the need for *communication*, the author describes a busy pedestrian crossing where a driver might communicate his intention by moving forward slowly into the crowd. The author then raises a concern regarding how an autonomous vehicle would behave in such a situation.

The *communication* necessity can also be seen from a different perspective. Prakken [208] argues that understanding communication cues in obeying traffic laws is important, but the current technology does not distinguish between the type of pedestrians which can be problematic when a *law enforcement* officer is present in the scene for directing the traffic. According to Prakken, autonomous vehicles should be able to interpret and distinguish communication messages produced by *law enforcement* personnel and regular pedestrians.

A number of empirical studies support the role of communication in autonomous driving.

A survey conducted by the League of American Bicyclists [211] shows that besides issues related to technological advancements, the inability to communicate and establishing eye contact are among major reasons that increase pedestrians and bicyclists *perceived risk* when interacting with autonomous vehicles.

Lagstrom and Lundgren [141], and, in a later study, Yang [194] evaluate the role of driver behavior when the vehicle is running autonomously. The authors used several scenarios of driver behavior when crossing an intersection including the driver making eye contact, staring straight at the front road, talking on the phone, reading a newspaper and sleeping (see Figure 4.1). In these experiments, the vehicles were operated by drivers (who were hidden from the view of pedestrians) using a right-hand steering wheel. Observing pedestrians' reactions, Lagstrom and Lundgren show that when the vehicle was stopping and the driver paid *attention* (made eye contact) to pedestrians, all pedestrians crossed the street. However, when the driver was busy on the phone, 20% of pedestrians did not cross and when the driver was reading a newspaper or not present in the vehicle, 60% of the pedestrians did not cross. In both studies surveys were conducted to measure the pedestrians' level of *perceived risk* in each situation. The results show that when a form of *attention* (eye contact) was present, the pedestrians felt most comfortable. Yang [194] also adds that *vehicle appearance* impacts the level of pedestrians' comfort. Her findings indicate that when the pedestrians could not see the driver (due to dark windows), they felt most uncomfortable.

4.2 Nonverbal Communication: How the Human Body Speaks to Us

Communication in traffic scenes is mainly nonverbal through the use of hand gestures, changing gaze direction or any postural movements. In general, nonverbal communication refers to communication styles that do not include the literal verbal content of communication [212], i.e. it is affected by means other than words [213]. Buck and Vanlear [214] argue that nonverbal communication comes in three types (see Figure 4.2):

1. Spontaneous: This form is based on a biologically shared signal system and nonvoluntary movements. Spontaneous communication may include facial expressions, micro gestural movements, and postures.

2. Symbolic communication: This type of communication is deliberate and has an arbitrary relationship with its referent and knowledge of what should be shared by the sender and receiver. For instance, symbolic communication may include a system of sign language, body movements or facial expressions associated with language.

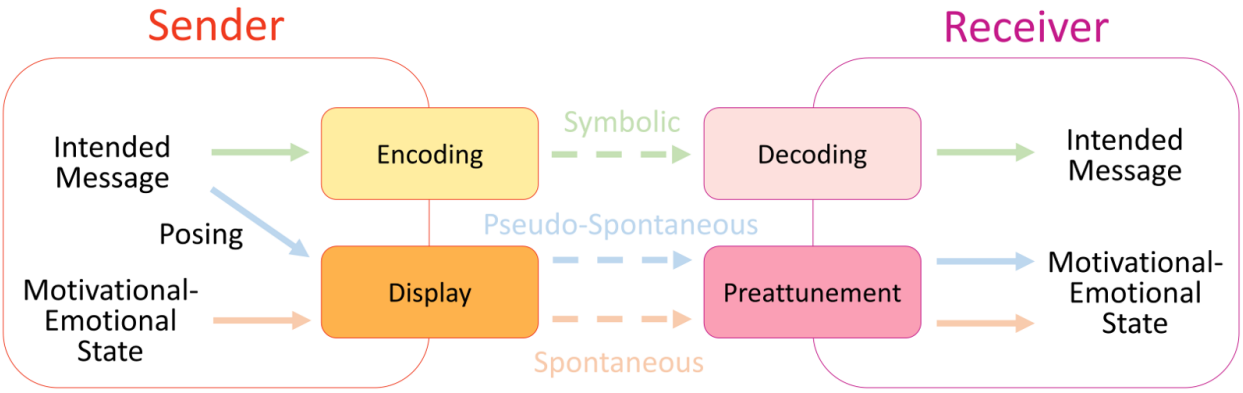


Figure 4.2: A simplified view of nonverbal communication forms. Symbolic communication is about transmitting an intended message to the receiver, for example, instructing the receiver to do a task. The other two forms of communication reflect (or posed to reflect) the internal state of the sender in order to change the emotional state of the receiver, e.g. acting. Source: [214]

3. Pseudo-spontaneous: This form involves the intentional and propositional manipulation by the sender of the expressions that are virtually identical to spontaneous displays from the point of view of the receiver. This may include acting or performing.

In the traffic context all three types of nonverbal communication are observable. It is intuitive to imagine the occurrence of the first two types of communication. For example, pedestrians may perform various spontaneous movements including yawning, scratching their head, stretching their muscles, etc. As for symbolic communication, humans use various forms of nonverbal signals to transmit their intentions such as waving hands, nodding, or any other forms of bodily movements. Symbolic communication in traffic interactions will be discussed in more detail in Section 4.3.

4.2.1 Studies of Nonverbal Communication

The modern study of nonverbal communication is dated back to the late 19th century. Darwin, in his book *Expression of the Emotions in Man and Animals* [215], was the first to focus on the possible modifying effects of body and facial expressions in communication. Darwin argues that nonverbal expressions and bodily movements have specific evolutionary functions, for instance, wrinkling the nose reduces the inhalation of bad odor.

In more recent studies, behavioral ethologists point that in humans, throughout their evolutionary history, these nonverbal bodily movements had gained communicative values [216]. In fact, it is estimated that 55% of communication between humans is through facial expressions [217]. According to Birdwhistell [218], humans are capable of making and

recognizing about 250,000 different facial expressions.

Scientists in behavioral psychology measured the importance of bodily movements in various interaction scenarios. For example, Dimatteo *et al.* [219] show that the ability to understand and express emotions through nonverbal communication significantly improves the level of patient satisfaction in a physician visit.

Comprehension or expression of nonverbal communication is linked to various factors. For instance, in a work by Nowicki and Duke [220], it is shown that the accuracy of emotional comprehension increases with age and academic achievement. Gender also plays a role in nonverbal communication. In general, women are found to engage in eye contact more often than men [221] and are also better at sending and receiving nonverbal signals [212]. Another important factor is culture which determines how people engage in nonverbal communication. For example, in Western culture, eye contact is much less of a taboo compared to Middle Eastern culture [221].

4.2.2 Methods of Studying Nonverbal Communication

Behavioral responses of subjects are measured by showing them a sequence of videos of human faces. Then the subjects are either asked about how comfortable they feel making eye contact with the human in the picture [221] or their emotions are directly observed [220]. In another method known as Profile of Nonverbal Sensitivity (PONS), in addition to the assessment of emotions, participants are asked to express certain emotions. The recorded expressions are then shown to independent observers who are asked to identify the emotions they represent, for example, whether they imply sadness, anger or happiness. The final score is the combination of both assessment and expression of emotions by the participants [219]. In some studies, fMRI is used to measure brain activities of participants during nonverbal communication [222].

In the context of autonomous driving, however, communication is mainly studied through naturalistic observations [112, 205]. The observation is sometimes combined with other methods to minimize subjectivity. For instance, pedestrians are instructed to perform a certain behavior, e.g. engage in eye contact, and then the behavior of the drivers (who are unaware of the scenario) are observed naturalistically [189]. The observees sometimes are interviewed to find out how they felt regarding the communication that took place between them and the other road users [127].

4.2.3 Eye Contact: Establishing Connection

Eye contact, perhaps, is the most important part and the foundation of communication and social interaction. In fact, scientists argue that eye contact creates a phenomenon in the observer called "eye contact effect" which modulates the concurrent and/or immediately following cognitive processing and/or behavioral response [222]. Putting it differently, direct eye contact increases physiological arousal in humans, triggering the sense of trying to understand the other party's intention by asking questions such as "why are they looking at me?" [119].

Depending on the context, in the course of social interaction, eye contact may serve different functions, which according to Argyle and Dean [221] can be one of the following:

1. Information-seeking: It is possible to obtain a great deal of feedback by careful inspection of other's face, especially in eye region. Various mental states such as noticing one, desire, trust, caring, etc. can be interpreted from the eyes [119].

2. Signaling that the channel is open: Through eye contact a person knows that the someone is attending to him, therefore further interaction is possible.

3. Concealment and exhibitionism: Some people like to be seen, and eye contact is evidence of them being seen. In contrast, some people don't like to be seen, and eye contact makes them feel depersonalized.

4. Establishment and recognition of social relationship: Eye contact may establish a social relationship. For example, if person A wants to dominate person B, he stares at B with the appropriate expression. Person B may accept person A's dominance by a submissive expression or deny it by looking away.

5. The affiliative conflict theory: People may engage in eye contact for both approaching or avoiding contact with others.

Since the communication between road users is a form of social interaction, eye contact in traffic scenes might serve any of the functions mentioned above. However, in the context of traffic interaction, the first two functions are particularly important. In most cases, prior to crossing, pedestrians assess their surroundings to check the state of approaching vehicles, traffic signals or road conditions. Likewise, drivers continuously observe the road for any potential hazards. It is also common that pedestrians engage in eye contact with drivers to transmit, for example, their intention of crossing.

Figure 4.3: The function of hand gesture depending on its lexical meaning.

4.2.4 Understanding Motives Through Bodily Movements

Besides eye contact, humans often rely on other forms of bodily movements for further message passing. For instance, hand gestures are commonly used during both nonverbal and verbal communication. Although all hand gestures are hand movements, all hand movements are not necessarily hand gestures. This depends on the movement and how the movement is done. Krauss *et al.* [216] group hand gestures into three categories (see Figure 4.3):

1. Adapters: These are also known as body-focused movements or self-manipulation. These are the types of gestures that do not convey any particular meaning and have pure manipulative purposes, e.g. scratching, rubbing or tapping.

2. Symbolic gestures: Purposeful motions to transfer a conversational meaning. Such motions are often presented in the absence of speech. Symbolic gestures are highly influenced by cultural background.

3. Conversational gestures: These are hand movements that often accompany speech.

In addition to hand gestures, body posture and positioning may also convey a great deal of information regarding one's intention. Scheflen [223] lists three functionalities of postural configuration in different aspects of communication:

1. Distinguishes the contribution of individual behavior in group activities.

2. Indicates how the contributions are related to one another.

3. Defines steps and order in interaction.

4.3 Pedestrian Nonverbal Communication: An Empirical Study

Although some past works studied pedestrian communication in traffic scenes, a systematic investigation of methods and meaning of communication and factors impacting them are

Figure 4.4: Examples of communication forms demonstrated by pedestrians in traffic scenes.

missing. In this section, we discuss our empirical study of pedestrian communication in traffic scenes and present some of the findings.

4.3.1 Joint Attention in Autonomous Driving (JAAD) Dataset

We collected a dataset of 346 high-resolution video clips (5-15s) showing various situations typical for urban driving. These clips were extracted from approx. 240 hours of driving videos collected in several locations including North America (Toronto, New York) and Europe (Kremenchuk, Lviv, Hamburg). Two vehicles equipped with wide-angle video cameras were used for data collection. Cameras were mounted inside the cars in the center of the windshield below the rear view mirror. Some examples of pedestrians communicating from our dataset are shown in Figure 4.4.

The video clips represent a wide variety of scenarios involving pedestrians and other drivers. Most of the data is collected in urban areas (downtown and suburban), and only a few clips are filmed in rural locations. The samples cover a variety of situations such as pedestrians crossing individually, or as a group, pedestrians occluded by objects, walking along the road and many more. The dataset contains fewer clips of interactions with other drivers, and most of them occur in uncontrolled intersections, in parking lots or when another driver is moving across several lanes to make a turn.

The videos are recorded during different times of the day and under various weather

and lighting conditions. Some of them are particularly challenging, for example, those that include sun glare. The weather can also impact the behavior of road users, for example, during heavy snow or rain people wearing hooded jackets or carrying umbrellas may have limited visibility of the road. Since their faces are obstructed it is also harder to tell if they are paying attention to the traffic from the driver's perspective.

We attempted to capture all of these conditions for further analysis by providing two kinds of annotations for the data: bounding boxes and textual annotations. Bounding boxes are provided for some of the cars and all pedestrians. All bounding box annotations are done using Piotr's annotation toolbox [224].

We save the following data for each video clip: weather, time of the day, age and gender of the pedestrians, location and whether it is a designated crosswalk. For each pedestrian, we provide the following labels pedestrians: walking, standing, looking, moving, etc. Labels for moving and changes in moving speed are mutually exclusive, but they can overlap with all other labels such as crossing, attention, and gestures. The behavioral annotations are created using BORIS software [225].

We call the dataset described above Joint Attention in Autonomous Driving (JAAD)[1] [2]. Overall, JAAD has approximately 700 pedestrian samples with behavioral annotations. We selected 565 pedestrian samples from the JAAD dataset using two criteria: the full event of crossing or near crossing is observable and in the cases where no crossing takes place, we have high confidence that the pedestrian intends to cross. For the purpose of this study, we collected additional 211 pedestrian samples from video footage recorded under similar conditions as JAAD. This brings our total number of pedestrian samples up to 776.

The data contains 521 crossing events and 255 non-crossing events. There are 332 samples of male pedestrians, 433 female pedestrians, and 11 children. 656 samples of females and males are adults (approximately between the age of $15 - 65$) and 109 are seniors (over the age of 65).

4.3.2 Method

We focus our analysis on pedestrian communication patterns during the crossing and non-crossing events (situations in which the pedestrian intends to cross but does not do so in front of the recording vehicle). We base our study on the following factors: pedestrians' demographics (*gender and age*), *crosswalk type*, pedestrian *group size* (the number of pedestrians intending to cross at the same time on both sides of crosswalk), *street width*, and

[1]The dataset and more details regarding the annotations can be found at http://data.nvision2.eecs.yorku.ca/JAAD_dataset/.

[2]Collection of this dataset was approved by the York Ethics Committee with certificate # 2016-203.

driver's action. The crosswalk type is divided into *non-designated (ND)*, *pedestrian crossing (PC)*, where only zebra crossing is available and *signalized (S)*, where either a traffic signal or a stop sign is present.

4.3.3 Pedestrian Forms of Communication and Meaning

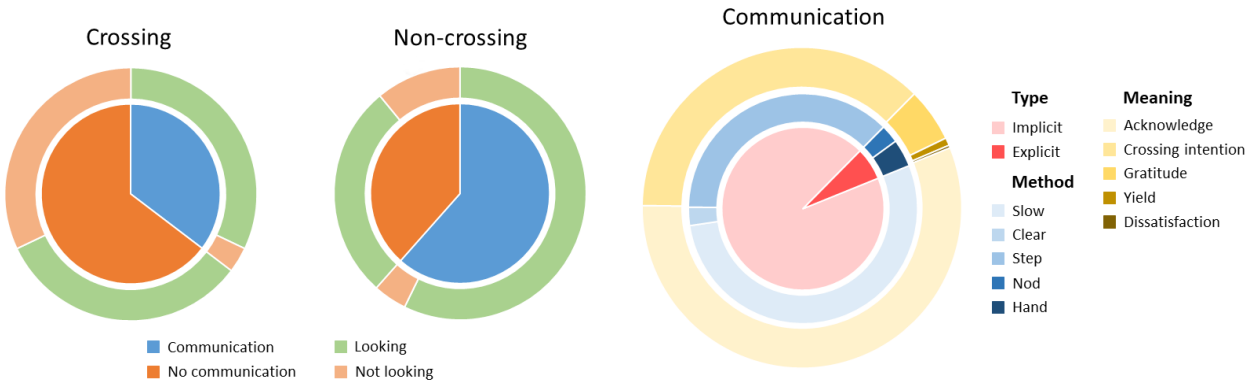


Figure 4.5: Frequency and types of communication, methods used to communicate and their meaning with respect to crossing and non-crossing events. *Left*, the frequency of communication in crossing and non-crossing events. The outer circle indicates how often a looking action was observed. For instance, in crossing events pedestrians communicated 34% of the time out of which 9% of the cases followed a looking action. *Right*, types (inner circle), methods (middle circle) and meaning (outer circle) of communication. For example, 6% of communication instances were explicit, out of which 60% involved hand gestures with the intention of yielding (20%), showing gratitude (73%) or dissatisfaction (7%).

Communication always follows a form of joint attention between the driver and pedestrian. In the context of driving, joint attention takes place when a pedestrian makes an eye-contact with the driver, and often on its own can be a strong indicator of pedestrian intention of crossing.

Other forms of communication between the driver and pedestrian can be either *implicit* or *explicit*. In our data, we observed three forms of implicit communication used by pedestrians, namely stepping onto the road *(step)*, indicating that pedestrian is intending to cross, clearing the vehicle's path *(clear)* and slowing down *(slow)* showing that the pedestrian is acknowledging the driver's action.

In comparison to implicit forms of communication, explicit communication, namely *hand gesture (hand)*, and *nodding (nod)*, is rarer and is used to signal gratitude/dissatisfaction regarding the driver's action or to ask for the right of way or yield to the driver. Figure 4.5 summarizes the types of communication observed in our data and their meaning.

4.3.4 How Often Pedestrians Communicate with Traffic

Overall, out of 776 samples, pedestrians communicated in 341 cases or about 44% of the time. In some events, we have observed multiple instances of communication, e.g. the pedestrian stepped onto the road, and once given the right of way, showed hand gesture as a sign of gratitude. Taking into account all instances of communication, we have 387 observations.

Communication frequency changes significantly depending on whether the pedestrian eventually crosses or not. In the event of crossing, in only 35% of the cases pedestrians engaged in some form of communication compared to 62% in non-crossing events.

It should be noted that in our analysis, two forms of communication, namely slowing down and clearing the vehicle's path, highly depend on the situation in which they have been observed. For instance, pedestrians slow down when they are approaching the crosswalk around the same time as the vehicle is passing, or they clear path if they are standing in the way of the vehicle. In these situations, the actions are in response to the driver's action.

After excluding pedestrian reaction cues, we observe communication in only 20% of the cases. Looking at the crossing and non-crossing events separately, communication had appeared in about 16% and 29% of the cases respectively.

We also found that communication in traffic scenes is predominantly implicit. Out of 387 instances of communication, we only observed 25 cases (6.5%) of explicit communication, out of which in only 3 cases the communication was instructive, i.e pedestrians were using hand gestures to yield to the driver.

4.4 Communication and Environmental Factors

4.4.1 When and Where do Pedestrians Look

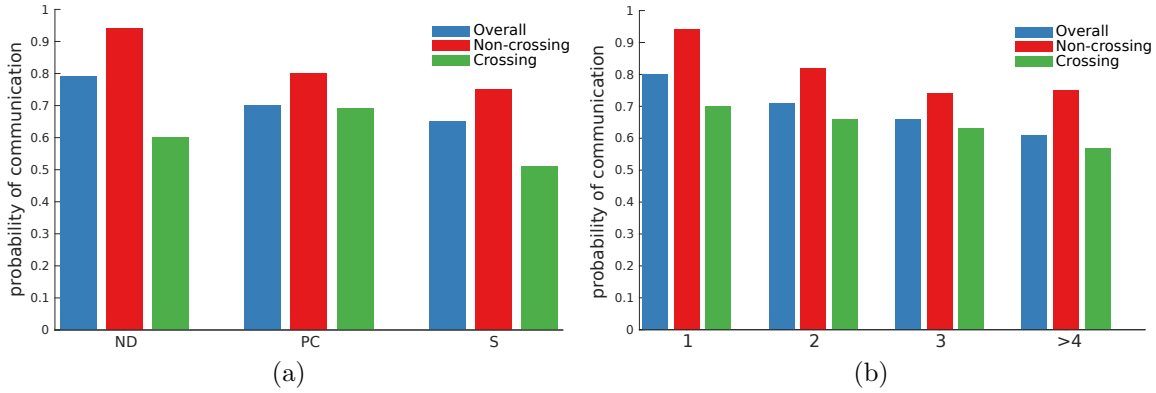


Figure 4.6: The effect of a) crosswalk type and b) pedestrian group size on looking frequency.

Figure 4.7: Examples of how only few pedestrians in larger groups look and the rest of the group follow.

Looking behavior is particularly important because it is an indicator of opening a communication channel between the pedestrian and the driver. Often looking suffices to convince the driver to give right of way. If it does not suffice, looking can be followed by a pedestrian's actions to request right of crossing by stepping onto the road or using hand gestures.

We have observed that at the point of crossing, pedestrians may look at traffic signals, other pedestrians, vehicles or drivers. The data collected from the driver's perspective showed us that pedestrians, even in the presence of traffic signals, often make eye-contact with drivers especially in cases where they are explicitly communicating with the driver or waiting for the driver's confirmation for giving right of way.

In terms of the frequency of looking behavior, we did not find any strong correlation between the pedestrian's gender, age and street width, and how likely pedestrians look.

Two factors, however, appear to have a strong influence on the likelihood of pedestrians looking towards vehicles. As shown in Figure 4.6a, crosswalk type alters looking frequency. As one might expect, at ND crossings, pedestrians require a confirmation from the driver prior to crossing, therefore they look more often. At designated crosswalks, on the other hand, they might assume the driver's compliance with the law, and as a result, they do not look as often. It is particularly important to note that when a traffic signal is present, pedestrians are less likely to look towards the vehicles because the signal has a higher priority in determining the right of way.

Larger pedestrian group size is generally found to make pedestrians more confident to cross (see Figure 4.6b). This is confirmed by our data as the likelihood of looking reduces with the larger groups of pedestrians. We also observe that in larger groups, typically pedestrians standing at the front of the group and closer to the approaching vehicles make eye-contact with the drivers or look at the vehicles, whereas the rest do not look and wait for the people at the front of the group to make a crossing decision (see Figure 4.7).

Although looking signals the intention of a pedestrian to cross the road, it can be an indicator of their awareness about their surroundings. This may impact how likely they will

attempt to cross the street. We expand more on this in Chapter 5.

4.4.2 Factors that Influence Communication

First, we evaluated the effect of demographics on communication. Although we observed different frequencies of communication for male (23%) and female (18%) pedestrians, we found gender not to be statistically significant. Pedestrian's age, however, was shown to influence the frequency of communication. We observed that adults, seniors, and children communicated in 22%, 11% and 0% of the cases respectively. It should be noted that our data was heavily skewed towards adult samples, therefore these results are not necessarily conclusive.

Similar to looking patterns, communication also varies depending on environmental factors. By analyzing different factors we found the strongest dependency between signal and communication. Our findings suggest that the frequency of communication according to street delineation is 36% in *ND*, 18% in *PC* and only 4% in *S*. In our data, we did not observe any form of explicit communication at signalized intersections. The major difference between signalized and other types of crosswalks is that the drivers are expected to comply with the signal, therefore pedestrians do not need to communicate.

Pedestrian group size is the second most significant factor and has an inverse relationship with the likelihood of communication, the larger the group size the lower the chance of pedestrians communicating. For group sizes of 1, 2, 3 and 4 or greater we saw 29%, 20%, 17% and 10% chances of communication respectively.

Street width has a direct relationship with the chance of communication, although with a lower significance. For streets with 1, 2, 3, and 4 lanes or wider we recorded 11%, 18%, 23% and 26% communication frequency.

It is generally difficult to connect pedestrian communication patterns to driver's action. This is particularly the case when pedestrians are performing implicit communication as it commonly indicates their intention of crossing regardless of what the driver might do.

However, pedestrians communicating explicitly do so to express gratitude towards the driver who complied by either slowing down or stopping for an extended amount of time. Hand gestures had a different meaning (yielding) if the driver gave right of way and the pedestrian slowed down or stopped. In a single case where the driver maintained its speed and did not yield, the hand gesture expressed dissatisfaction.

4.5 Summary

In this chapter, we studied the factor of communication in the context of pedestrians and drivers/vehicles interactions. We argued that communication is necessary not only for pedestrian-driver interactions but also for the interactions involving autonomous vehicles.

We determined that communication in traffic is predominantly nonverbal. Our review of the past studies suggests that there are three forms of nonverbal communication: eye contact, which is used to establish a connection, gestures and postural configuration which are used to transfer a meaning regarding one's intentions.

Moreover, we introduced a newly collected large-scale naturalistic driving dataset of pedestrians at the time of crossing. To enhance the diversity of the dataset, the data was collected under different weather and lighting conditions, different types of roads, and various geographical locations across North America and Europe.

Using the proposed dataset, we conducted an empirical study on pedestrian communication in traffic. Our study confirms some of the findings of the past literature on pedestrian communication presented earlier in Chapter 3. In the majority of the cases pedestrians look towards the traffic prior to crossing. The looking action includes establishing eye-contact in cases when pedestrians intend to resolve traffic ambiguities. In addition, we showed that pedestrians' looking frequency varies significantly depending on whether they end up crossing the street.

Our findings also agree with previous works showing that communication is quite frequent and important in traffic interactions. We found that, similar to the way drivers communicate, pedestrians mainly communicate implicitly by changing their movement patterns, for example, stopping, clearing the way or changing their walking speed. Explicit forms of communication, albeit rare, can be quite important for resolving traffic ambiguities by expressing the intention of pedestrians, e.g. yielding or asking for the right of way.

For the first time, we looked at some of the contextual factors that might impact the way pedestrians communicate with traffic. We argued that factors such as the size of the group that pedestrians are part of or the width of the road they are intending to cross can have a direct impact on how likely pedestrians would look towards the traffic or communicate. Some of the other important factors include pedestrian demographics (i.e. age and gender) and the presence of traffic signals or designated crosswalks.

Chapter 5

Understanding Pedestrian Crossing Behavior: An Empirical Study

In Chapter 3, we talked about traffic context and enumerated environmental and social factors that impact pedestrian behavior and crossing decision. Later in that chapter, we pointed out that the majority of the studies on pedestrian behavior focus either on an isolated small set of factors or study pedestrian behavior in a limited context (e.g. at the particular intersection). In addition, none of these studies highlight the patterns of behavior pedestrians exhibit when making a crossing decision. In this section, we address these shortcomings. Using a large naturalistic dataset, JAAD, we aim to have a more comprehensive look at pedestrian behavior, identify factors that impact pedestrian crossing decision and discuss how some of these factors are interconnected. Another objective of this section is to highlight some of the major challenges in developing practical systems capable of pedestrian behavior understanding and prediction.

5.1 Pedestrian Behavioral Data

As mentioned earlier, in this chapter, we make use of the JAAD dataset (see Section 4.3.1), in particular, the behavioral annotations associated with it. We annotated each pedestrian in the JAAD dataset the following behavioral tags: First set of tags capture pedestrian movement behavior, i.e. whether the pedestrian is *walking* or *standing*, whether the pedestrian is *moving slow* or *moving fast*, or whether the pedestrian is reacting to the driver's actions by *slowing down*, *speeding up* or *stopping*. Another set of tags show whether the pedestrian is *looking* towards the traffic, *nodding* or showing *hand gesture* to the driver. Labels for moving and changes in moving speed are mutually exclusive, but they can overlap with all other labels such as crossing, attention (looking), and gestures.

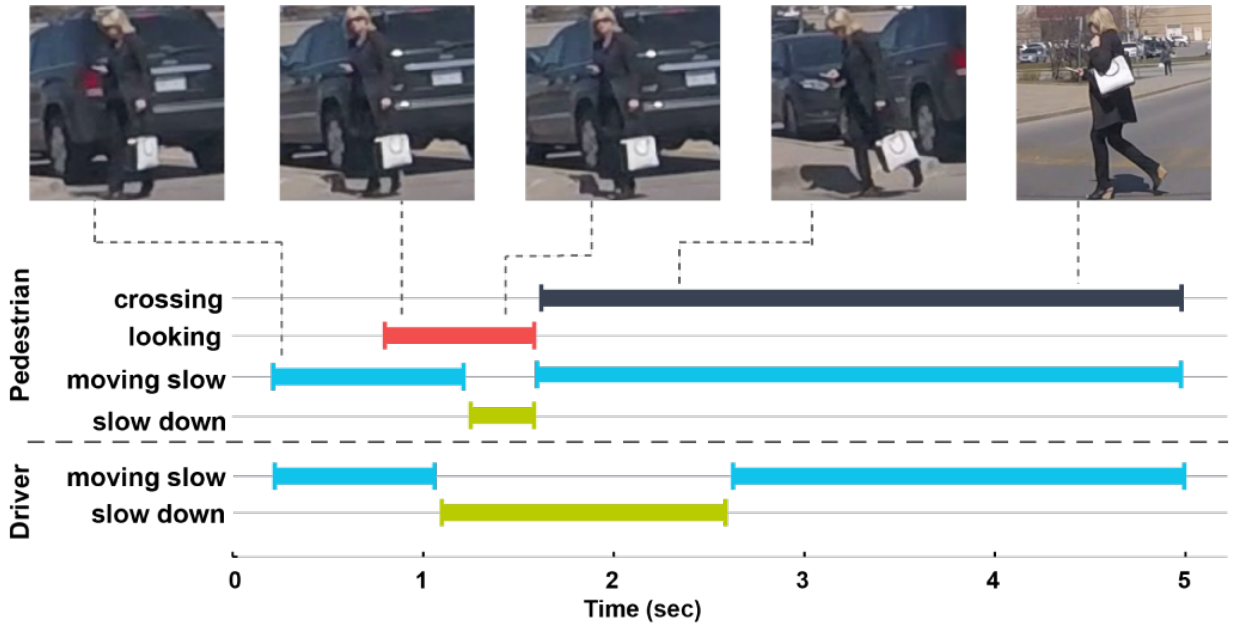
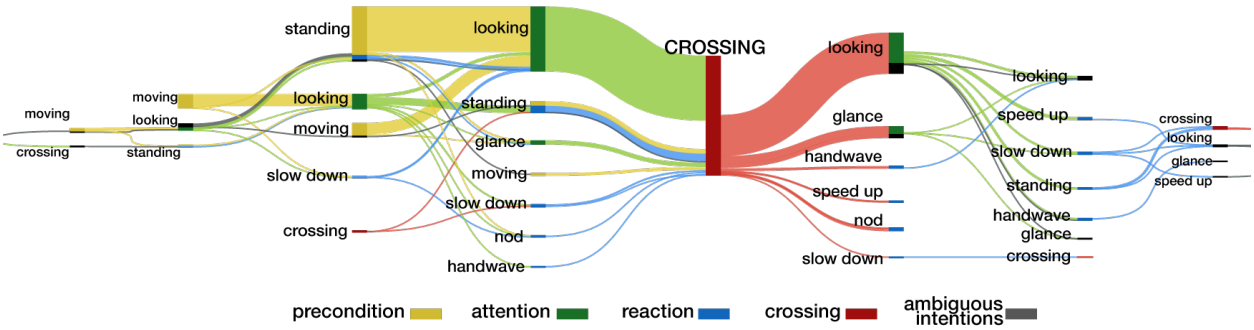


Figure 5.1: The timeline of events is recovered from the behavioral data and shows a single pedestrian crossing the parking lot. Initially, the driver is moving slow and, as he notices the pedestrian ahead, slows down to let her pass. At the same time the pedestrian crosses without looking first then turns to check if the road is safe, and, as she sees the driver yielding, continues to cross.

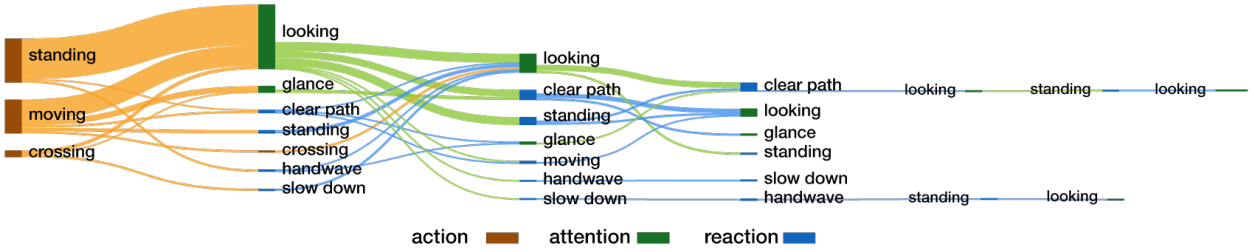
JAAD also contains behavioral tags for the ego-vehicle's driver. The driver's behavior is captured based on observable changes in the motion of vehicle in videos. The driver's actions are summarized with the following tags: *speeds* (when the driver either maintains the current speed or speeds up), *slows down* and *stops*. An example of annotations for a pedestrian and the driver's actions is illustrated in Figure 5.1.

5.2 Pedestrian Behavior at the Time of Crossing

In our data, we observed high variability in the behaviors of pedestrians at the point of crossing/no-crossing with more than 100 distinct patterns of actions. For instance, Figure 5.2a shows sequences of actions during the completed crossing scenarios found in the dataset. Two typical patterns, "standing, looking, crossing" and "crossing, looking", cover only half of the situations observed in the dataset. Similarly, in one third of non-crossing scenarios (Figure 5.2b) pedestrians are waiting at the curb and looking at the traffic. Otherwise, the behaviors vary significantly both in the number of actions before and after crossing and in the meaning of particular actions (e.g. standing may be both a precondition and a reaction to driver's actions).



(a) crossing events



(b) no crossing events

Figure 5.2: Pedestrian motifs at the time of crossing. Diagram a) shows a summary of 345 sequences of pedestrians' actions before and after crossing. Diagram b) shows 92 sequences of actions when pedestrians did not cross. Vertical bars represent the start of actions color-coded as the *precondition* to crossing, *attention*, *reaction* to driver's actions, *crossing* or *ambiguous* actions. Curved lines between the bars show connections between consecutive actions. The thickness of lines reflects the frequency of the action in the 'crossing'or 'non-crossing'subset. The sequences longer than 10 actions (e.g. when the pedestrian hesitates to cross) are extremely rare, and are truncated.

For further analysis, we split these behavioral patterns into 9 groups depending on the initial state of the pedestrian and whether the attention or the act of crossing is happening. We list these actions and the number of samples in Table 5.1. Here attention refers to the first moment the pedestrian is assessing the environment and expressing his/her intention to the approaching vehicles.

Visual attention takes two forms: looking and glancing. Looking refers to the scenarios in which the pedestrian inspects the approaching car (typically for 1 second or longer), assesses the environment and in some cases establishes eye contact with the driver. The other form of attention, glance, usually lasts less than a second and is used to quickly assess the location or speed of the approaching vehicles. Pedestrians glance when they have a certain level of confidence in predicting the driver's behavior, e.g. the vehicle is stopped or moving very slowly or otherwise is sufficiently far away and does not pose any immediate danger.

Table 5.1: The behavioral patterns observed in the data.

Behavior Sequence	**Meaning**	**# Samples**
Crossing	The pedestrian is observed at the point of crossing and no attention is taking place	152
Crossing + Attention	The pedestrian is observed at the point of crossing and some form of attention is occurred	64
Crossing + Attention + Reaction	The pedestrian is observed at the point of crossing and some form of attention is occurred and the pedestrian changes behavior	29
PreCondition + Crossing	The pedestrian is walking/standing and crosses without paying attention	37
Precondition + Attention + Crossing	The pedestrian is walking/standing and crosses after paying attention	160
Precondition + Attention + Reaction + Crossing	The pedestrian is walking/standing, pays attention and changes behavior prior to crossing	64
Action	The pedestrian is walking/standing and his/her intention is ambiguous	56
Action + Attention	The pedestrian is about to cross and pays attention	43
Action + Attention + Reaction	The pedestrian is about to cross, pays attention and responds	49
Total		654

5.3 Analyzing Pedestrian Crossing Behavior

Our data contains various scenarios in which pedestrians are observed during or prior to crossing. Two categories from Table 5.1, *crossing* and *action*, are omitted from the analysis. In the *crossing* scenarios, pedestrians are not observable prior to the crossing event, therefore it is difficult to assess the behavior of the pedestrians at time or prior to crossing. In *action* cases the intentions of the pedestrians are ambiguous, because, for example, pedestrians are not approaching the curb or are standing far away from the crossway.

5.3.1 Attention Occurrence Prior to Crossing

In Section 4.3, we talked about looking action which can signal the intention of the pedestrian to cross. Besides its communication properties, identifying pedestrian looking action (or as we call it attention here) is important as it shows how much the pedestrian is aware of the surroundings. In our data, we observed that about 90% of pedestrians look towards the traffic (whether it is for an extended amount of time or just a glance). To investigate the

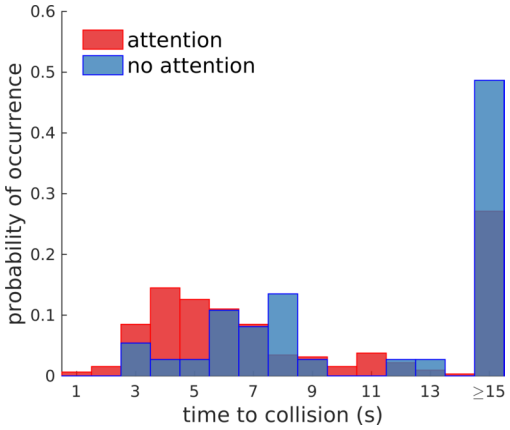
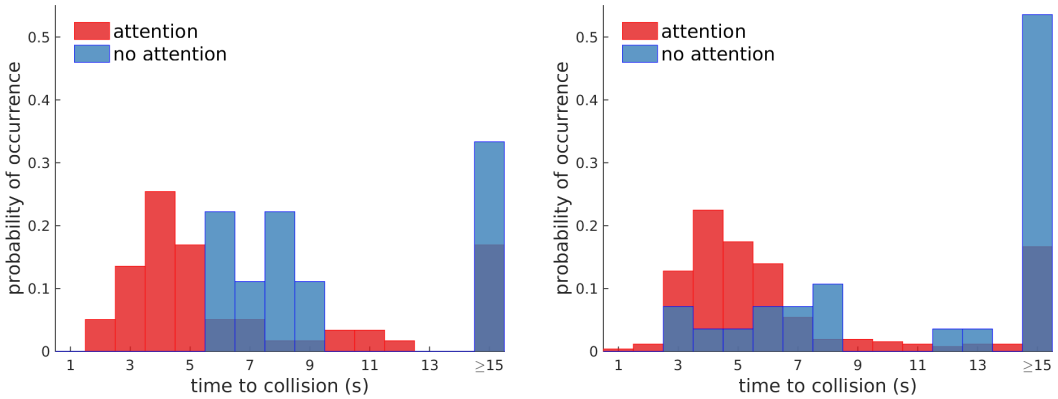


Figure 5.3: Relationship between TTC and probability of attention occurring prior to crossing. Pedestrians more likely look at the traffic prior and during crossing when the crossing time gap is lower.



(a) (b)

Figure 5.4: The pedestrian attention at a) non-designated and b) designated crosswalks. When pedestrians intend to cross at non-designated crosswalks, the tend to be more conservative, and therefore, look at the upcoming traffic more often.

probability of attention occurrence, one important factor to consider is Time To Collision (TTC) (at the start of crossing) or how long it takes the approaching vehicle to arrive at the position of the pedestrian, given that they maintain their current speed and trajectory.

The relationship between attention occurrence and TTC is illustrated in Figure 5.3. Crossing without attention comprises only about 10% of all crossing scenarios out of which more than 50% of the cases occurred when TTC is above 10s (including situations where the approaching vehicle is stopping). There are also no cases of crossing without attention when TTC is less than $2s$.

The context in which the crossing takes place also plays a role in crossing behavior. The context can be described by factors such as weather condition, street structure, and the

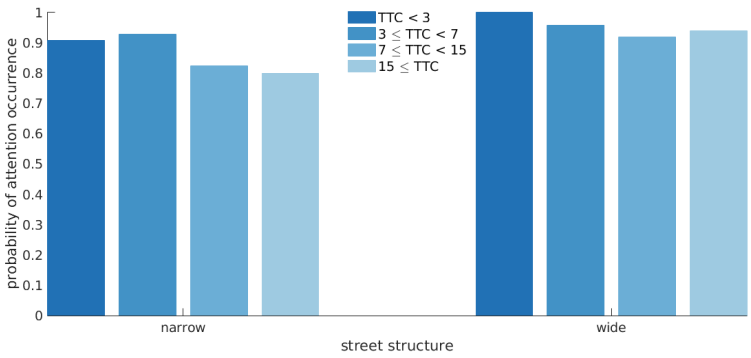


Figure 5.5: Attention occurrence with respect to the number of lanes and TTC. Pedestrians tend to look at the traffic more often when intending to cross wide streets (> 2 lanes) and when TTC is lower.

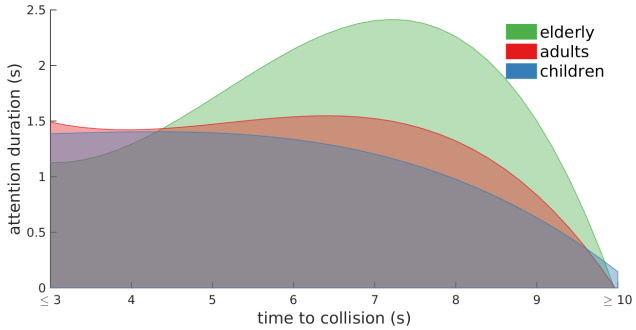


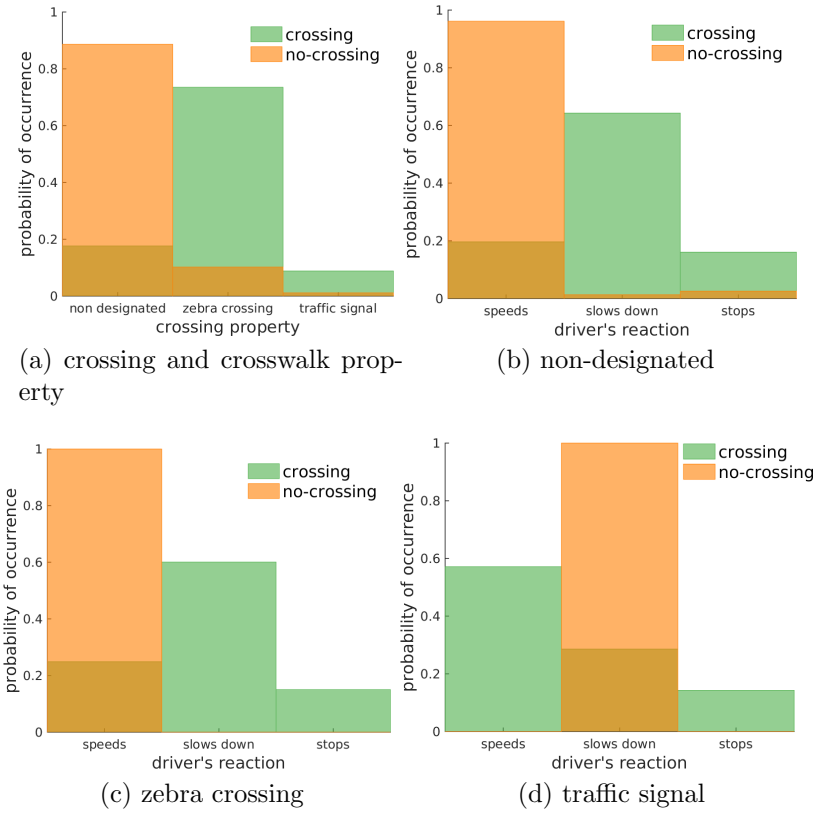
Figure 5.6: Average duration of the pedestrian's attention prior to crossing based on TTC for different age groups.

driver's reaction. Since analyzing all these factors is beyond the scope of this analysis, here we only look at the effect of the street structure.

There are two factors that characterize a crosswalk: whether it is designated (there is a zebra crossing or traffic signal) and its width (measured as the number of lanes). In our samples, crossing without attention only happened in non-designated crosswalks when TTC was higher than 6 seconds (see Figure 5.4).

The full crossing events happen in street with widths ranging from 1 (narrow one-way streets) to 4 lanes (main streets). We report on the data by dividing the results into 4 intervals with respect to the TTC values and in each category, we group them based on the number of lanes (see Figure 5.5). As illustrated, when TTC is below 3s there is no occurrence of crossing without attention in wide streets (more than 2 lanes). For higher TTC values, crossing without attention may still occur in wider streets, however, it happens less frequently than in narrow streets.

The duration of attention or how fast pedestrians tend to begin crossing from the moment they gaze at the approaching car also may vary. As illustrated in Figure 5.6, the duration of



(a) crossing and crosswalk property

(b) non-designated

(c) zebra crossing

(d) traffic signal

Figure 5.7: Pedestrians crossing behavior at crosswalks with different properties.

looking depends on time to collision. The further away the vehicle is from the pedestrians, the longer it will take them to assess the intention of the driver, hence they will attend longer. The gaze duration increases up to a maximum safe TTC threshold (from 7s for adults and up to 8s for elderly) after which it dramatically declines when the vehicle is either far away or stopped. In addition, the elderly pedestrians in comparison to adults and children tend to be more conservative and spend on average about 1s longer on looking prior to crossing.

5.3.2 Crossing Action Post Attention Occurrence

Although the pedestrian's head orientation and attentive behavior are strong indicators of crossing intention, they are not always followed by a crossing event. In addition to TTC, which reflects both the approaching driver's speed and their distance to the contact point, the structure of the street and the driver's reaction can impact the pedestrians' level of confidence to cross.

To investigate this we divide the crosswalks into three categories: *non-designated*, without zebra or traffic signal, *zebra-crossing*, with either zebra or/and a pedestrian crossing sign and *traffic signal* with a signal such as traffic light or stop sign which forces the driver to stop.

Figure 5.7a shows that pedestrians are less likely to cross the street after communicating their intention if the crosswalk is not designated and more likely to cross if some form of signal or dedicated pathway is present.

To understand under what circumstances the crossing takes place in different crosswalks, we look at the driver's reaction to the pedestrian's intention of crossing. Figures 5.7b and 5.7c show that when there is no traffic signal present, in the majority of the cases pedestrians cross if the driver acknowledges their intention of crossing by slowing down or stopping. In a few scenarios, the pedestrian still crosses the street even though the vehicle accelerates. In these cases, either TTC is very high (average of 25.7 s) or the car is in traffic congestion and the pedestrian anticipates that the car would shortly stop. Moreover, crossing also might not take place when the driver slows down or stops (even in the presence of a traffic signal) (see Figure 5.7b and 5.7d). In these cases either the pedestrian hesitates to cross or explicitly (often by some form of hand gesture) yields to the driver.

5.4 What Makes Understanding Pedestrian Actions Difficult

In section 4.3 and earlier in this chapter, we conducted behavioral studies of pedestrians in traffic. We highlighted the ways pedestrians behave and the factors that impact their behaviors. Here, we look at the problem from a computational perspective and analyze the challenges in understanding and predicting pedestrian behavior.

5.4.1 Identifying Actions

(a) Looking (b) Hand gesture

Figure 5.8: Pedestrian actions with the same meaning and different appearances.

One of the major challenges in traffic scenarios is interpreting pedestrian actions such as head movement or gestures. This is primarily due to the fact that actions with the

same meaning can appear in very different ways. For example, pedestrian head orientation while looking can be very subtle or rather extreme, involving major changes in body posture (Figure 5.8a). Likewise, pedestrians' hand gestures as a sign of gratitude can appear very differently (Figure 5.8b).

5.4.2 Identifying Relevant Elements

Figure 5.9: Selecting relevant objects in the scene with the aid of communication cues. Relevant and irrelevant pedestrians are shown with *green* and *red* boxes respectively.

Computational resources in practical systems, such as those used in autonomous vehicles are limited as the vehicles have to deal with various tasks such as visual perception, control, mapping, etc. Given such limitations, reasoning about all road users in the traffic scenes is infeasible. Understanding how pedestrians behave and show their intention of crossing (e.g. stepping onto the road or looking) can help eliminate the ones that are irrelevant to the current driving task, and as a result, reduce computational load. For example, in Figure 5.9, it may seem that all pedestrians close to the curb are relevant and the ones farther away can be ignored. Taking into account pedestrians' attention towards the traffic, we can see that only 4 pedestrians are potentially relevant including the woman with a child on the right who are about to change their direction.

5.4.3 Interpreting Behavior

The Role of Context

Besides the challenges associated with detecting communication cues due to the high variability of pedestrian gestures (see Figure 5.8), understanding the underlying meaning of communication is not always intuitive and requires the knowledge of the context. As depicted in Figure 5.10, by merely relying on gestures, it is hard to infer the intention of the pedestrian.

For instance, in Figure 5.10, the first image from the left, the driver has been standing at the intersection for an extended period of time, therefore the pedestrian is showing his

Figure 5.10: Making sense of hand gesture based on context. From *left* to *right*: showing gratitude, dissatisfaction, yielding and greeting another person (irrelevant).

gratitude. In the second image, the driver does not slow down and the pedestrian is waving his hand as a sign of dissatisfaction. In the third image, both the driver and the pedestrian slow down, but the pedestrian shows a yielding gesture. The last case is quite different as the pedestrian is waving his hand at someone on the other side of the road.

Identifying Relevant Actions

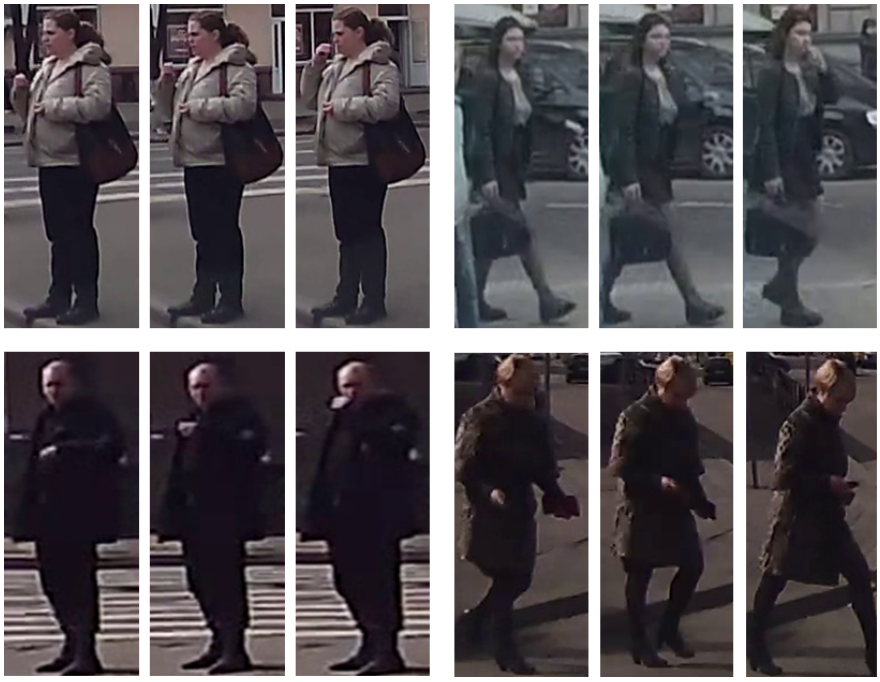


Figure 5.11: Various pedestrian hand movements without any symbolic meaning. From the *left* to *right* and *top* to *bottom*, pedestrians are snacking, touching face, cleaning face and looking at a cellphone.

Not all hand movements are hand gestures. In a course of an interaction, only the types of body movements that contain a symbolic meaning (i.e. an intended message) should be

considered as gestures. In traffic scenes, most observed body movements have no particular meaning, for example, pedestrians eating or bringing hands to their faces, using mobile phones, adjusting their clothes or bags, etc. (see Figure 5.11 for examples). From a practical point of view, these body movements can be easily confused with symbolic hand gestures. To remedy this, once again the knowledge of the context is needed as well as an understanding of the ways pedestrians communicate their intention under different conditions.

5.4.4 Other Road Users

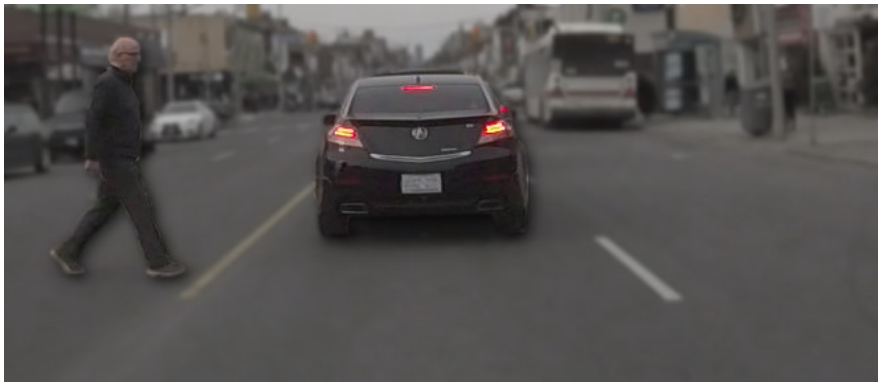


Figure 5.12: The pedestrian is crossing the street regardless of the action of the ego-vehicle' driver because the pedestrian anticipates that the vehicle will stop due to traffic congestion.

In some circumstances, the behavior of pedestrians might be influenced by elements that are not directly involved in the interaction. For instance, as illustrated in Figure 5.12, the pedestrian starts crossing because he expects the ego-vehicle to slow down due to the traffic congestion. This means that when making a crossing decision, the pedestrian relied on the expected behavior of the ego-vehicle due to the environmental conditions instead of its current action.

5.5 Summary

In this chapter, we looked at pedestrian crossing behavior in traffic scenes. Unlike the previous studies, instead of focusing on a few specific factors, we took a broader look at traffic context and its impact on pedestrian behavior.

Through an empirical study, for the first time, we illustrated the wide range of behaviors that pedestrians exhibit at the time of crossing. We concluded that such a diverse range of behaviors makes predicting pedestrians' future actions extremely challenging.

Our study on the role of context on pedestrian behavior suggests that there are numerous elements present in a scene that can help predict what a pedestrian is going to do next. Street properties, such as width, the presence of zebra crossings or traffic signals, can determine pedestrians' level of confidence while crossing. In addition, the driver's dynamic state with respect to pedestrians is important. Factors such as TTC, which reflects the speed and the position of the vehicle, should be considered. These results agree with some of the findings presented in Chapter 3.

As part of this study, we identified the interrelationships between different contextual elements. For instance, although the majority of pedestrians tend to look at the traffic prior to crossing, they do so less when the street is narrow or when TTC is high. This is also true if the crosswalk is signalized because pedestrians feel safer and are, therefore, less cautious while crossing. This suggests that studies of traffic context and its impact on pedestrian behavior should be conducted on a broader scope by including not few but many different contextual elements. This is something that was not often considered in the past works.

At the end of this chapter, we enumerated some of the practical challenges in understanding and predicting pedestrian behavior. Some of these challenges include the high dependency of action interpretability on context, the diversity of action appearances, the presence of irrelevant objects and non-symbolic actions, and the influence of other road users' behaviors.

Chapter 6

Detecting Pedestrians in Cluttered Traffic Scenes

The first step towards understanding pedestrian behavior is to detect relevant objects in the environment. Among typical objects present in traffic scenes, pedestrians are particularly challenging for identification because they assume different poses, have high variability of appearances and can be easily confused with other objects with similar properties [226].

In the past decades numerous pedestrian detection algorithms [227, 228, 226, 229, 230] have been proposed, the majority of which have been tested on the publicly available datasets such as Caltech [231] and KITTI [232]. Although these datasets contain an adequately large amount of data for evaluating the performance of pedestrian detection algorithms, they lack sufficient variability in scene properties such as different lighting conditions and pedestrians' appearance corresponding to different weather conditions. Examples of errors caused by the changes in data properties are illustrated in Figure 6.1.

Given the dynamic nature of driving and the fact that autonomous vehicles should be able to handle a wide range of conditions robustly, there is a need to examine the performance of pedestrian detection algorithms and measure their limitations under various visual conditions.

To this end, we examine the performance of state-of-the-art pedestrian detection algorithms with respect to dataset properties and highlight changes in their behavior with respect to different training and testing samples. We perform a cross-evaluation of the state-of-the-art algorithms on the JAAD (in an extended format) and Caltech datasets to measure the generalizability of algorithms and datasets based on different properties of the data. In addition, as part of this study, we contribute a software framework (which has been made public) for experimentation and benchmarking classical and state-of-the-pedestrian detection algorithms.

(a) Localization errors (b) False positives (c) False negatives

Figure 6.1: Different sources of detection errors due to the variability in the appearance of the pedestrians and scenes: a) shows localization errors due to the presence of bags, backpack and umbrellas which are commonly associated with pedestrians observed in the scenes; b) false positives caused by various environmental factors such as reflections on wet surfaces, over-exposure as well as the presence of objects resembling pedestrians; and c) false negatives due to variation in shape, e.g. children who have different aspect ratio compared to adults, and appearance, e.g. pedestrians wearing hooded jackets, holding umbrellas or carrying bulky backpacks.

6.1 A Literature Review on Pedestrian Detection

6.1.1 Pedestrian Detection Algorithms

Pedestrian detection is a well-studied field. Over the years, a large number of algorithms have been developed, ranging from models based on hand-crafted features [228, 227, 233] to modern convolutional neural networks [230, 229, 234], and hybrid algorithms benefiting from a combination of both of these techniques [235, 236].

The modern pedestrian detection algorithms use various techniques to overcome the challenges of identifying pedestrians in the wild. For example, Tian *et al.* [237] propose a part-based detection algorithm to deal with occlusion. The model consists of a number of part detectors, combinations of which determine the existence of a pedestrian in a given location. In [226], the authors use semantic information of the scene in the form of pedestrian attributes, e.g. carrying a backpack, and scene attributes such as trees or vehicles to distinguish the pedestrians from the background.

In [234] the authors use bootstrapping techniques to mine hard negative samples to minimize confusions caused by background while detecting pedestrians. The proposed algorithm uses features learned by a region proposal network (RPN) to train a cascaded boosted forest

for the final hard negative mining and classification. In a more recent approach, Brazil *et al.* [230] show that jointly training a Faster R-CNN network and semantic segmentation network on pedestrian bounding boxes can improve the overall detection results.

The focus of more recent algorithms is on occlusion handling in detection using methods such as part-based detection [238, 239, 240], novel loss functions [241, 242], attention mechanisms [243], and feature transformation techniques [244]. Some algorithms are concerned with pedestrian detection at far distances [245] and in crowds [246]. A group of algorithms proposes various architectural and computational techniques to improve the overall performance, e.g novel feature generation methods [247, 248, 249], novel anchor generation methods [250] and data synthesis methods to generate more training samples [251]. Even though many of the recent works adopted more diverse datasets such as CityPersons [252] for benchmarking, most of these algorithms still heavily use datasets such as Caltech [231] and KITTI [232]. Given that the performance of state-of-the-art pedestrian detection algorithms on benchmark datasets began to saturate (e.g. 7-9% miss rate reported on Caltech [231]), attention has shifted towards the effects of data properties on detection performance.

6.1.2 Pedestrian Detection Datasets

Besides datasets such as KITTI [232], Waymo [253], nuScenes [254] and ArgoVerse [255] that have multiple object classes, there are a number datasets that are specifically designed for training pedestrian detection algorithms including MIT [256], INRIA [257], Daimler datasets [258, 259, 260, 261], Penn-Fudan [262], ETHZ [263], Caltech [231], TUD Brussels [264], Berkley [265], CUHK [266], PETA [267], Kaist [268], CityPersons [252], and EuroCity Person [269]. Among these datasets, Caltech and CityPersons (more recently) are widely used. The Caltech dataset contains a very large number of pedestrian samples (350K) along with occlusion information in the form of bounding boxes that cover only the visible portions of occluded pedestrians. This dataset, however, is collected under uniform weather conditions and on similar roads which make this dataset very monotonous in terms of visual properties. CityPersons, on the other hand, has been collected in different geographical conditions and has more diverse samples. This dataset, however, does not contain samples under severe weather conditions and does not also have pedestrian attribute information.

6.1.3 Data Properties and Pedestrian Detection

A recent study on generic object recognition tasks shows that an order of magnitude increase in the size of training samples can enhance performance even in the presence of up to 20% error in ground truth annotation [270]. For example, factors such as the effect of occlusion

and sample size [271], the balance between negative and positive samples [272], and the cleanness of ground truth annotations [273] have been investigated. Zhang *et al.* [274], for instance, demonstrate that the percentage of mis-classifications and localization errors varies significantly depending on the algorithm. Through experimental evaluations, the authors show that simply by improving the quality of ground truth annotations, localization errors can be significantly reduced resulting in an overall performance boost of more than 7% miss rate in state-of-the-art pedestrian detection algorithms. What is missing from these studies is that neither of them explores the effects of data properties such as the changes in visual representations, e.g. due to lighting or weather conditions or visual appearances of pedestrians, on the performance of the pedestrian detection algorithms.

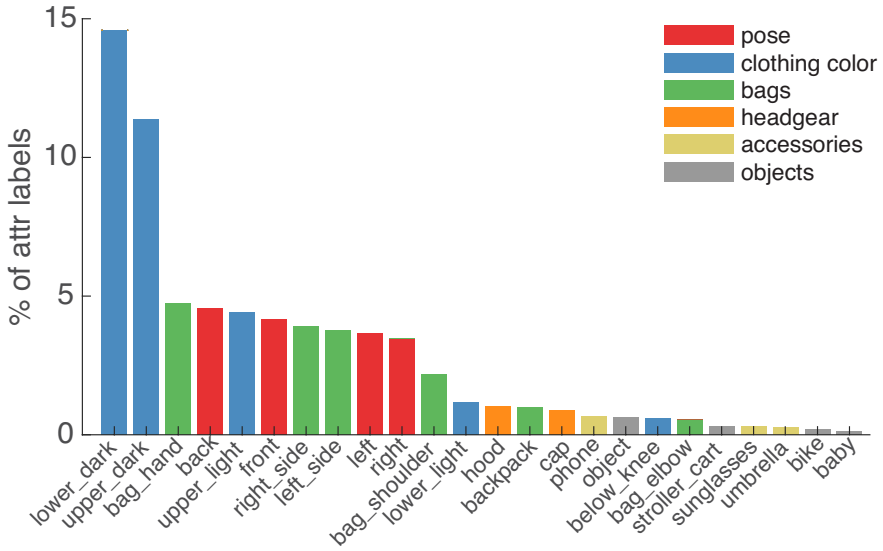
6.2 A Pedestrian Detection and Attribute Dataset

There are a number of existing pedestrian attribute datasets that provide fine-grained attributes (e.g. RAP [275], PETA [267]). These datasets primarily cater to applications such as surveillance and identification tasks, and, as a result, often contain indoor scenes or are recorded using on-site security cameras. Such characteristics make these datasets unsuitable for analyzing pedestrian detection algorithms for applications such as autonomous driving. This is because some of these datasets are recorded from a viewing angle that is not representative of driving scenarios (e.g. bird's-eye-view for surveillance), there is often no camera motion as in driving scenarios, and visual representations, such as lighting conditions, object types, etc., are not the similar to those present in traffic scenes.

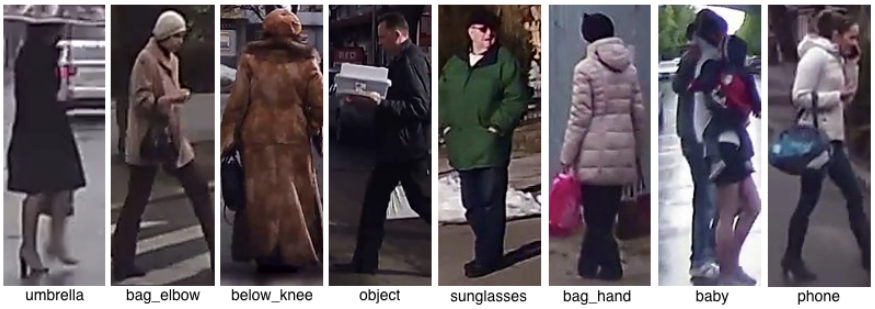
For the context of pedestrian detection in traffic, Tian *et al.* [226] introduced pedestrian attribute information for the Caltech dataset. The authors augmented the dataset with 9 attributes on 2.7K pedestrian samples. These augmented annotations, however, lack attribute diversity we require for our study purposes as the Caltech dataset has insufficient variability of weather, lighting and scenery properties.

To investigate the effect of pedestrian attributes and data properties on detection algorithms, we utilized the JAAD dataset (introduced earlier in Section 4.3.1). We further extended the annotations of JAAD by annotating all pedestrians in the scenes and adding 16 attributes for each of the 392K pedestrian samples, a total of 900K new attribute labels, summarized in Figure 6.2a. There are attributes for coarse pose (*left*, *right*, *back*, *front*), clothing color (*upper_dark* and *lower_dark*) and length (*below_knee* for long coats and skirts) (see Figure 6.2b for examples).

There are also several attributes for the presence and location of bags and their type: whether they are worn on the *left_side* or *right_side* relative to pedestrian's body and carried



(a)



(b)

Figure 6.2: a) Types and frequency of new attribute labels in the JAAD dataset color-coded based on the attribute type (e.g. pose, clothing color, accessories); b) Samples of pedestrians with select attribute labels shown.

on the shoulder (*bag_shoulder*), elbow (*bag_elbow*), back (*backpack*) or held in the hand (*bag_hand*). In addition, we add labels for hooded clothing (*hood*) and caps (*cap*), accessories (e.g. *phone*, *sunglasses*) and various objects that pedestrians can hold in their hands (e.g. *object*, *baby*).

The attributes were selected based on their appropriateness for the driving tasks. For instance, the pose of the pedestrian and color of their clothing affect visibility. Long clothing obscures the shape and movement of the human body. Caps, hoods, and sunglasses occlude pedestrian's face and may limit their view of the traffic scene as well. Carrying large bags, backpacks or other objects may not only change the appearance and shape of the pedestrian but limit their mobility. Holding a phone does not change the pose significantly, but can be used to determine pedestrian's distraction as illustrated in [276].

Clothing color and pose are the only attributes provided for all bounding boxes in the JAAD dataset and form the minimum attribute set. As can be seen from the bar plot in Figure 6.2a, most pedestrians in the dataset are wearing dark clothes, for instance, nearly 70% of pedestrians have both *upper_dark* and *lower_dark* attributes present.

Pose attributes, *left*, *right*, *back*, and *front*, are nearly equally distributed. Aside from clothing color and pose, the *bags* category is the most represented. In fact, nearly 50% of all pedestrians carry a bag or a backpack. In the following sections, we will consider the effect of the diversity and uneven distribution of attributes in the training data on detection.

6.3 Experiment Setup

6.3.1 Pedestrian Detection Algorithms

For the experimental evaluations in this chapter we choose three classical algorithms as baselines including ACF+ and its variation LDCF [227], and LDCF++ [233], and deep learning algorithms including RPN+BF [234], MS-CNN [277], and SDS-RCNN [230]. From RPN+BF algorithm, we only report the results of its RPN component to highlight how the weak segmentation approach proposed in the RPN component of SDS-RCNN would behave under different conditions.

The algorithms were trained on the subsets of the JAAD dataset using the default parameters proposed by the authors for the Caltech dataset. The only exception is that we modified the width parameters of training and testing images to maintain the aspect ratio of the images in JAAD. For cross-evaluation with the Caltech dataset, we used the pre-trained models published by the authors of corresponding algorithms.

6.3.2 Data

The JAAD dataset contains HD quality images with dimensions of 1080×1920 pixels. To maximize the performance of the detection algorithms using default parameters tuned on Caltech, we resized all images to half-scale of 540×960. For evaluation and training, we selected samples with reasonable scale (bounding box height of 50 pixels or more) with partial occlusion (visibility of 75% or more).

For experimental evaluations, we divided JAAD into four different train/test subsets according to the property of the data in terms of weather conditions including *clear*, *cloudy*, *cloudy+clear* (*c+c*) and *mix*. As the names imply, *clear* and *cloudy* subsets only include training images collected under clear and cloudy skies with no rain/snow, and *mix* contains all weather conditions including clear and cloudy, and more extreme weather conditions such

as rain/snow. It should be noted that we excluded the videos from the JAAD dataset that were collected under very poor visibility conditions such as nighttime and heavy rain.

The training images for each subset are generated by uniformly sampling 50% of the videos that are recorded under the given condition. Each training subset contains approximately 6.5K pedestrian samples. The remainder of the videos (which may include all weather conditions) are also uniformly sampled and divided into validation and test set.

6.3.3 Metrics

To report the performance of the algorithms, we use log-average miss rate over the precision range of $[10^{-2}, 10^{0}]$ (MR_2) and $[10^{-4}, 10^{0}]$ (MR_4) false positives per image (FPPI) as in [234, 274]. We also follow [274] and apply two oracle test cases to measure the contributions of background and localization errors. The localization oracle excludes all false positives that overlap with ground truth from evaluation thus reflecting the contribution of background error. The background oracle does not count all false positives that do not overlap with ground truth hence showing the amount of localization error. All of our results are presented using the matching criterion of intersection over union (IoU) ≥ 0.5, unless otherwise stated.

6.4 Data Properties and Detection Accuracy

6.4.1 Weather

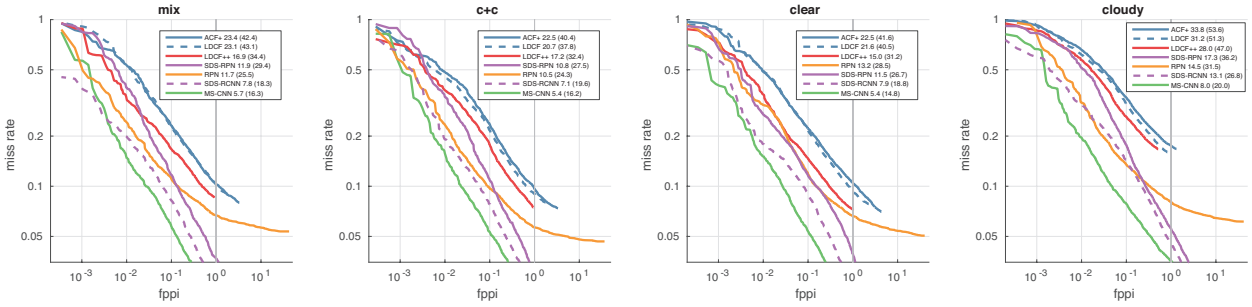


Figure 6.3: ROC curves for all algorithms trained and tested on *mix*, *clear*, *cloudy* and *c+c* (clear and cloudy) datasets with detection threshold set to 0.5 IoU. Legends for each plot show the names of algorithms together with $MR_2(MR_4)$ measures. In each plot legend the algorithms are sorted by MR_2 in the descending order.

Weather conditions have multiple effects on the visibility of the pedestrians (e.g. due to rain) and their appearances (e.g. presence of sunglasses or umbrellas). In addition, the appearance of the scene itself may be altered by different lighting conditions, precipitation,

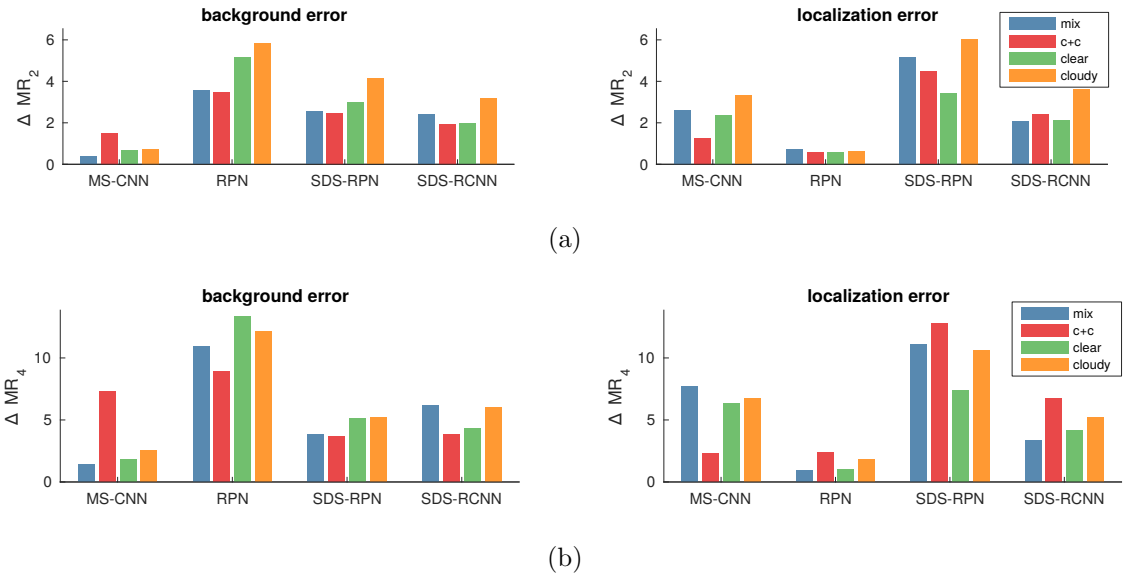


Figure 6.4: The relative contribution of background and localization errors to the performance of state-of-the-art pedestrian detection algorithms. The errors are calculated as changes in a) MR_2 and b) MR_4 measures for algorithms trained and tested on different subsets of JAAD.

reflections, sharp shadows, etc., leading to detection errors as illustrated in Figure 6.1. In order to quantify these effects, we trained and tested all pedestrian detection algorithms on different subsets of JAAD dataset split by weather conditions as explained earlier.

We begin by reporting the ROC curves along with MR_2 and MR_4 metrics. As can be seen in Figure 6.3, despite the changes in the overall performance of the algorithms, the rankings are the same across different subsets. The only exception is in the *clear* case where SDS-RPN outperforms RPN.

The main difference between SDS-RPN and regular RPN is that the former adds a weak segmentation component utilizing binary masks from bounding box annotations. It is apparent that using this technique is only effective under clear weather conditions which correspond to the properties of the Caltech dataset that this algorithm was originally tested on (see Table 6.3). Under different weather conditions, however, the weak segmentation results in a poorer performance compared to the regular RPN.

Another observation is that the MS-CNN algorithm (which according to [249] is not among top five performing algorithms on Caltech) achieves the best performance by a large margin (up to 2% on *mix, clear* and *c+c* subsets and more than 5% on *cloudy*) compared to state-of-the-art SDS-RCNN.

To further understand the underlying factors impacting the performance of each algorithm, we report background and localization errors under different weather conditions. As

Table 6.1: The performance of pedestrian detection algorithms in the presence of individual attributes. The results are reported as MR_4 metric. The top performing algorithms for each attribute are highlighted in bold.

Algorithms	Attributes										
	female	*male*	*pose_back*	*pose_front*	*pose_left*	*pose_right*	*child*	*backpack*	*bag*	*cap_hood*	*umbrella*
ACF+	38.96	34.66	39.71	38.28	34.70	33.91	60.92	38.88	36.00	40.21	69.18
LDCF+	37.02	33.84	35.27	37.24	32.90	30.94	55.02	33.50	33.94	38.27	68.16
LDCF++	30.09	28.30	34.41	31.79	26.44	26.71	55.16	32.76	26.69	33.29	56.64
MS-CNN	**13.49**	**14.03**	17.77	**14.00**	15.20	**11.19**	45.37	16.01	**10.77**	**14.08**	31.06
RPN	21.99	25.79	28.03	26.82	22.72	21.34	53.59	24.59	19.48	28.97	37.35
SDS-RPN	24.31	22.57	26.58	23.67	21.51	22.74	52.54	19.50	20.12	24.61	31.68
SDS-RCNN	14.30	15.77	**17.72**	15.29	**14.46**	13.60	**43.14**	**15.85**	12.25	15.68	**25.57**

depicted in Figure 6.4, testing and training on the subsets of JAAD with different properties reveal inconsistencies in the performance of each detection algorithm as well as their relative performance compared to other algorithms. For example, in the case of *c+c*, MS-CNN reaches its highest background error while at the same time it achieves the lowest localization error compared to others.

For RPN-based models the same trend does not hold as they all perform poorly in terms of localization error, when trained and tested on *c+c*. Comparatively, MS-CNN has the lowest background error on the *mix, clear* and *cloudy* subsets and the second worst on *c+c*.

Likewise, on average, RPN performs best in terms of localization error, however, it is the worst in terms of background error. One interesting observation is the added benefit of the weak segmentation component to RPN (in SDS-RPN) which helps improve the background error but at the price of reducing its localization accuracy.

6.4.2 Pedestrian Attributes

In this section we evaluate the contribution of select attributes (shown in Table 6.1) on the performance of detection algorithms trained and tested on the *mix* dataset. Due to the fact that many attributes often appear together in various combinations, it is very hard to disentangle the effect of the individual attributes on the overall detection accuracy of each algorithm. However, major differences can be observed in the relative performances of the algorithms in the presence of certain attributes in the scene.

As one would expect, the performance of classical models is inferior compared to the CNN-based algorithms, particularly with respect to some of the rarely occurring attributes such as *child* and *umbrella*. The performance of the state-of-the-art also varies on different attributes. For example, MS-CNN, which shows the highest results on *mix*, underperforms compared to SDS-RCNN on *umbrella, backpack, child, pose-back* and *pose-left*.

To investigate the common causes of error for MS-CNN and SDS-RCNN we group false

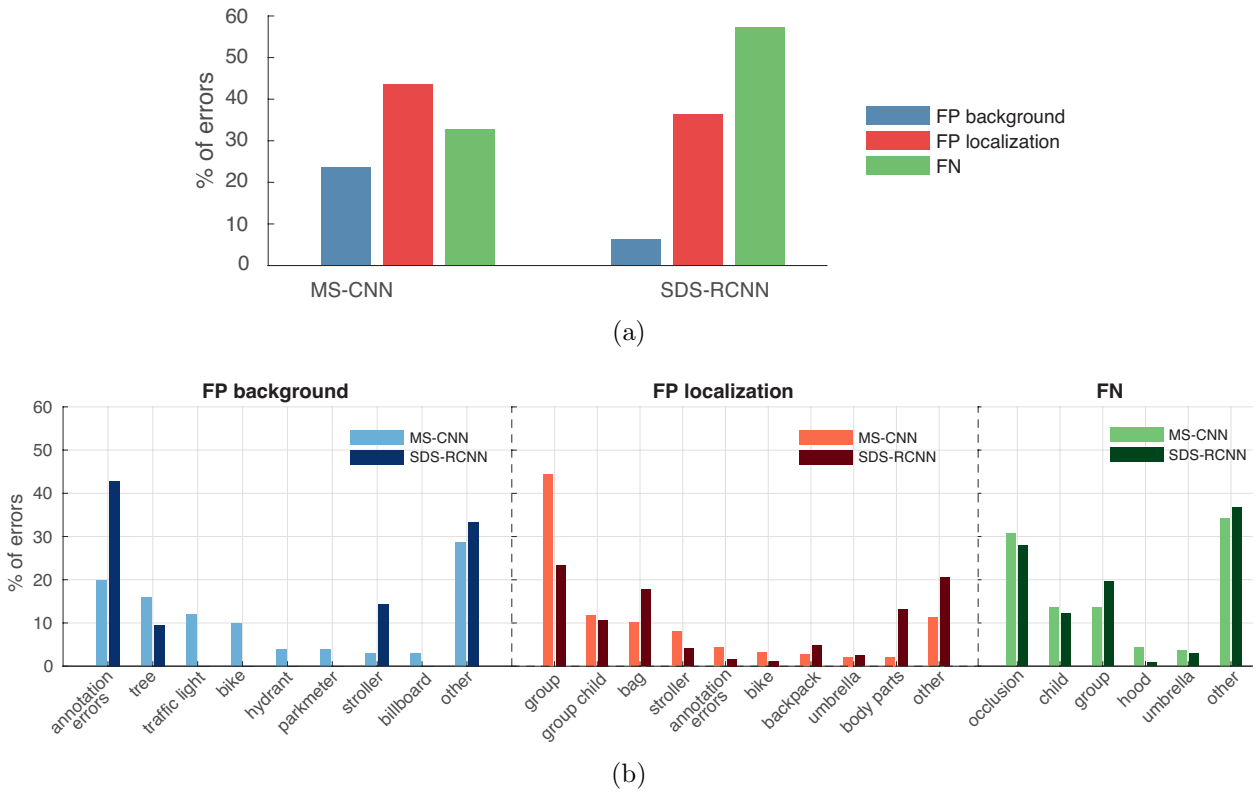


Figure 6.5: Error analysis for MS-CNN and SDS-RCNN trained and tested on the *mix* reasonable subset of JAAD. Plot a) shows the relative percentages of false positives (FP) and false negatives (FN) for each algorithm at 0.1 FPPI. FP is further split into localization and background errors depending on whether the detected bounding box overlaps with the ground truth or not. Plot b) shows a detailed breakdown of false positive and false negative errors grouped by the corresponding attributes.

positive (FP) and false negative (FN) detections at 0.1 FPPI by the object present in the bounding box as shown in Figure 6.5.

Beginning with FP, SDS-RCNN and MS-CNN differ greatly not only in the relative contributions of background and localization errors but also in terms of the objects they commonly confuse with pedestrians. Aside from annotation errors, MS-CNN is much more distracted by elongated objects often found in the street scenes, such as tree trunks, hydrants and parking meters.

Many of the localization errors for both MS-CNN and SDS-RCNN are caused by not being able to distinguish pedestrians in groups of 2 or more, particularly when children are also present (attribute *group child* in Figure 6.5b). SDS-RCNN also has a higher tendency to place bounding boxes on body parts of the pedestrians or objects they carry (e.g. bags) than MS-CNN. Finally, for both MS-CNN and SDS-RCNN, partially occluded pedestrians, groups of pedestrians and children stand out as main sources of false negative detections.

Table 6.2: The performance of state-of-the-art pedestrian detection algorithms on the Caltech and JAAD *mix* datasets. The table shows the results for algorithms trained and tested on the same dataset. In the table, for example, $C \rightarrow C$ means that the models were trained and tested on Caltech. The performances on the Caltech test set are reported on both old (MR^O) and new (MR^N) annotations. The best results are highlighted with blue color.

	$C \rightarrow C$ $MR_2^N(MR_2^O)$	$mix \rightarrow mix$ MR_2
ACF+	26.27 (30.55)	23.36
LDCF+	23.07 (25.79)	23.07
LDCF++	13.66 (16.10)	16.90
RPN	11.71 (14.33)	11.71
MS-CNN	9.47 (11.21)	5.70
SDS-RPN	8.15 (9.27)	11.89
SDS-RCNN	6.58 (7.59)	7.78

Note that despite individual sensitivities to certain attributes, both MS-CNN and SDS-RCNN have trouble detecting children and pedestrians with infrequently occurring attributes such as backpacks, umbrellas, hooded clothing, etc. There is also evidence that algorithms may learn the appearance of common attributes such as bags instead of the pedestrian itself leading to poor localization.

The former issue may be addressed by increasing the variability of the training data either by explicitly ensuring the presence of certain hard attributes or implicitly, by gathering data under different weather conditions, which in turn affect the appearance of the pedestrians. On the other hand, explicitly learning the attributes may also help, as demonstrated by [226].

6.4.3 Generalizability Across Different Datasets

Here, our goal is to identify the link between the generalizability of the dataset and its properties, i.e. we want to measure whether using training data from a diverse dataset can improve the performance of detection algorithms on other datasets with more uniform properties.

For this purpose, we employed the widely used Caltech dataset [231] and JAAD. We evaluated the algorithms trained on Caltech using the test data from the *mix* subset of JAAD, and also the models trained on different subsets of JAAD using Caltech test set. All the tests are done on a reasonable set of pedestrians with a height of 50 pixels and above. The minimum allowable visibility is set to 75% on the Caltech test set to match the partial occlusion of the JAAD dataset.

Given that a large portion of the original bounding box annotations in the Caltech dataset

Table 6.3: The performance of state-of-the-art pedestrian detection algorithms on the Caltech and different subsets of the JAAD dataset. The results show the performance of the algorithms trained on Caltech and tested on JAAD ($C \rightarrow mix$) and trained on different subsets of JAAD and tested on Caltech ($J \rightarrow C$). The performances on the Caltech test set are reported on both old (MR^O) and new (MR^N) annotations. The best and second best results are highlighted with blue and green color respectively.

	$C \rightarrow mix$ MR_2	$J \rightarrow C$ $MR_2^N(MR_2^O)$			
		mix	c+c	cloudy	clear
ACF+	77.94	46.97 (53.63)	49.52 (55.06)	70.79 (74.06)	49.99 (55.23)
LDCF+	54.82	43.61 (49.93)	44.89 (50.85)	59.18 (64.11)	47.29 (52.54)
LDCF++	47.94	37.66 (46.04)	40.41 (48.54)	54.86 (60.72)	44.77 (51.93)
RPN	40.15	27.80 (41.19)	25.74 (38.18)	34.67 (47.34)	28.75 (40.05)
MS-CNN	35.09	22.87 (34.83)	26.30 (38.11)	31.55 (46.35)	29.49 (41.64)
SDS-RPN	43.40	24.24 (30.84)	26.64 (33.61)	35.62 (42.90)	30.85 (38.52)
SDS-RCNN	25.45	21.47 (27.73)	25.29 (32.69)	35.20 (42.35)	23.81 (31.75)

are poorly localized, following the advice of [274], we report the results on both the original and newly clean Caltech test set. We denote the miss rate results as MR^O and MR^N for old and new annotations respectively. All detections are calculated on $IoU \geq 0.5$.

The results of the evaluations of the algorithms trained and tested on the same dataset are summarized in Table 6.2 and the results of cross-evaluation between algorithms trained and tested on Caltech and subsets of JAAD are shown in Table 6.3.

The first observation is that the performance of algorithms on a uniform dataset compared to a diverse one varies significantly. SDS-RCNN algorithm that achieves state-of-the-art performance on Caltech is the second best in JAAD and its counterpart, SDS-RPN, which has the second-best performance on Caltech, performs even worse compared to the regular RPN algorithm. MS-CNN, on the other hand, performs best on the *mix* subset, even though on Caltech it is the third best in our evaluation and not even in the top five in the latest benchmarks [230].

As was mentioned earlier, the Caltech dataset contains images collected during daylight under clear sky. Surprisingly, we observe that the *clear* subset of JAAD that has similar properties does not generalize best to Caltech. Besides having the second-best performance on SDS-RCNN models, it ranks third in other cases. In fact, we can see that diversifying the data by training on *c+c* and further adding extreme weather conditions such as rainy and snowy samples achieves the best results on the Caltech dataset.

Partly, such performance improvement is owing to better localization. For instance, MS-CNN and SDS-RCNN on average have IoUs of 0.73 and 0.75 respectively when trained on

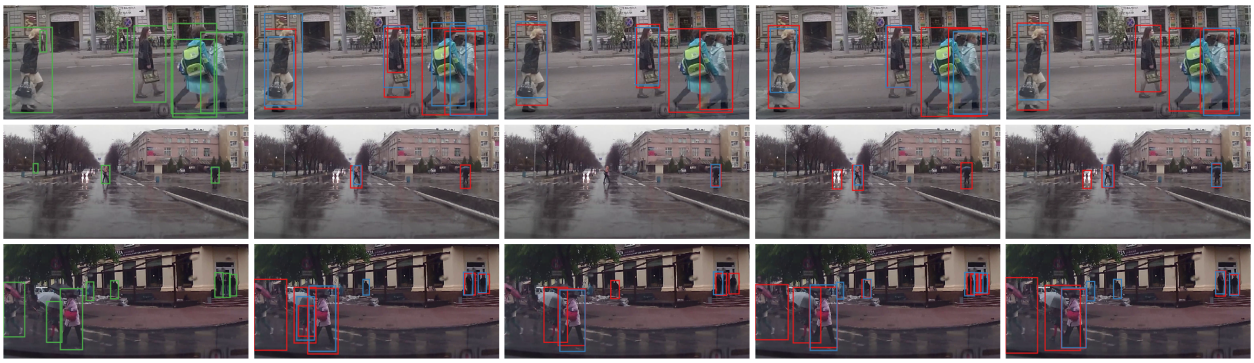


Figure 6.6: Examples of the performance of state-of-the-art pedestrian detection algorithms on samples with different weather conditions and pedestrian attributes. From left to right, the ground truth (GT) and the results of algorithms trained on different subsets of the JAAD dataset are shown. Colors green, red and blue correspond to the ground truth, MS-CNN and SDS-RCNN respectively. The results show that the behaviors of both detection algorithms are affected based on the changes in the training data, but in different and somewhat unpredictable ways. For instance, in the example in the second row, SDS-RCNN performs better when trained on the *mix* subset whereas MS-CNN does so on when trained on the *clear* subset.

JAAD *clear* and 0.74 and 0.76 when trained on JAAD *mix*. The same models trained on Caltech, on the other hand, have an average IoU of 0.73.

It should be noted that the CNN-based models in the table are trained on Caltech10x [228] which contains over 45K images with more than 16K training samples. The diverse *mix* dataset contains less than 7K samples, yet generalizes better on Caltech than vice versa.

6.5 Summary

In this chapter, we conducted a series of experiments to investigate the effect of dataset diversity on the performance of pedestrian detection algorithms (see some qualitative examples in Figure 6.6). Using the extended JAAD dataset, we showed that the performance measures reported on the classical benchmark datasets, such as Caltech, do not necessarily reflect the true potential of detection algorithms in dealing with a wider range of environmental conditions. For instance, MS-CNN which ranks fifth in the recent state-of-the-art benchmarks, outperforms the current top-ranking algorithm, SDS-RCNN, by a significant margin on all subsets of the JAAD dataset.

We showed that the changes in relative performance can be attributed to different properties of the datasets, e.g. depending on what types of weather conditions are represented

in the training data. For example, SDS-RPN outperforms the classical RPN on the Caltech dataset owing to the use of a weak segmentation technique, however, it shows inferior results on the JAAD dataset under all weather conditions except clear (which is the most similar to Caltech).

Similar fluctuations in the performance of detection algorithms can be seen with respect to pedestrian attributes. Particularly, rarely occurring attributes such as *child, backpack* and *umbrella* are associated with the highest miss rate for all algorithms. On the other hand, some of the most frequently occurring attributes such as handbags are shown to be commonly localized instead of pedestrians.

The diversity of training data also leads to better generalization of pedestrian detection algorithms across different datasets. Our empirical results suggest that mixing samples with different properties can improve the performance of algorithms even on a more uniform dataset. For example, the MS-CNN algorithm trained on the *mix* subset of JAAD had 7% and 3% lower miss rates on Caltech compared to the models trained on the *clear* and *c+c* subsets respectively.

A carefully selected dataset can also reduce the need for a large volume of training data. For example, the models trained on the *mix* subset of JAAD using only 7K training samples performed better on the Caltech dataset compared to models that were trained on more than 16K training samples from Caltech and tested on the JAAD *mix*.

In conclusion, our study shows that the selection of benchmark datasets for the evaluation of pedestrian detection algorithms for practical applications such as autonomous driving should be revisited to properly assess their performance and limitations under different conditions and to better reflect their generalizability.

Using larger datasets certainly benefits the training of the algorithms as does balancing the data with respect to underrepresented weather conditions and pedestrian categories. On the other hand, overrepresented attributes in the data can cause detection errors which should be taken into account when designing pedestrian detection algorithms.

Chapter 7

Understanding Pedestrians' Intentions and Their Role in Predicting Trajectories

In Section 2.1.3 we talked about the necessity of understanding intentions in the context of social interaction and coordination. We showed that in addition to sharing the focus of attention, parties involved in an interaction should have the intention of doing so for the sake of accomplishing a common task. For predicting pedestrian behavior in traffic scenes, intention of crossing can serve as a source of contextual information. In this chapter we look at pedestrian intention from a practical point of view and investigate the importance of intention estimation in predicting pedestrian trajectory.

Most current approaches to pedestrian action prediction are trajectory-based [278, 279, 280], meaning that they rely on the past observed motion of the pedestrians and/or vehicle dynamics to predict the future locations of the pedestrians. These approaches, however, are effective when the pedestrians are already crossing or are about to do so, i.e. these algorithms react to an action already in progress instead of anticipating it. For example, scenarios, where a pedestrian is standing at the intersection or walking alongside the road prior to crossing can be challenging for trajectory-based approaches. Moreover, the past trajectory of a pedestrian might not necessarily reflect their ultimate objective. For instance, a pedestrian waiting at a bus stop might step on the road to check for the bus. This action might be interpreted as a crossing event by a trajectory-based approach.

A remedy for the common drawbacks of trajectory-based algorithms is to anticipate the action by estimating its underlying cause or intention. Intention estimation allows one to predict a future situation using expected behaviors rather than merely rely on scene dynamics [281]. In the context of intelligent driving, a pedestrian's intention reflects their principal

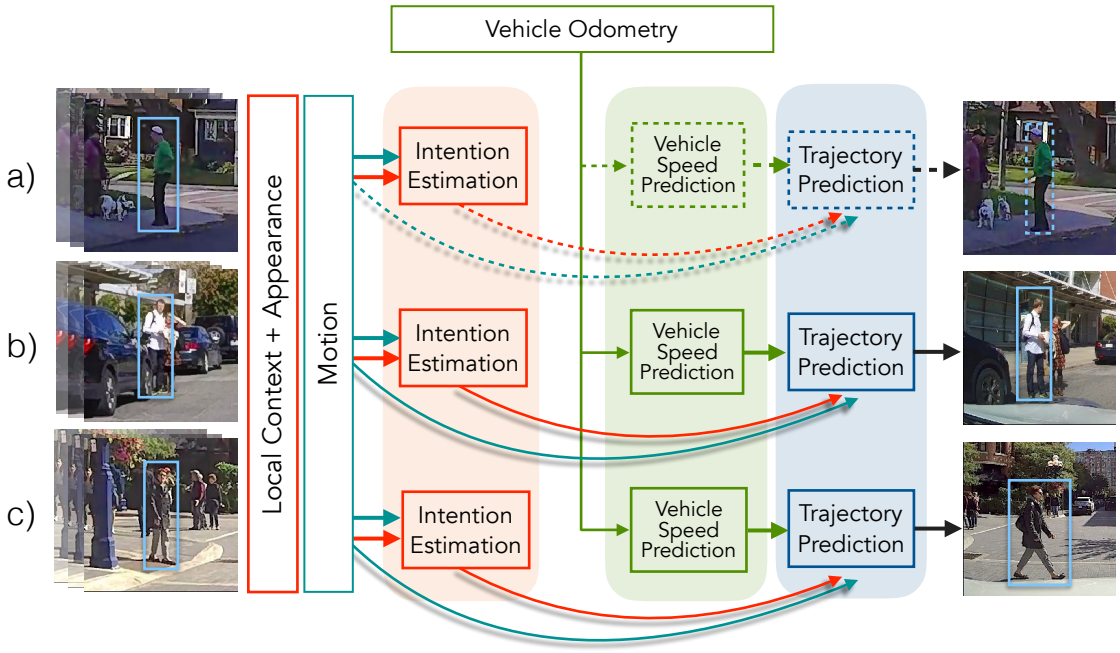


Figure 7.1: Processing stages for different sources of information required for understanding and predicting pedestrian behavior. Three examples are shown: a) a pedestrian who does not intend to cross, b) a pedestrian who intends to cross but does not cross and c) a pedestrian who intends to cross and crosses the street. Observations of pedestrians' appearance and movement in combination with local context help estimate whether they intend to cross the street. Intention can be used to filter out irrelevant pedestrians (eliminating the need for further processing as shown with dashed lines) and/or to improve trajectory prediction.

goal of crossing the street (see Figure 7.1). The pedestrian might not have any intention to cross (e.g. they could be waiting for a bus, talking to someone, or taking a photo), or they intend to cross and may or may not act on it depending on the traffic conditions. Detecting pedestrians' intentions can potentially reduce the cognitive load of an intelligent driving system allowing it to identify those pedestrians whose actions will be relevant to their own behavior planning. This may also grant such systems a better ability to anticipate pedestrian behavior [281].

As part of our contribution presented in this chapter, we introduce a newly collected dataset that is the first large-scale dataset for pedestrian intention estimation and trajectory prediction. We propose a baseline model for pedestrian intention estimation and show how intention can be used to improve the performance of pedestrian trajectory prediction.

7.1 A Literature Review of State-of-the-Art

7.1.1 Vision-Based Trajectory Prediction

As the name implies, these algorithms are designed to predict the future trajectories of objects (e.g. pedestrians), i.e. the future positions of the objects over time. These algorithms are particularly important for applications such as intelligent driving in which predicted locations can be used for route planning or predicting future events such as anomalies, events (e.g. accidents) or actions (e.g. pedestrian crossing).

In recent years, many trajectory prediction algorithms rely on classical reasoning methods including Gaussian mixture models [282, 283] and processes [284, 285], Markov decision processes (MDPs) [286, 287, 288, 289, 290, 291, 292, 293, 294], Markov chains [295, 296] and other techniques [297, 298, 299, 300, 301, 302, 303, 304]. However, in this review we focus on deep learning approaches given their popularity in recent years.

Trajectory prediction applications like many other sequence prediction tasks heavily rely on recurrent architectures such as LSTMs [255, 305, 306, 307, 308, 309, 310, 311, 312, 313, 314, 315, 316, 317, 318, 319, 320, 321, 322, 323, 324, 325, 326, 327, 279], and GRUs [328, 329, 330, 331, 332, 333]. These methods often use an encoder-decoder architecture in which a network, e.g. an LSTM encodes single- or multi-modal observations of the scenes for some time, and another network generates future trajectories given the encoding of the observations. Depending on the complexity of input data, these algorithms may rely on some form of pre-processing for generating features or embedding mechanisms to minimize the dimensionality of the data.

The feedforward algorithms [328, 334, 335, 336, 337, 338, 339, 340, 341, 342, 343, 344] often use whole views of the scenes (i.e. the environment and moving objects) and encode them using convolutional layers followed by a regression layer to predict trajectories. A few algorithms use hybrid approaches in which both convolutional and recurrent reasonings are applied [345, 346].

Depending on the prediction task, algorithms may rely on single- or multi-modal observations. For example, in the context of visual surveillance where a fixed camera provides a top-down or bird's eye viewing angle, many algorithms only use past trajectories of the agents in either actual 2D frame coordinates or velocities of agents calculated by the changes from each time step to another [346, 307, 311, 312, 316, 318, 347, 320, 322, 331, 324, 348, 319, 327, 279]. In addition to observations of individual trajectories of agents, these algorithms focus on modeling the interaction between the agents and how they affect each other. For example, Zhang et al. [307] use a state refinement module that aligns all pedestrians in the scene with a message passing mechanism that receives as input the current locations of the subjects and

their encodings from an LSTM unit. In [311] a graph-based approach is used where pedestrians are considered as nodes and the interactions between them as edges of the graph. By aggregating information from neighboring nodes, the network learns to assign a different level of importance to each node for a given subject. The authors of [320, 279] perform a pooling operation on the generated representations by sharing the state of individual LSTMs that have spatial proximity.

As shown in some works, other sources of information are used in surveilling objects [305, 306, 308, 310, 313, 330, 321, 323, 324, 326, 332, 343, 344]. For example, in addition to encoding the interactions with the environment, Liang et al. [305] use the semantic information of the scene as well as changes in the poses of the pedestrians. In [306, 308, 310, 313, 330, 326, 332, 344] the visual representations of the layout of the environment and the appearances of the subjects are included. The authors of [324] use an occupancy map which highlights the potential traversable locations for the subjects. The method in [321] takes into account pedestrians' head orientations to estimate their fields of view in order to predict which subjects would potentially interact with one another. To predict interactions between humans, in [323] the authors use both poses and trajectories of the agents. Ma et al. [343] go one step further and take into account the pedestrians' characteristics (e.g. age, gender) within a game-theoretic perspective to determine how the trajectory of one pedestrian impacts another.

In the context of traffic understanding, predicting trajectories can be more challenging due to the fact that there is camera ego-motion involved (e.g. the prediction is from the perspective of a moving vehicle), there are interactions between different types of objects (e.g. vehicles and pedestrians), and there are certain constraints involved such as traffic rules, signals, etc. To achieve robustness, many methods in this domain take advantage of multi-modal data for trajectory prediction [345, 255, 328, 309, 329, 336, 314, 337, 349, 350, 330, 317, 335, 339, 280, 333, 340, 341, 332]. In addition to using past trajectories, all these algorithms account for the road structure (whether from the perspective of the ego-vehicle or a top-down view) often in the form of raw visual inputs or, in some cases, as an occupancy map [309, 340]. The scene layout can implicitly capture the structure of the road, the appearances of the objects (e.g. shape) and the dynamics (e.g. velocity or locations of subjects). Such implicit information can be further augmented by explicit data such as the shapes of the objects (in the case of vehicles) [345], the speed [337, 280] and steering angle [337, 280] of the ego-vehicle, the distance between the objects [255, 350], traffic rules [350] and signals [341], and kinematic constraints [342]. For example, the method in [280] uses a two-stream LSTM encoder-decoder scheme: the first stream encodes the current ego-vehicle's odometry (steering angle and speed) and the last observation of the scene

and predicts future odometry of the vehicle. The second stream is a trajectory stream that jointly encodes location information of pedestrians and the ego-vehicle's odometry and then combines the encoding with the prediction of the odometry stream to predict the future trajectories of the pedestrians. Chandra et al. [345] create embeddings of contextual information by taking into account the shape and velocity of the road users and their spatial coordinates within a neighboring region. These embeddings are then fed into some LSTM networks followed by a number of convolutional layers to capture the dynamics of the scenes. In [309] the authors use separate LSTMs for encoding the trajectories of pedestrians and vehicles (as oriented bounding boxes) and then combine them into a unified framework by generating an occupancy map of the scene centered at each agent, followed by a pooling operation to capture the interactions between different subjects. Lee et al. [332] predict the future trajectories of vehicles in two steps: First, an encoder-decoder GRU architecture predicts future trajectories by observing the past ones. Then a refinement network adjusts the predicted trajectories by taking into account the contextual information in the form of social interactions, dynamics of the agents involved, and the road structure.

A recent trend in trajectory prediction algorithms is the use of attention modules [305, 306, 307, 311, 313, 330, 319, 348, 331]. For example, in [305, 311], the attention module jointly measures spatial and temporal interactions. The authors of [306, 307, 313, 348] propose the use of social attention modules which estimate the relative importance of interactions between the subject of interest and its neighboring subjects. Xue et al. [319] propose an attention mechanism to measure the relative importance between different data modalities, namely the locations and velocities of subjects.

Overall, as it becomes obvious from the review, the recurrent architectures are strongly favored in this domain primarily due to the fact that they are flexible in terms of dealing with variable lengths of data sequence and also incorporating multi-modal data.

7.1.2 Intention Estimation

In the computer vision and robotics literature, the term intention is often used in the context of action classification or path refinement. In [351], the authors assume that pedestrians want to cross and decide whether the crossing takes place in front of the vehicle and when. Intention, defined as the potential goal (destination) of pedestrians, is used to refine predicted trajectories [352, 353, 294, 325]. These approaches rely heavily on the motion history of the pedestrians and predict the trajectory of every individual.

To the best of our knowledge, there is only one previous work that defines pedestrian crossing intention as their principal goal to cross [281]. The authors propose to infer pedes-

trian crossing intention from their movement patterns and their proximity to various road elements, e.g. curbside, bus stop, ego-vehicle lane. Their algorithm, however, does not contain a perception mechanism and relies on ground truth information for reasoning.

7.1.3 Datasets for Pedestrian Trajectory Prediction

A number of datasets for trajectory prediction contain videos collected from a top-down view [354, 355, 356, 357] or surveillance camera perspective [358, 359, 360]. There are relatively fewer datasets that are specifically catered for pedestrian behavior prediction from a moving vehicle perspective. Publicly available pedestrian detection datasets [231, 232, 252] can potentially be used for such a purpose, however, they lack necessary characteristics such as ego-vehicle information [231], temporal correspondence [252], or enough pedestrian samples with long tracks [232]. These datasets also do not include any form of pedestrian behavior annotations that can be used for action prediction.

The JAAD that we introduced in Section 4.3.1, contains a large number of pedestrian samples with temporal correspondence, a subset of which are annotated with behavior information. However, for the purposes of intention estimation and trajectory prediction, this dataset has a number of drawbacks. The dataset does not have ego-vehicle information, the videos are divided into short discontinuous chunks, and the majority of pedestrian samples with behavioral annotations have the intention of crossing.

7.2 Pedestrian Intention Estimation (PIE) Dataset

The PIE dataset[1] [2] consists of over 6 hours of driving footage captured with calibrated monocular dashboard camera Waylens Horizon equipped with 157° wide-angle lens. All videos are recorded in HD format (1920×1080 px) at 30 fps. The camera was placed inside the vehicle below the rear-view mirror. For convenience, videos are split into approx. 10 minute long chunks and grouped into 6 sets. The entire dataset was recorded in downtown Toronto, Canada during the daytime under sunny/overcast weather conditions.

Our dataset represents a wide diversity of pedestrian behaviors at the point of crossing and includes locations with high foot-traffic and narrow streets as well as wide boulevards with fewer pedestrians. PIE provides long continuous sequences and annotations for a wide range of applications.

[1]The dataset and more details regarding the annotations can be found at http://data.nvision2.eecs.yorku.ca/PIE_dataset/.

[2]Collection of this dataset was approved by the York Ethics Committee with certificate # 2016-203.

	PIE	JAAD
# of frames	911K	82K
# of annotated frames	293K	75K
# of pedestrians	1.8K	2.8K
# of pedestrians with behavior annot.	1.8K	686
# of pedestrian bboxes	740K	391K
Avg. pedestrian track length	401	140
Pedestrian intention	yes	no
Ego-vehicle sensor information	yes	no
Scene object annotations	bboxes+text	text

Table 7.1: Properties of the PIE dataset compared to JAAD.

7.2.1 Annotations

For each pedestrian close to the road that can potentially interact with the driver we provide the following annotations: bounding boxes with occlusion flags, as well as crossing intention confidence and text labels for pedestrians' actions ("walking", "standing", "looking", "not looking", "crossing", "not crossing"). Each pedestrian has a unique id and can be tracked from the moment of appearance in the scene until going out of the frame. An occlusion flag is set to partial occlusion if between 25 and 75% of the pedestrian is not visible and to full if $> 75\%$ of the pedestrian is not visible. Crossing intention confidence is a numeric score estimated from human reference data (see Section 7.3).

Spatial annotations are provided for other relevant objects in the scene, including infrastructure (e.g. signs, traffic lights, zebra crossings, road boundaries) and vehicles that interact with pedestrians of interest[3].

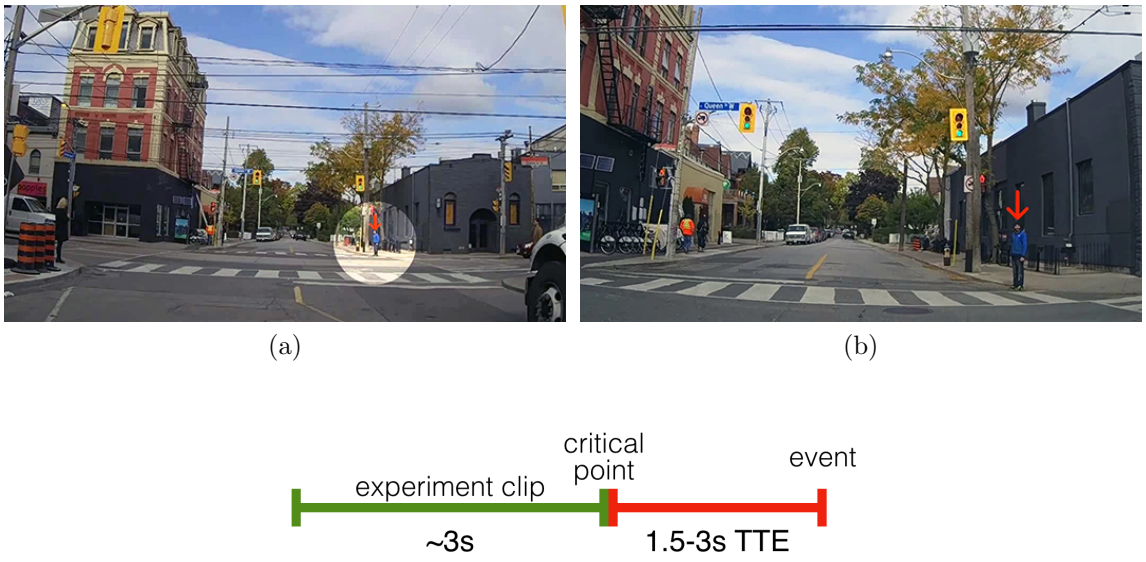
Using an on-board diagnostics (OBD) sensor synchronized with the camera we provide GPS coordinates and vehicle information, such as accurate speed and heading angle, for each frame of the video. Table 7.1 summarizes the properties of PIE and JAAD datasets. JAAD has bounding box annotations for all pedestrians, which makes it suitable for detection and tracking applications. However, it lacks accurate vehicle information, spatial annotations for traffic objects and pedestrian intentions which are vital for pedestrian action prediction.

[3]We used the CVAT tool (https://github.com/opencv/cvat) for all spatial annotations and behavior labels.

7.3 A Human Study on Predicting Pedestrian Crossing Intention

As mentioned earlier, research in the field of pedestrian behavior understanding largely focuses on the problem of action and behavior prediction, while the topic of intention estimation remains relatively unaddressed. Partly this is due to the fact that establishing ground truth for crossing intention is infeasible since it would require interviewing people on the street and observing their actions after the vehicle passed by them [281]. However, this data is necessary for identifying and focusing on the most relevant pedestrians on the street, pedestrian behavior understanding and prediction, including trajectory estimation. In order to determine human reference data for samples in the PIE dataset we conducted a human experiment described below.

7.3.1 Experiment Description



(a) (b) (c)

Figure 7.2: An example of the first (a) and the last (b) freeze-frames from a video clip used in the human experiment. c) The timeline showing how the clips were cropped from the pedestrian track.

The experiment involved watching short videos from the PIE dataset. We asked the participants to observe a single pedestrian highlighted in the first few seconds and, after viewing each video *once*, answer the following question: "Does this pedestrian **want** to cross

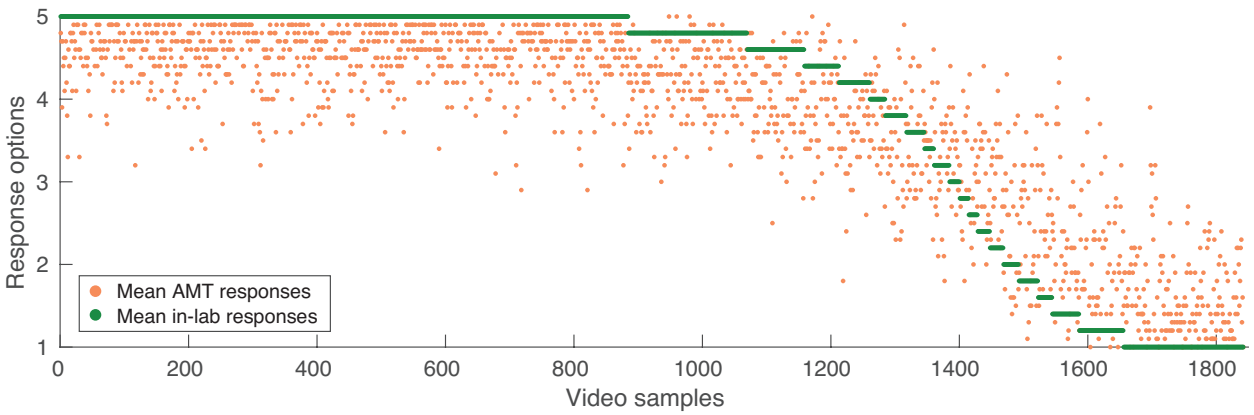


Figure 7.3: A plot of average responses to the question "Does this pedestrian want to cross?" for each of the 1842 video samples containing a single pedestrian of interest. Answer option 5 is selected for the presence and option 1 for the absence of crossing intention respectively. Answer options in between represent various levels of uncertainty. In-lab and AMT responses are shown as green and red dots respectively. Average responses are sorted in descending order for clarity.

the street?". The options were set on a 5-interval scale (the outer intervals for definite 'yes' or 'no' and 3 intervals expressing varying degrees of uncertainty in between).

Videos used in the experiment were generated for each of the 1842 labeled pedestrians in the PIE dataset. Using GPS information and vehicle speed we created short clips showing $\approx$ 3s before the vehicle reaches $1.5-3$s time-to-event. In cases when ego-vehicle was stationary the video was cropped at 3s before the pedestrian began crossing. The first and the last frames of each video clip were frozen for 4s to allow the subjects to get familiar with the scene. The pedestrian of interest was highlighted with a red arrow pointing down for the duration of freeze-frames in the beginning and at the end of the video (see Figure 7.2 for an example).

7.3.2 Procedure

We first ran the experiment in a lab setting with 5 subjects (ages $27-62$) each of whom viewed the entire set of 1842 videos. We then repeated the same experiment on Amazon Mechanical Turk (AMT) to gather additional 10 answers per video. For the AMT experiment, we grouped videos into sets of 10 for each HIT (Human Intelligence Task). We limited our study to participants residing in Canada and the USA to ensure that they are familiar with the rules of the road, signs, road delineation, etc. and to reduce any cultural bias. In total, we collected $27,630$ responses from over 700 subjects (ages $19-88$).

7.3.3 Results

A plot of aggregated responses from lab and AMT participants is shown in Figure 7.3. Since ground truth data was not available, we focused on analyzing the agreement among subjects to validate our results. First, we computed an intraclass correlation coefficient (ICC), a measure of inter-rater consistency, commonly used to analyze subjective responses from a large population of raters in the absence of ground truth data [361]. Despite an inherent degree of subjectivity of estimating pedestrian intention, the measured ICC[4] is 0.97 and 0.93 for the lab and AMT subjects respectively, which suggests a very high degree of agreement within both groups of raters (ICC = 1 for absolute agreement). The slightly lower agreement among the AMT workers is likely due to the much larger and diverse group of subjects and the presence of factors that we could not control for (e.g. viewing conditions, distractions, etc.).

Despite some noise present in the AMT data, the Pearson correlation coefficient between the average responses of the lab and AMT subjects is 0.90 suggesting that both groups answer similarly. For instance, 14 out of 15 raters agreed on the same answer in nearly 17% of cases. On the other hand, there were only 10 cases in the entire dataset where raters did not reach an agreement with respect to the pedestrian's intention, resulting in an average score of exactly 3 ('Not sure'). The samples in question included pedestrians who were close to the curb or already stepped onto the road but were distracted, e.g. by their phone or by interacting with another person. Bus stops in close proximity to the pedestrian crossings were another source of confusion, making it difficult to distinguish between pedestrians waiting for the bus and those waiting to cross. However, the number of these borderline cases was very low ($\approx 3\%$).

The PIE dataset contains 898 examples of people who intend to but do not cross, 512 pedestrians with the intention to cross who eventually cross in front of the vehicle and 430 pedestrians with no crossing intention. Interestingly, there are only 2 samples where the pedestrian crossed the street but responses from human subjects did not indicate crossing intention. Since this type of false negative is a potential safety concern, it is reassuring that human participants are particularly good at interpreting others' intentions.

[4]We use ICC(3,k) and ICC(1,k) for lab and AMT data respectively. The first measure assumes that a fixed number of raters k (in this case $k = 5$ for in-lab participants) rate all targets and the second measure assumes that a subset of k raters ($k = 10$) from a large population rates all targets. Ratings are aggregated across raters in both cases.

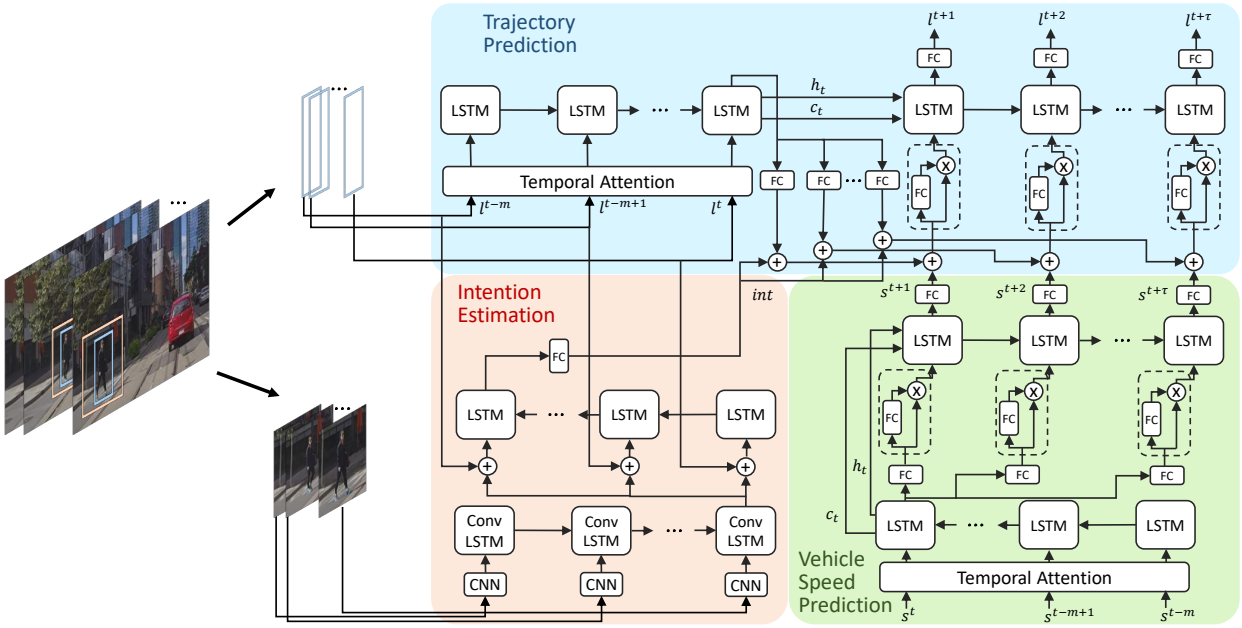


Figure 7.4: The proposed intention estimation and trajectory prediction framework. The system receives as input a sequence of images and the current speed of the ego-vehicle. The intention estimation model's encoder receives as input a square cropped image around the pedestrians, produces some representation which is concatenated with their observed locations (bounding box coordinates) before feeding them to the decoder. The speed model predicts future speed using an encoder-decoder scheme followed by a series of self-attention units. The location prediction unit receives location information as encoder input and the combination of encoder representations, pedestrian intention and future speed as decoder input, and predicts future trajectory. In the diagram, FC refers to fully-connected layers, $s^{1:m}$. $\bigoplus$ to concatenation operation and $s^{1:m}$. $\bigotimes$ to element-wise multiplication. Location, intention and speed are denoted by l, int and s respectively.

7.4 Methods for Intention Estimation and Trajectory Prediction

In this work, we address the problem of pedestrian behavior prediction on two levels: *Early anticipation* in the form of estimating pedestrians' intention of crossing and *trajectory prediction* as late forecasting of the future trajectory of pedestrians based on observed scene dynamics. The former primarily serves as a refinement procedure to change the focus of an intelligent system to those pedestrians that matter, or potentially will interact with the vehicle. Intention estimation may also benefit trajectory prediction by implying the types of motion patterns that are more probable in the scene. For instance, someone with no intention of crossing will not perform a lateral movement across the street in front of the vehicle.

7.4.1 Pedestrian Intention Estimation

We represent pedestrian intention for each sample as an average response of human experiment participants, rescaled to range $[0, 1]$ and rounded. Then we define the task as a binary classification problem of predicting whether the pedestrian i has an intention of crossing the street $int_i \in \{0, 1\}$ given a partial observation of local visual context around pedestrian $C_{obs} = \{c_i^{t-m}, c_i^{t-m+1}, ..., c_i^t\}$ and trajectory $L_{obs} = \{l_i^{t-m}, l_i^{t-m+1}, ..., l_i^t\}$, where l is a 2D bounding box around the pedestrian defined by top-left and bottom-right points $([(x_1, y_1), (x_2, y_2)])$.

It has been shown that pose, implicitly encoded in the appearance (e.g. whether the person is leaning forward or turned towards the road), immediate local surroundings (e.g. location relative to the curb) and motion, convey vital information about the intention to cross. Other context elements, such as street signs, traffic signals as well as the behavior of the ego-vehicle, may influence pedestrian's actions, e.g. whether they will attempt to cross, but will not have an effect on their initial intention to cross the street.

For the task of the intention estimation, we employ an RNN encoder-decoder architecture (see Figure 7.4), where encoder receives a sequence of feature representations corresponding to the image areas around the detected pedestrian. The output of the encoder is then concatenated with the sequence of bounding box coordinates which capture pedestrian dynamics. We use a binary cross-entropy loss function for training.

7.4.2 Pedestrian Trajectory Prediction

We address the problem of future trajectory prediction as an optimization process in which the objective is to learn the distribution $p(L_{pred}|L_{obs}, S_{pred}, Int_i)$ for multiple pedestrians $1 \leq i \leq n$, where $L_{pred} = \{l_i^{t+1}, l_i^{t+2}, ..., l_i^{t+\tau}\}$ are the predicted trajectories of pedestrians, $L_{obs} = \{l_i^{t-m}, l_i^{t-m+1}, ..., l_i^t\}$ are the observed locations of pedestrians, $S_{pred} = \{s^{t+1}, s^{t+2}, ..., s^{t+\tau}\}$ refers to predicted future speed of the ego-vehicle, and Int_i is the crossing intention of pedestrian i estimated by the intention estimation stream. The locations, l are 2D bounding boxes around pedestrians defined by top-left and bottom-right corner points $[(x_1, y_1), (x_2, y_2)]$

As depicted in Figure 7.4, the proposed model is based on an RNN encoder-decoder architecture where the inputs to the encoder are the observed locations of pedestrians for some time t and the output of the decoder is the future trajectory prediction up to time $t + \tau$. We use two types of attention: a *temporal attention* module applied to the encoder inputs and a *self-attention* unit applied to the decoder inputs. The former focuses on finding the most relevant information (key frames) in the observed sequence, whereas the latter is applied at feature-level and focuses on the parts of the encoding representation that are

relevant to current prediction. The self-attention units are preceded by embedding units for dimensionality reduction of encodings in order to minimize the effect of noise. The final predictions are generated by a linear transformation of the decoder's output.

The vehicle speed estimation stream follows a similar scheme, except it learns $p(S_{pred}|S_{obs})$, where S_{obs} refers to observed speed of the vehicle up to time t. At training time, both sequence prediction models use a mean squared error loss function defined as $MSE = \frac{1}{N}\sum_{j=1}^{\tau}||loc_i^{t+j} - \hat{loc_i}^{t+j}||$.

7.4.3 Implementation

Intention Estimation. We use Convolutional LSTM with 64 filters and kernel size of 2×2 with stride 1 as encoder and for decoder an LSTM with 128 hidden units, *tanh* activation, dropout of 0.4 and recurrent dropout of 0.2. VGG16 [362] (without fc layers) pretrained on ImageNet [363] is used to encode image features. We experiment with two different types of visual information. The first is img_{bbox} which is input image cropped to the size of the bounding box, resized so that the larger dimension matches the VGG input size of 224×224 and padded with zeros to preserve the aspect ratio. The second type of input is local context around the pedestrian ($img_{context}$) which is input image cropped to $2\times$ the size of the bounding box, squarified and resized to 224×224.

Trajectory Prediction. We use LSTMs with 256 hidden units and $softsign$ activation in our trajectory and speed prediction streams. Compared to *tanh* activation, we observed faster training and performance improvement of up to 5% when using $softsign$ activation. The embedding layer in the trajectory prediction stream is a fully-connected network with 64 output nodes and no dropout[5].

Training. Models are trained separately and combined at test time. Intention and trajectory models are trained using RMSProp [364] optimizer with learning rate of 10^{-5} and 10^{-2} respectively. The intention model was trained for 300 epochs using a batch size of 128 with $L2$ regularization of 0.001. We trained the trajectory model for 60 epochs using a batch size of 64 with $L2$ regularization of 0.0001.

[5]The full implementation can be found at https://github.com/aras62/PIEPredict.

7.5 Experimental Evaluations

7.5.1 Datasets

Pedestrian Intention Estimation (PIE). There are 1842 pedestrian samples divided into train, test and validation sets with the ratios of 50%, 40% and 10% respectively. We sample the tracks with an overlap ratio of 0.5. For trajectory prediction training, the tracks below the minimum length of 2 seconds (observation + prediction) are discarded. We use the OBD sensor readings for speed information.

Joint Attention in Autonomous Driving (JAAD). For trajectory prediction evaluation using only bounding boxes we use pedestrian tracks from the JAAD dataset. Given the smaller number of samples and shorter tracks in this dataset, we use all pedestrian samples with overlap ratio of 0.8. We use the same train/test split as in Section 6.3.2 using JAAD mix subset and exclude the low-resolution and low-visibility videos (13 out of 346) from the evaluation.

7.5.2 Metrics

For intention estimation we report *accuracy* and $F1$-score defined as $2{\cdot}precision{\cdot}recall/(precision{+}recall)$. The following metrics are used for evaluation of the proposed trajectory prediction algorithm: MSE over bounding box coordinates [280], C_{MSE} and CF_{MSE} which are the MSEs of the center of the bounding boxes averaged over the entire predicted sequence and only the last time step $(t+\tau)$ respectively. All results of the bounding box predictions are in pixels.

7.5.3 Pedestrian Intention Estimation

Method	Input data	*acc*	$F1$
LSTM	loc	0.63	0.73
LSTM_{ed}	loc	0.67	0.76
	img_{bbox}	0.60	0.78
PIE_{int}	img_{bbox}	0.69	0.79
	$img_{context}$	0.71	0.82
	$img_{bbox} + loc$	0.73	0.82
	$img_{context} + loc$	**0.79**	**0.87**

Table 7.2: Pedestrian intention estimation results for various combinations of input data: loc - bounding box coordinates, img_{bbox} - image cropped to the size of bounding box, and $img_{context}$ - image cropped to $2\times$ size of the bounding box to show local context.

Table 7.2 summarizes the results of various models trained on different combinations of input data over 0.5s of observation. The following models are used in the evaluation: a vanilla LSTM trained on normalized bounding box coordinates (loc) as a baseline, an LSTM encoder-decoder (LSTM$_{ed}$) trained on normalized bounding box coordinates or img_{bbox} and the proposed model PIE$_{int}$ trained on 4 different types of input data, img_{bbox}, $img_{context}$, $img_{bbox} + loc$ and $img_{context} + loc$.

The baseline LSTM achieves 63% accuracy. In comparison, LSTM encoder-decoder (LSTM$_{ed}$), performs better using the same information, however, it does worse using only img_{bbox} even though it has a higher $F1$-score. This can be due to the fact that pedestrian appearance in the absence of dynamics is not informative enough.

PIE$_{int}$ overall performs better than the other two models on all input types. Its performance on appearance features (img_{bbox}) and motion data (loc) is approx. 4% above the baseline performance. Adding local context ($img_{context}$) offers a small performance improvement. This suggests that, despite using different representations, motion or appearance features on their own may not be effective in estimating intention. As expected, combining different sources of information results in improved performance. We see that motion improves intention estimation on samples that are relatively far away or occluded, where visual information is unreliable. However, in situations where the pedestrian was more visible, their pose and context elements were also very important. Overall, the combination of appearance, local context and motion offer the most advantage boosting the final accuracy to 79%. Figure 7.5 shows some examples of the proposed algorithm's performance.

7.5.4 Trajectory Prediction

Method	***PIE***					***JAAD***				
	MSE			C_{MSE}	CF_{MSE}	MSE			C_{MSE}	CF_{MSE}
	0.5s	1s	1.5s	1.5s	1.5s	0.5s	1s	1.5s	1.5s	1.5s
Linear	123	477	1365	950	3983	223	857	2303	1565	6111
LSTM	172	330	911	837	3352	289	569	1558	1473	5766
B-LSTM [280]	101	296	855	811	3259	159	539	1535	1447	5615
PIE$_{traj}$	**58**	**200**	**636**	**596**	**2477**	**110**	**399**	**1248**	**1183**	**4780**

Table 7.3: Location (bounding box) prediction errors over varying future time steps. MSE in pixels is calculated over all predicted time steps, C_{MSE} and CF_{MSE} are the MSEs calculated over the center of the bounding boxes for the entire predicted sequence and only the last time step respectively.

We begin by evaluating the proposed model using only location (bounding box) information. For this purpose we report the results on the following models: two baseline models,

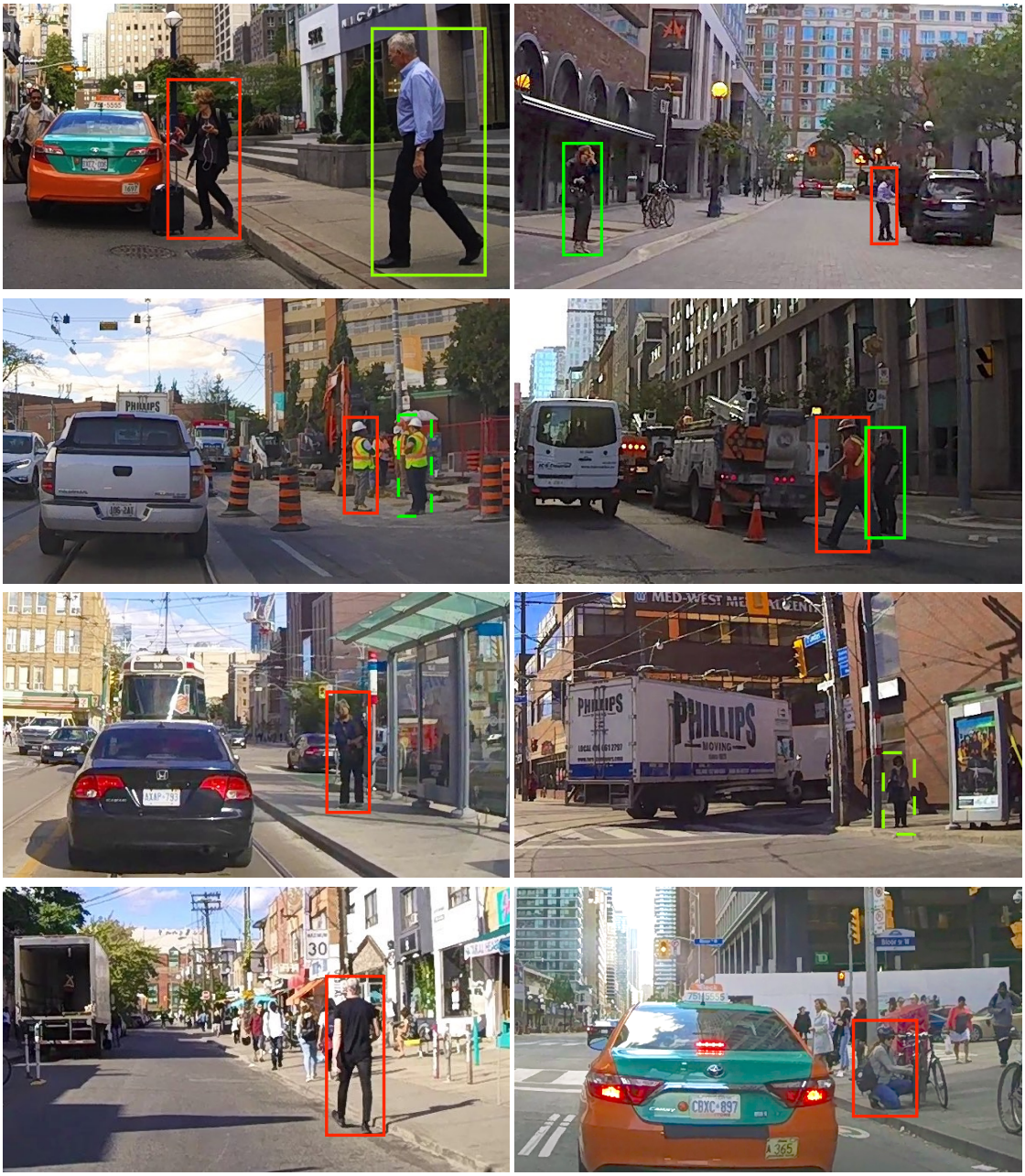


Figure 7.5: Results of pedestrian intention estimation overlaid on top of frames from the PIE dataset (cropped for better visibility). Bounding boxes are colored depending on the presence (green) or absence (red) of crossing intention as detected by our model. Dashed bounding boxes represent incorrectly estimated intention.

a linear Kalman filter [365] and a vanilla LSTM model, state-of-the-art algorithm, Bayesian LSTM [280] (B-LSTM), and the proposed model PIE$_{traj}$. Each model is trained and tested on 0.5s (15 frames) observation, and predicts trajectories over 0.5, 1 and 1.5 seconds in future.

Table 7.3 summarizes the results of the predictions using only bounding box information. As shown in the table, the proposed method achieves state-of-the-art performance on all metrics, by up to 26% on the PIE dataset and 18% on JAAD compared to B-LSTM. The performance of all models is generally poorer on the JAAD dataset which can be partially attributed to the smaller number of samples, scales and shorter tracks all of which reduce the diversity of the dataset. The deterioration of linear model performance for long-term predictions indicates the complexity of human motion patterns that cannot be explained with simple linear interpolation. As expected, the performance of all models is generally better on bounding box centers due to the fewer degrees of freedom.

7.5.5 Ego-Vehicle Speed Prediction

	MSE			
Method	0.5s	1s	1.5s	last
Linear	0.87	2.28	4.27	10.76
LSTM	1.50	1.91	3.00	6.89
PIE$_{speed}$	**0.63**	**1.44**	**2.65**	**6.77**

Table 7.4: Speed prediction errors over varying time steps on the PIE dataset. *Last* stands for the last time step. The results are reported in km/h.

We first evaluate the proposed speed prediction stream, PIE$_{speed}$, by comparing this model with two baseline models, a linear Kalman filter and a vanilla LSTM model. We use MSE metric and report the results in km/h. Table 7.4 shows the results of our experiments. The linear model achieves reasonable performance in short-term which is better than the vanilla LSTM over 0.5s. This indicates that the speed variation often is insignificant in the short-term, especially in urban environments which is the case in the proposed PIE dataset. In long-term, however, LSTM-based models perform significantly better. The proposed PIE$_{speed}$ achieves the best performance by up to 10% over vanilla LSTM model.

7.5.6 Intention in Trajectory Prediction

Earlier we argued that pedestrian intention can serve as an early prediction stage in addition to trajectory prediction. Here, we examine whether estimating pedestrians' intention of crossing can improve trajectory prediction. We report the results on our trajectory prediction model PIE$_{traj}$ which receives as input the context information provided by PIE$_{speed}$ and PIE$_{int}$. We report the results on 0.5s observation and 1.5s prediction.

As shown in Table 7.5, conditioning trajectory prediction on pedestrian intentions can improve the results by up to 4%. This is due to the fact that intention may imply certain

Method	Input	MSE	C_{MSE}	CF_{MSE}
PIE_{traj}	loc	636	596	2477
	loc+PIE_{int}	611	570	2414
	loc+PIE_{speed}	572	535	2204
	loc+PIE_{int}+PIE_{speed}	**559**	**520**	**2162**
	$loc + int + speed$	473	435	1741

Table 7.5: Location (bounding box) prediction errors of the proposed model PIE_{traj} on 0.5s observation and 1.5s prediction using different inputs. *loc*, *int* and *speed* stand for location, intention and vehicle speed. PIE_{int} and PIE_{speed} are the outputs of the intention and vehicle speed estimation models. MSE is reported in pixels and calculated over all predicted time steps. C_{MSE} and CF_{MSE} are the MSEs over the center of the bounding boxes for the entire predicted sequence and only the last time step respectively.

patterns of motion. For instance, someone with the intention of crossing might have a lateral movement across the street whereas someone without intention might stand still. As one would expect, the ego-vehicle's speed improves the trajectory prediction, and when combined with pedestrian intention, the best results are achieved with more than 11% improvement over baseline using only bounding boxes.

Figure 7.6 illustrates the performance of our proposed algorithm using different contextual information on the PIE dataset. Even though speed has a dominant effect in improving trajectory prediction it may also fail in certain cases, when the vehicle is stationary or when the pedestrian has no intention of crossing.

7.6 Summary

In this chapter, we presented a novel large-scale dataset for studying pedestrian crossing intention and behavior with extensive multimodal annotations for visual reasoning tasks. Since there is no ground truth data for crossing intention, we conducted a large-scale experiment to determine human reference data for this task. Our data shows that a large number of human experiment subjects have a high degree of agreement in their answers.

As part of this work, we proposed a trajectory prediction algorithm for an on-board camera. Our model outperforms the state-of-the-art by a significant margin. In addition, we proposed a baseline intention estimation model and by evaluating various input data combinations we showed that the appearance of pedestrians and their local surrounding context in conjunction with the changes in their movements are good predictors for estimating crossing intention. In the end, we presented empirical results that suggest that combining various sources of information such as ego-vehicle speed and pedestrian intention with motion history can improve the performance of the trajectory prediction algorithm.

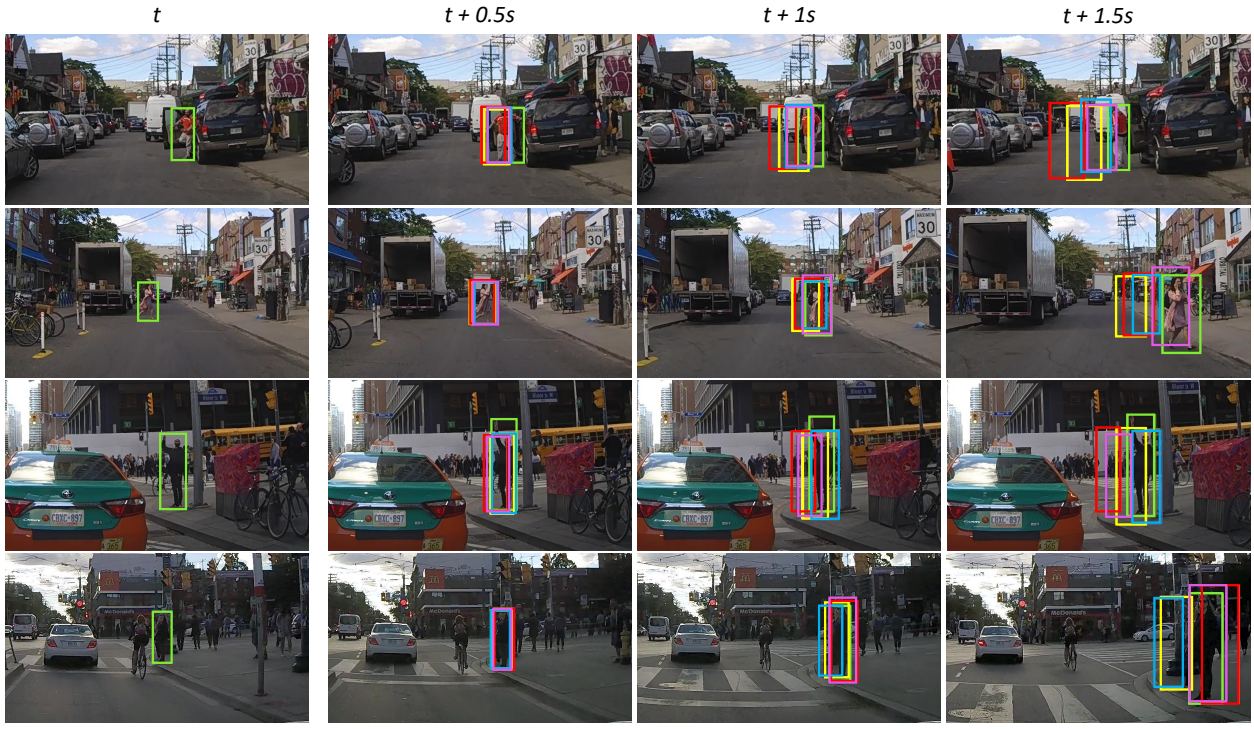


Figure 7.6: Examples of trajectory prediction algorithm using the proposed model PIE$_{traj}$ with different input combinations. The color and model combinations are: *loc* (yellow), *loc*+PIE$_{int}$ (blue), *loc*+PIE$_{speed}$ (red), and *loc*+PIE$_{int}$+PIE$_{speed}$ (purple). Ground truth annotations are shown in green. The sequences depict different traffic scenarios. From top to bottom: A man leaving his vehicle, a woman crossing the street, a man hailing a taxi, and a woman waiting to cross.

Chapter 8

Anticipating Pedestrian Crossing Action Using Contextual Cues

Figure 8.1: Examples of pedestrians prior to making crossing decisions. Green and red colors indicate whether the pedestrian will or will not cross.

In Chapter 7, we talked about trajectory prediction algorithms as one of the important components for planning in autonomous driving systems. Another common approach for planning is event prediction, in particular, predicting pedestrian crossing actions. This allows autonomous cars to make an assessment of pedestrians' behavior and act accordingly.

Similar to the trajectory prediction tasks, merely relying on pedestrian dynamics is not sufficient for making sense of pedestrian behavior and predicting their upcoming actions as they are often subject to error. For example, a pedestrian intending to cross the street could be standing at the intersection (with no motion history), walking alongside the road or abruptly changing their walking pattern prior to crossing [130] (see Figure 8.1). In addition, pedestrians exhibit highly variable motion patterns which can be influenced by various environmental factors such as signals, the ego-vehicle motion, road structure, etc. All of these factors add to the complexity of predicting pedestrian actions (see Chapter 3 for more

information). Thus statistical inference on pedestrian trajectories alone may not be sufficient for predicting their actions.

In this chapter we examine the role of context on pedestrian action prediction. Given that the main point of interaction between autonomous vehicles and pedestrians is at the time of crossing, here we particularly focus on the pedestrian crossing anticipation, i.e. we determine whether an observed pedestrian will cross in front of the vehicle. For this purpose, we discuss a subset of contextual information that can potentially impact pedestrian behavior and analyze various computational models to incorporate such information for reasoning about future actions.

8.1 A Review of Action Prediction Algorithms

Action prediction algorithms can be categorized into two groups: Next action or event prediction (or action anticipation) [366, 367, 368, 369, 370, 371], and early action prediction [372, 373, 374, 375, 323, 376, 377, 378, 379, 380]. In the former category, the algorithms use the observation of current activities or scene configurations and predict what will happen next. Early action prediction algorithms, on the other hand, observe parts of the current action in progress and predict what this action is. Our focus will be on the former category in which we intend to predict what will happen next without observing any parts of the upcoming action. Some recent works use traditional learning methods for behavior prediction [381, 382, 383, 384, 385, 386], however, given the dominance of deep learning algorithm, we focus our review on this class of algorithms.

Action prediction algorithms are used in a wide range of applications including cooking activities [366, 367, 368, 387, 369, 370, 371, 388], traffic understanding [389, 390, 391, 392, 393, 392, 394, 341, 395, 396], accident prediction [397, 398, 399, 400], sports [401, 402] and other forms of activities [305, 372, 403, 404, 405, 406, 407, 408, 409, 386]. Although the majority of these algorithms use sequences in which the objects and agents are fully observable, a number of methods rely on egocentric scenes [366, 367, 370, 403, 401, 386] which are recorded from the point of view of the acting agents and only parts of their bodies (e.g. hands) are observable.

The methods used in action prediction predominantly use a variation of RNN-based architectures including LSTMs [305, 367, 368, 387, 392, 369, 370, 398, 403, 393, 392, 404, 405, 401, 400, 402, 388, 407, 410, 395, 396], GRUs [372, 389, 371], ConvLSTMs [394], and Quasi-RNNs (QRNNs) [399]. For instance, in [372, 369] the authors use a graph-based RNN architecture in which the nodes represent actions and the edges of the graph represent the transitions between the actions. The method in [371] employs a two-step approach: using a

recognition algorithm, the observed actions and their durations are recognized. These form a one-hot encoding vector which is fed into GRUs for the prediction of the future activities, their corresponding start time and length. In the context of vehicle behavior prediction, Ding et al. [389] uses a two-stream GRU-based architecture to encode the trajectory of two vehicles and a shared activation unit to encode the vehicles mutual interactions. Scheel et al. [411] encode the relationship between the ego-vehicle and surrounding vehicles in terms of their mutual distances. The vectorized encoding is then fed into a bi-directional LSTM. At each time step, the output of the LSTM is classified, using a softmax activation, into a binary value indicating whether it is safe for the ego-vehicle to change lane. In [399] the authors use a QRNN network to capture the relationships between road users in order to predict the likelihood of a traffic accident. To train the model, the authors propose an adaptive loss function that assigns penalty weights depending on how early the model can predict accidents.

As an alternative to recurrent architectures, some algorithms use feedforward architectures using both 3D [366, 390, 391] and 2D [397, 371, 341, 402, 406, 408, 409, 386] convolutional networks. For example, in the context of pedestrian crossing prediction, in [390] the authors use a generative 3D CNN model that produces future scenes and is followed by a classifier. The method of [391] detects and tracks pedestrians in the scenes, and then feeds the visual representations of the tracks, in the form of an image sequence, into a 3D CNN architecture, which directly classifies how likely the pedestrian will cross the road. To predict the time of traffic accidents, the method in [397] processes each input image using a 2D CNN model and then combines the representations followed by a fully-conntected (*fc*) layer for prediction. Farha et al. [371] create a 2D matrix by stacking one-hot encodings of actions for each segment of observation and use a 2D convolutional net to generate future actions encodings. Casas et al. [341] use a two-stream 2D CNN, each processing the stacked voxelized LIDAR scans and the scene map. The feature maps obtained from each stream are fused and fed into a backbone network followed by three headers responsible for the detection of the vehicles and predicting their intentions and trajectories. For sports forecasting, Felsen et al. [402] concatenate 5 image observations channel-wise and feed the resulting output into a 2D CNN network comprised of 4 convolutional layers and an *fc* layer.

Although some algorithms rely on a single source of information, e.g. a set of preprocessed features from RGB images [372, 398, 399, 405, 400, 402, 388, 407, 410] or trajectories [389], many algorithms use a multimodal approach by using various sources of information such as optical flow maps [367, 369, 370, 393, 386], poses [305, 369, 404, 396], road structure [403, 341], text [368], speed (e.g. ego-vehicle or surrounding agents) [392, 393, 411, 401], and gaze [403, 404]. For example, Gammulle et al. [387] propose a two-stream LSTM net-

work with external neural memory units. Each stream is responsible for encoding visual features and action labels. Farha et al. [371] use a two-layer stacked GRU architecture which receives as input a feature tuple of the length of the activity and its corresponding one-hot vector encoding. In [393], the method uses a two-stage architecture: First information regarding the appearance of the scene, optical flow (pre-processed using a CNN) and vehicle dynamics are fed into individual LSTM units. Then, the output of these units is combined and passed through an *fc* layer to create a representation of the context. This representation is used by another LSTM network to predict future traffic actions. Jain et al. [396] use a fusion network to combine head pose information of the driver, outside scene features, GPS information, and vehicle dynamics to predict the driver's next action.

Similar to trajectory prediction algorithms, in the field of action anticipation, recurrent architectures are strongly preferred. Compared to feedforward algorithms, recurrent methods have the flexibility of dealing with variable observation lengths and multi-modal data, in particular, when they are significantly different, e.g. trajectories and poses.

In terms of the use of context, many of the pedestrian behavior prediction algorithms for traffic scenes rely on dynamic information. Those that include visual context either focus on very limited information, e.g. head orientation, or encode the entire scene which can be quite noisy and uninformative for reasoning. Here, we focus on a subset of visual context and encode what may matter for reasoning, both implicitly and explicitly.

8.2 Anticipating Crossing Using Multi-Modal Data Fusion

We define pedestrian crossing prediction as a binary classification problem in which the objective is to determine whether a pedestrian i will cross the street given the observed context up to some time m. The prediction relies on five sources of information including the local context $\{C_{p_i}, C_{s_i}\}$, where $C_{p_i} = \{c_{p_i}^1, ..., c_{p_i}^m\}$ and $C_{s_i} = \{c_{s_i}^1, ..., c_{s_i}^m\}$ refer to visual features of the pedestrian and their surroundings respectively, pedestrian pose $P_i = \{p_i^1, ..., p_i^m\}$, 2D bounding box locations $B_i = \{b_i^1, ..., b_i^m\}$, where b_i is a two-point coordinate $[(x1_i, y1_i)(x2_i, y2_i)]$ corresponding to the top-left and bottom-corner of the bounding box around the pedestrian, and the speed of the ego-vehicle $S = \{s^1, ..., s^m\}$.

8.2.1 Context for crossing prediction

In this section we describe how we incorporated some of the contextual elements that we discussed in Chapter 3. It should be noted that due to algorithmic limitations and lack

of data we cannot include all these factors. For example, identifying pedestrians abilities to asses the environment, their belief system or how long they have been waiting prior to crossing is not possible to detect. In addition, our dataset does not have 3D information which limits our abilities to reason about traffic flow, distance, and other dynamic factors. Annotated data is also limited to pedestrians who likely want to cross. This means that training a model to explicitly detect all traffic objects and reason about the relationships between them is extremely challenging.

To accommodate these limitations, we define five sources of information to capture many of contextual elements discussed in Chapter 3.

Local context refers to the visual representations of pedestrians and their surroundings. We make use of pedestrians' appearances within the predefined bounding boxes. This information reflects pedestrian *age*, *state*, *gender* or any other relevant attributes. We also consider a region around the pedestrians proportional to the size of their bounding boxes. The size of the region is set in a way to include information regarding the presence of *signals*, *zebra crossing lines*, pedestrians *proximity to curbs*, and whether or not the pedestrians are in a *group*. By relying on visual representations in this way, we intend to implicitly identify the potential relevant components.

At each time step of the observation, for each pedestrian, we use their appearance and surroundings. The former is captured using images cropped to the size of 2D bounding boxes around the pedestrian in the frame. For the surroundings, we extract a region around the pedestrian by scaling up the 2D bounding box coordinates, and squarifying the dimensions so the width of the scaled bounding box matches its height. This gives us a wider viewing angle of the scene around the pedestrian which may include street, other pedestrians, signals or traffic. In the surround crop, we suppress the pedestrian appearance by setting the pixel values in the original bounding box coordinate to neutral gray. Both appearance and surround crops are processed using a convolutional neural network (CNN) which produces two feature vectors $vc_p^{1:m}$ and $vc_s^{1:m}$.

Pose. We explicitly make use of pedestrians' poses. The pose information shows whether the pedestrian is *looking* at the traffic or distracted, whether she is *walking*, *standing*, or *sitting* and to some extent can capture certain *disabilities*.

The pose network used for this purpose generates 18 body joints coordinates, each corresponding to a point in 2D space, for each pedestrian. The joint coordinates are normalized and concatenated into a 36D feature vector $vp^{1:m}$.

2D bounding box. The bounding box information captures the *trajectories* of pedestrians in the scene. In addition, using box coordinates provides a sense of scale which helps estimate *distances* of pedestrians to the ego-vehicle. In conjunction with the speed of the

ego-vehicle, *time to collision (TTC)* can also be approximated.

We transform the bounding box coordinates into relative displacement from the initial position forming a feature vector $vb^{1:m}$. This can be seen as the velocity of the pedestrian at every time step.

Speed of the ego-vehicle is necessary to reason about the *dynamics* of the scene which has a direct impact on pedestrian behavior. As was mentioned in Chapter 3 and 4, changes in the speed of the vehicle can be an indicator that the driver is *communicating* his intention to pedestrians.

We present this information as a vector of the ego-vehicle speed recordings for each time step $vs^{1:m}$ in km/h.

8.2.2 Architecture

Recurrent neural networks (RNNs) are extensions of feedforward networks. RNNs have recurrent hidden states allowing them to learn temporal dependencies in sequence data. This inherent temporal depth has been shown to greatly benefit tasks, such as pedestrian trajectory prediction, that apply single-layer RNNs to point coordinates in a space. In addition to temporal depth, the spatial depth of RNNs can also be increased by stacking multiple layers of RNN units on top of one another. This approach is an effective way of improving sequential data modeling in complex tasks [412], in particular, video sequence analysis [413, 414] in which the network models dependencies between visual features of consecutive video frames.

Given the multimodal nature of pedestrian action anticipation which relies on both dynamics and visual scene information, we employ a hybrid approach. We use a stacked RNN architecture similar to [414] in which we gradually fuse the features at each level according to their complexity. In other words, we input the visual features of the scene that can benefit more from spatial depth of the network at the bottom layers and the dynamics features, e.g. trajectories and speed, at the higher levels of the network (see Figure 8.2).

Multimodal feature fusion. For the joint modeling of our sequence data, we use gated recurrent units (GRUs) [415] which are simpler compared to LSTMs and, in our case, achieve similar performance. Recalling the equation of GRU, the j^{th} level of the stack is given by,

$$\begin{aligned} r_j^t &= \sigma(W_j^{xr} x_j^t + W_j^{hr} h_j^{t-1}), \\ z_j^t &= \sigma(W_j^{xz} x_j^t + W_j^{hz} h_j^{t-1}), \\ \tilde{h}_j^t &= tanh(W_j^{xh} x_j^t + W_j^{hh}(r_j^t \odot h_j^{t-1}), \\ h_j^t &= (1 - z_j^t) \odot h_j^{t-1} + z_j^t \odot \tilde{h}_j^t), \end{aligned} \tag{8.1}$$

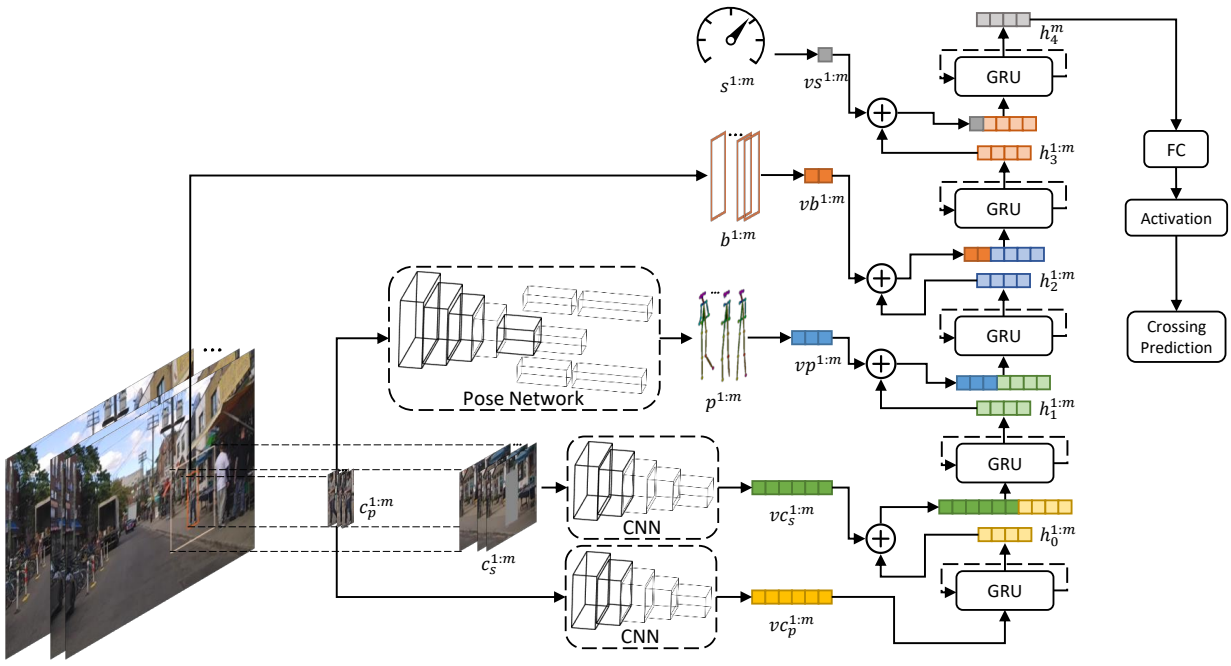


Figure 8.2: The architecture of the proposed algorithm *SF-GRU* comprised of five GRUs each of which processes a concatenation of features of different modalities and the hidden states of the GRU in the previous level. The information is fused into the network gradually according to the complexity of the features. Each feature input consists of m sequential observations. From bottom to top layers features are fused as follows: pedestrian appearance $c_p^{1:m}$, surrounding context $c_s^{1:m}$, poses $p^{1:m}$, bounding boxes $b^{1:m}$ and ego-vehicle speed $s^{1:m}$. $\bigoplus$ refers to concatenation operation.

where $\sigma(.)$ is the sigmoid function, r^t and z^t are reset and update gates, and matrices $W^{\cdot\cdot}$ are weights between two units. For $j = 0$ (the bottom level of the stack), $x_0^t = vc_p^t$ and for $j > 0$, $x_j^t = h_{j-1}^t + vy^t[j-1]$ where $y^t = \{vc_s^t, vp^t, vb^t, vs^t\}$. The final prediction is achieved by a linear transformation of h_n^t where n is the number of levels (in our case 5) in the proposed stacked architecture. In the training phase we use the *binary cross-entropy* loss function.

8.2.3 Implementation

In our architecture, we use GRUs [415] with 256 hidden units. For local context, we crop the pedestrian samples C_p using the 2D bounding box annotations, resize them so the larger dimension is equal to 224 and pad them with zeros to preserve the aspect ratio. For surround context, C_s, we use $2.5x$ (set empirically) scaled version of the 2D bounding boxes. The parts of the cropped images that include pedestrians of interest are suppressed by neutral gray with RGB of $(128, 128, 128)$. We resize these images to 224×224. The local context images are processed using VGG16 [362] (without fully-connected (fc) layers) pretrained on ImageNet

[363] followed by an average global pooling generating a feature vector of size 512 per crop. For pedestrian poses, we use [416] which is pretrained on the COCO dataset [417]. The network generates 18-joint pose per pedestrian sample [1].

Training. The model is trained using ADAM [418] optimizer with a learning rate of 5×10^{-6} for 60 epochs with batch size of 32 and $L2$ regularization of 0.0001. The context and pose features are precomputed. In addition, we augment the data at training time by horizontally flipping the images and sub-sampling the over-represented class to equalize the number of crossing and non-crossing samples.

8.3 Experimental Evaluations

8.3.1 Dataset

There are not many datasets suitable for the purpose of pedestrian crossing prediction. The JAAD dataset introduced in Section 4.3.1 contains videos of pedestrians prior or during crossing. Unfortunately, the number of samples in this dataset is small, no vehicle information is available and sequences are short snippets which are not suitable for long-term predictions. Therefore, we only use our PIE dataset (see Section 7.2) which comprises 1842 pedestrian tracks captured using an on-board monocular camera while driving in urban environments with various street structures and crowd densities. The samples represent people who are close to the curbs or are at intersections and may or may not have the intention of crossing, e.g. waiting for a bus. Overall, the ratio of non-crossing to crossing events is 2.5 to 1. All video sequences are collected during daylight under clear weather conditions. The videos are continuous allowing us to observe the pedestrians from the moment they appear in the scene until they go out of the field of view of the camera.

For each pedestrian sample we identified an event point. For those who cross in front of the ego-vehicle, the event is the moment they start crossing. For other samples, the events are set at the time when the pedestrians go out of the field of view of the camera. We randomly split the data into train-test sets with ratio of 60-40 respectively.

8.3.2 Metrics

As in [399], we report all the evaluation results using the following metrics: *Accuracy*, *F1* score, *precision* and *recall*. We also use *Area Under Curve (AUC)* metric which, in the case of binary event anticipation, reflects the balanced accuracy of the algorithms.

[1] The full implementation can be found at https://github.com/aras62/SF-GRU.

8.3.3 Predicting Crossing Events

We evaluate the performance of our proposed algorithm, stacked with multilevel fusion GRU (***SF-GRU***), against baseline models and state-of-the-art sequence analysis approaches:

Static. This model serves as a baseline. It has two VGG16 branches (without *fc* layers and with a global pooling layer) pretrained on ImageNet. One network processes the local context corresponding to the pedestrian crop c_p^m and the other processes the surroundings c_s^m at the last frame of the observation. The outputs of both networks are combined and fed into a *fc* layer for the final prediction.

GRU. A single-layer GRU [415] trained and tested only on pedestrians' appearances C_p and their surroundings C_s. We also use this model with all sources of information which are concatenated and fed into the network at the same time.

Multi-stream GRU (M-GRU). Following the approach in [280], this architecture processes different types of features separately using different GRUs, and feeds the concatenation of the last hidden states of all GRUs into a dense layer for prediction.

Hierarchical GRU (H-GRU). This model has a hierarchical structure similar to [419]. H-GRU processes each feature type using a separate GRU, concatenates the hidden states of all units and then feeds them into another GRU whose last hidden state is used for prediction.

Stacked GRU (S-GRU). This is a five-level stacked GRU architecture as described in [414] which receives the feature inputs at the bottom layer. The inputs to the subsequent GRUs in the higher levels are the hidden states of the GRUs in the previous layers.

All evaluations are done on observation sequences of 0.5s (15 frames) duration. The samples are selected with 2s time to event (TTE), the minimum time within which pedestrians make crossing decisions according to [130].

Models	*Features*	*Acc*	*AUC*	*F1*	*Prec*	*Recall*
Static	c_p^m, c_s^m	0.592	0.589	0.419	0.328	0.582
GRU	C_p, C_s	0.681	0.644	0.475	0.407	0.570
GRU	C_p, C_s, P, B, S	0.811	0.812	0.685	0.593	**0.812**
M-GRU	C_p, C_s, P, B, S	0.804	0.792	0.665	0.585	0.770
H-GRU	C_p, C_s, P, B, S	0.819	0.805	0.685	0.612	0.776
S-GRU	C_p, C_s, P, B, S	0.801	0.770	0.643	0.588	0.709
SF-GRU (ours)	C_p, C_s, P, B, S	**0.844**	**0.829**	**0.721**	**0.657**	0.800

Table 8.1: Evaluation results of the algorithms using observation length of 0.5s and time to event (TTE) of 2s. Abbreviations in *features* column are: pedestrian appearance C_p, surround context C_s, pose P, bounding box B, and ego-vehicle speed S. c_p^m and c_s^m stand for appearance and surround context in the last observation frame respectively.

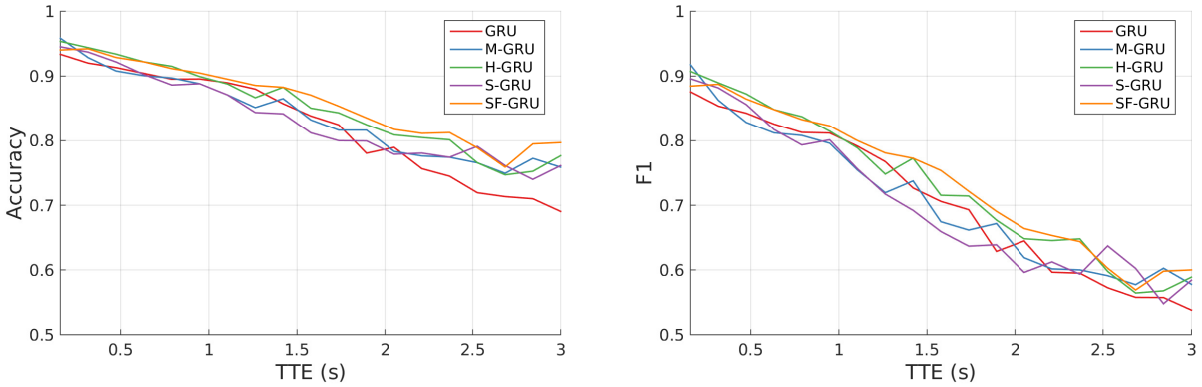
Figure 8.3: Examples of the predictions produced by the proposed algorithm *SF-GRU* and top competing methods, namely *GRU*, *M-GRU*, *H-GRU*, and *S-GRU*. In the examples, *GT* stands for ground truth and *green* and *red* colors indicate whether the pedestrian will cross in front of the ego-vehicle or not respectively. The instances where the color of the algorithm labels matches the GT means that their predictions are correct.

The results are summarized in Table 8.1. We can see that using the visual information of the local context, even as a single image in the static method can lead to approximately 60% accuracy which can be improved by 9% by performing temporal reasoning using a GRU.

Using all sources of information, the proposed algorithm *SF-GRU* performs best on all metrics except recall. For this metric single-layer *GRU* performs slightly better (by 1.2%) at the expense of more than 6% drop in precision. In addition, the results show that no performance improvement is achievable by simply adding layers to the network or separating the processing of features with different modalities. Example predictions made by the *SF-GRU* method are shown in Figure 8.3.

8.3.4 When to Predict Crossing Events

The prediction of crossing events may vary depending on TTE as the scene dynamics changes, in particular, when the ego-vehicle motion impacts the way people make a crossing decision. Here we examine the prediction ability of the temporal algorithms with respect to TTE. We alter TTE points from 0s to 3s with steps of approximately 0.16s, a total of 19 different points. To maintain the consistency of data across different time frames, we only sample from pedestrian tracks equal to or longer than 3.5s (the maximum TTE time in the experiment + observation length). All other parameters including the observation sequence length remain the same as before.



(a) (b)

Figure 8.4: The performance of the algorithms with respect to varying time to event (TTE) points with 0.5s observation length.

The proposed algorithm *SF-GRU* performs best for the most part at different TTE points (see Figure 8.4). At early TTE times where the intention of pedestrians becomes obvious, all algorithms perform similarly well. However, as expected, the performance of the algorithms degrades gradually (some at a faster rate than others) as the observations are moved further away from the time of the event. We can also see that the single-layer *GRU* only performs better than *M-GRU* and *S-GRU* up to 2s TTE after which its performance drops rapidly.

8.3.5 The Effect of Observation Length on Prediction

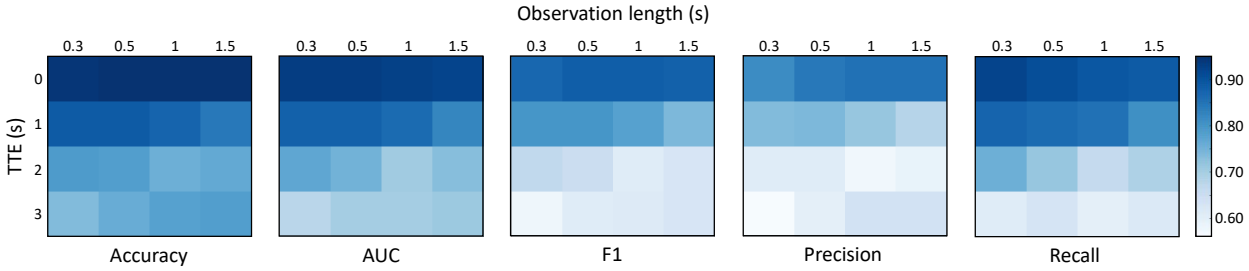


Figure 8.5: The changes in the performance of *SF-GRU* according to varying observation length and time to event (TTE).

Longer observation time can potentially provide more information but at the same time may add noise. We examine the effect of observation length on the proposed algorithm *SF-GRU* with respect to different TTE points. For the same reason as mentioned in the previous experiment, we only sample from tracks with length equal to or longer than 4.5s (the longest observation length + the largest TTE value). In total, we examine 16 different combinations.

As shown in Figure 8.5, on most metrics the improvement gain is only on samples very close (0s) or far away (3s) from the event. In these cases, precision can be improved by longer observations at the expense of reducing the recall. In critical decision regions of $1-2$s, however, a small gain is achieved by increasing observation from 0.3s to 0.5s after which point the performance drops rapidly. This could be due to noise in longer observations caused by accumulation of the changes in the scene dynamics. For instance, within 1.5s observation window, the speed of the vehicle can change significantly which can have a considerable effect on predicting pedestrian crossing behavior.

8.3.6 Feature Types and Prediction Accuracy

Features	*Acc*	*AUC*	*F1*	*Prec*	*Recall*
C_p	0.660	0.622	0.448	0.380	0.546
C_{p+s}	0.666	0.650	0.483	0.397	0.618
C_p, C_s	0.692	0.645	0.475	0.417	0.552
C_p, C_s, P	0.745	0.705	0.554	0.498	0.624
C_p, C_s, P, D	0.796	0.765	0.636	0.580	0.703
C_p, C_s, P, B	0.816	0.781	0.661	0.619	0.709
C_p, C_s, P, B, S	**0.844**	**0.829**	**0.721**	**0.657**	**0.800**

Table 8.2: The impact of different sources of information on the performance of *SF-GRU*. The feature types are as follows: C_p pedestrian context (appearance), C_s surround context, C_{p+s} full context , P pose, D displacement (center coordinates), B bounding box, and S speed.

We examine the contribution of each feature type on the performance of the proposed algorithm. In addition to the features discussed earlier, we also evaluate two other types of features: displacement D (the center coordinates of the bounding boxes) and full context C_{p+s} which is the pedestrian appearance and surround context in a single frame, not as decoupled features as proposed earlier.

As shown in Table 8.2, we can see that adding contextual information in addition to pedestrian appearance to the network improves the overall performance by more than 18%. We also see that decoupling appearance and surround context boosts the accuracy by almost 3% owing to precision gain. Another observation is that using bounding box coordinates instead of center coordinates improves the results by 2%. This can be due to the fact that the changes in the scale of the bounding boxes in a sequence can add another layer of information, e.g. the movement of pedestrian or the changes in their distance to the ego-vehicle.

8.3.7 The Order of Fusion and Performance

Features	*Acc*	*AUC*	*F1*	*Prec*	*Recall*
P, S, B, C_p, C_s	0.753	0.737	0.590	0.509	0.703
S, B, C_p, C_s, P	0.784	0.759	0.624	0.557	0.709
B, C_p, C_s, P, S	0.798	0.776	0.647	0.579	0.733
S, C_p, C_s, P, B	0.810	0.785	0.661	0.602	0.733
C_p, B, C_s, S, P	0.813	0.803	0.679	0.619	0.788
C_p, C_s, P, B, S	**0.844**	**0.829**	**0.721**	**0.657**	**0.800**

Table 8.3: Feature fusion strategies and their impact on the performance of the proposed algorithm *SF-GRU*. The feature types are as follows: C_p pedestrian context (appearance), C_s surround context, P pose, B bounding box, and S speed.

In this experiment, we investigate how different fusion strategies alter the performance. Since reporting on all possible permutations of different sources of information is prohibitive, we only include a subset of these permutations to show the fluctuations in the overall performance.

A summary of the results is provided in Table 8.3. Here, it is shown that when more complex features such as local context are infused into higher levels of the network, the performance gets worse. By inputting different feature types in the right order, that is by moving simpler features, such as speed, to the higher levels of the stack, the performance improves by up to 9% on accuracy, 10% on recall and more than 15% on precision. This can be due to the fact that more complex visual features, which benefit more from deeper spatial analysis, are inputted at the bottom layers of the network while simpler features such as trajectory coordinates are entered at the higher levels.

8.4 Does Intention Help Action Prediction?

In Chapter 7 we discussed the pedestrian intention factor and showed how estimating intention of pedestrians can improve predicting their trajectories. In this section, we examine whether intention can benefit the task of pedestrian crossing prediction. For this purpose, we use a simplified version of the model presented earlier.

We use a single layer GRU architecture which as input receives a feature vector generated by concatenating surround context features C_s as before but instead keep the pedestrian visual features included, normalized bounding boxes, speed of the ego-vehicle and intention scores. The last hidden state of the GRU is fed into a *fc* layer to reduce the dimensionality of representations before entering another *fc* for the final classification of actions. The training and testing samples are selected the same as before using the default data split as in Section

Features	Acc	AUC	F1	Precision	Recall
C	0.60	0.58	0.55	0.56	0.57
C+B	0.78	0.78	0.75	0.74	0.78
C+B+S	0.80	0.81	0.78	0.76	0.80
C+B+S+Int	**0.83**	**0.85**	**0.81**	**0.79**	**0.85**

Table 8.4: Pedestrian crossing action prediction using different contextual features, namely Context (C), Bounding boxes (B), Speed (S) and Intention (Int).

7.5.1. The learning rate is set to 10^{-6} and the model is trained for 100 epochs using a batch size of 32 and $L2$ regularization of 0.0001.

A summary of action prediction results is presented in Table 8.4. As we can see, using intention probability as input, the best results on all metrics are achieved. In our experiment, accuracy is improved by 3% and recall by 5%. The balanced accuracy represented by Area Under the Curve (AUC) is improved by 4% showing that intention can be helpful for the correct prediction of both non-crossing and crossing events.

8.5 Summary

In this chapter, we examined the role of different contextual information and architectural design approaches to pedestrian crossing prediction. We presented a novel stacked RNN architecture in which different sources of contextual information including pedestrian and the vehicle dynamics, pedestrians' appearances and their surroundings are fused gradually at different levels of processing. Using empirical evaluations we showed that the proposed approach performs best compared to alternative RNN architectures.

In addition, we demonstrated how different sources of contextual information and data fusion strategies within the network can impact crossing action prediction. We highlighted that the performance of action prediction algorithms can be improved when adding more complex features to the bottom layers of the network and the simpler ones at the higher levels. Although the proposed architecture was presented in the context of pedestrian crossing prediction, other applications of similar nature e.g. activity recognition may also benefit from using this approach.

Chapter 9

Final Remarks

9.1 Dissertation Summary

Social interaction with traffic participants, in particular pedestrians as the most vulnerable ones, is fundamental for autonomous driving systems designed for urban environments. To effectively interact with pedestrians, it is necessary to understand the behavior and what they are going to do next. In this dissertation, we focused on the problem of pedestrian behavior understanding and prediction in traffic scenes. In particular, we studied the effect of context on pedestrian behavior and investigated the importance of including contextual information in developing behavior prediction algorithms.

We began by discussing theoretical foundations of social interaction and coordination, and based on past behavioral and philosophical studies, argued that behavior prediction and understanding one's intentions are necessary in social interactions.

A meta-analysis of the impact of traffic context on pedestrian behavior prediction was presented. This study identified a large number of factors that influence pedestrian behavior and showed how these factors are interconnected and in what ways they can impact the future behaviors of pedestrians.

We further investigated driver-pedestrian communication as one of the factors that influence pedestrian behavior. We analyzed the ways pedestrians transmit their intentions and the factors that impact the likelihood of pedestrians communicating. For the purpose of this study, we introduced a large-scale dataset of pedestrian crossing events in traffic scenes. We call this dataset Joint Attention in Autonomous Driving or JAAD. The data was annotated with bounding box and behavioral information which make the dataset suitable for both behavioral studies and developing practical applications.

Using the JAAD dataset, we conducted an empirical study of pedestrian crossing actions. We highlighted what behaviors pedestrians exhibited at the time of crossing, identified the

contextual factors that impact pedestrian crossing decision-making processes and discussed the challenges in designing practical systems capable of predicting pedestrian crossing action.

Next, we focused on the practical aspects of pedestrian behavior prediction. We started by evaluating the performance of pedestrian detection algorithms. We showed how various data properties (e.g. lighting conditions or pedestrian attributes) impact the performance of these algorithms, and how diversifying the data can improve the generalizability of pedestrian detection algorithms. As part of this evaluation, we provided approximately 900k attribute labels for samples from JAAD.

For practical model of pedestrian behavior prediction, we began by framing it as the problem of future trajectory prediction. As part of this work, we introduced a novel large-scale dataset, called Pedestrian Intention Estimation (PIE), comprised of 6 hours of driving footage with annotated traffic object elements, behavioral information as well as vehicle sensor data. Using the PIE dataset, we conducted an extensive human experiment in which we asked the subjects to rank the crossing intention of pedestrian observed in videos. Moreover, we proposed a state-of-the-art trajectory prediction algorithm and showed that including various contextual information such as pedestrian intention and the ego-vehicle speed can positively impact the prediction of pedestrian trajectory.

The last chapter was dedicated to pedestrian crossing action prediction. For this, we examined the impact of various sources of contextual information and learning architectures on crossing prediction. We showed that a hierarchical architecture with multi-level contextual information fusion achieves the best performance for crossing action prediction. In addition, we evaluated the impact of changing observation properties and the order of fusing features within the network architecture on the accuracy of crossing prediction.

9.2 Study Limitations

The behavioral studies conducted as part of this dissertation were based on observations pedestrians in a naturalistic setting. Therefore, there is a possibility of some bias in judging pedestrian behavior and intentions. We tried to overcome this problem in later studies by increasing the number of people used for annotations.

Some of the behavioral factors such as pedestrians looking towards the traffic or making eye-contact were based on the judgments of the people involved in collecting the data.

The behavioral datasets collected as part of this work only contain images of pedestrians and lack the recordings of the drivers of the vehicles. Having the full recordings of both pedestrians and drivers could help to interpret some of the observed behaviors.

We collected the behavioral datasets using vehicles driven by humans. Using an actual

autonomous driving system can potentially impact the way pedestrians would behave.

Last but not least, in the chapters involving practical systems, we evaluated the performance of the algorithms in isolation without the presence of prerequisite methods for detecting and tracking pedestrians or any other aspects of an intelligent driving system. Although such an isolated evaluation strategy is necessary for understanding the limitations of the algorithms, in practice there are many sources of noise and interference from other modules that can potentially impact the performance of the prediction algorithms.

9.3 Future Work

Even though we covered various aspects of pedestrian behavior understanding and prediction, we barely scratched the surface of this problem. On the theoretical side, there is still a need for many behavioral studies of pedestrians. These studies should be conducted on a larger scope both in terms of the number of samples and the diversity of locations to capture effects of the environment and cultural norms. In addition, more studies involving autonomous vehicles are needed. Although classical traffic analyses involving pedestrian-driver interactions can shed light on understanding some of the fundamental aspects of pedestrian behavior, the ways pedestrians would behave can be quite different when facing autonomous vehicles.

To design better prediction algorithms we need to answer many questions. What are the best sources of information for prediction? What learning strategies are most suited (e.g. supervised vs reinforcement learning) for learning pedestrian behavior? What architectures should be used (e.g. feedforward vs recurrent networks) to make predictions? How the task(s) should be defined (e.g. should we be focusing on actions, trajectories or both) in the context of traffic prediction?

On the practical side of the problem, pedestrian behavior prediction should be examined in a driving system as a whole. In addition to social and environmental factors, there are other properties of the traffic scenes that can impact pedestrian behavior. For example, one should take into account the behavior of the ego-vehicle and other road users. In addition, noise present in various driving modules, such as perception and planning, should be considered when predicting pedestrian behavior.

Bibliography

[1] F. Kröger, "Automated driving in its social, historical and cultural contexts," in *Autonomous Driving*, 2016, pp. 41–68.

[2] T. Winkle, "Safety benefits of automated vehicles: Extended findings from accident research for development, validation and testing," in *Autonomous Driving*, 2016, pp. 335–364.

[3] T. Litman, "Autonomous vehicle implementation predictions," *Victoria Transport Policy Institute*, vol. 28, 2014.

[4] E. D. Dickmanns and A. Zapp, "A curvature-based scheme for improving road vehicle guidance by computer vision," in *Cambridge Symposium_Intelligent Robotics Systems*, 1987.

[5] M. Darms, P. E. Rybski, and C. Urmson, "A multisensor multiobject tracking system for an autonomous vehicle driving in an urban environment," in *International Symposium on Advanced Vehicle Control*, 2008.

[6] D. Radovanovic and D. Muoio, "This is what the evolution of self-driving cars looks like," Online, 2017-05-28. [Online]. Available: http://www.businessinsider.com/what-are-the-different-levels-of-driverless-cars-2016-10/#-1

[7] R. Bishop, "Intelligent vehicle applications worldwide," *IEEE Intelligent Systems and Their Applications*, vol. 15, no. 1, pp. 78–81, 2000.

[8] "Automated driving levels of driving automation are defined in new SAE international standard j3016," Online, 2017-05-28. [Online]. Available: https://www.sae.org/misc/pdfs/automated_driving.pdf

[9] "Unmanned ground vehicle," Online, 2017-05-28. [Online]. Available: http://www.wikiwand.com/en/Unmanned_ground_vehicle

[10] A. Oagana, "A short history of mercedes-benz autonomous driving technology," Online, 2017-05-28. [Online]. Available: https://www.autoevolution.com/news/a-short-history-of-mercedes-benz-autonomous-driving-technology-68148.html

[11] A. Broggi, M. Bertozzi, A. Fascioli, C. G. L. Bianco, and A. Piazzi, "The argo autonomous vehicles vision and control systems," *International Journal of Intelligent Control and Systems*, vol. 3, no. 4, pp. 409–441, 1999.

[12] "Self driving car," Online, 2017-05-28. [Online]. Available: http://stanford.edu/~cpiech/cs221/apps/driverlessCar.html

[13] "De 1977 nos jours, beaucoup de progrs!" Online, 2017-05-28. [Online]. Available: http://voitureautonome-2014.kazeo.com/de-1977-a-nos-jours-beaucoup-de-progres-a124503004

[14] "Vislab intercontinental autonomous challenge: Inaugural ceremony milan, italy," Online, 2017-05-28. [Online]. Available: http://manonthemove.com/2010/07/21/vislab-intercontinental-autonomous-challenge-inaugural-ceremony-milan-italy/

[15] "Watch Stanfords self-driving vehicle hit 120mph: Autonomous Audi proves to be just as good as a race car driver," Online, 2017-05-28. [Online]. Available: http://www.dailymail.co.uk/sciencetech/article-3472223/Watch-Stanford-s-self-driving-vehicle-hit-120mph-Autonomous-Audi-proves-just-good-race-car-driver.html

[16] A. Davieg, "We take a ride in the self-driving Uber now roaming Pittsburgh," Online, 2017-05-28. [Online]. Available: https://www.wired.com/2016/09/self-driving-autonomous-uber-pittsburgh/#slide-8

[17] "W. Grey Walters Tortoises Self-recognition and narcissism," Online, 2017-05-26. [Online]. Available: http://cyberneticzoo.com/cyberneticanimals/w-grey-walter-tortoises-picture-gallery-2/attachment/la-scienza-illustrata-1950_10-walter-tortoise-2-x640/

[18] "1960 Stanford Cart (American)," Online, 2017-05-26. [Online]. Available: http://cyberneticzoo.com/cyberneticanimals/1960-stanford-cart-american/

[19] "Elsie (electro-mechanical robot, light sensitive with internal and external stability," Online, 2017-05-26. [Online]. Available: http://cyberneticzoo.com/cyberneticanimals/elsie-cyberneticanimals/elsie/

[20] H. Moravec, "Obstacle avoidance and navigation in the real world by a seeing robot rover." DTIC Document, Tech. Rep., 1980.

[21] H. P. Moravec, "The Stanford Cart and the CMU rover," *Proceedings of the IEEE*, vol. 71, no. 7, pp. 872–884, 1983.

[22] E. Ackerman, "Self-driving cars were just around the cornerin 1960," Online, 2020-02-23. [Online]. Available: https://spectrum.ieee.org/tech-history/heroic-failures/selfdriving-cars-were-just-around-the-cornerin-1960

[23] R. Dang, "History of autonomous driving," Online, 2020-02-23. [Online]. Available: https://futurama.io/history-of-autonomous-driving/

[24] M. Mueller-Freitag, "Germany asleep at the wheel?" Online, 2017-05-28. [Online]. Available: https://medium.com/twentybn/germany-asleep-at-the-wheel-d800445d6da2

[25] S. Tsugawa, T. Yatabe, T. Hirose, and S. Matsumoto, "An automobile with artificial intelligence," in *International Joint Conference on Artificial Intelligence*, 1979.

[26] B. D. Mysliwetz and E. Dickmanns, "Distributed scene analysis for autonomous road vehicle guidance," in *Mobile Robot II*, 1987.

[27] E. D. Dickmanns, B. Mysliwetz, and T. Christians, "An integrated spatio-temporal approach to automatic visual guidance of autonomous vehicles," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 20, no. 6, pp. 1273–1284, 1990.

[28] D. A. Pomerleau, J. Gowdy, and C. E. Thorpe, "Combining artificial neural networks and symbolic processing for autonomous robot guidance," *Engineering Applications of Artificial Intelligence*, vol. 4, no. 4, pp. 279–285, 1991.

[29] D. Pomerleau, "Progress in neural network-based vision for autonomous robot driving," in *Intelligent Vehicles Symposium*, 1992.

[30] D. A. Pomerleau, "Neural network vision for robot driving," in *The Handbook of Brain Theory and Neural Networks*, 1996.

[31] S. Baluja, "Evolution of an artificial neural network based autonomous land vehicle controller," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 26, no. 3, pp. 450–463, 1996.

[32] C. Thorpe, M. Herbert, T. Kanade, and S. Shafer, "Toward autonomous driving: the CMU Navlab. i. perception," *IEEE Expert*, vol. 6, no. 4, pp. 31–42, 1991.

[33] T. M. Jochem, D. A. Pomerleau, and C. E. Thorpe, "Vision-based neural network road and intersection detection and traversal," in *International Conference on Intelligent Robots and Systems*, 1995.

[34] Y. U. Yim and S.-Y. Oh, "Three-feature based automatic lane detection algorithm (TFALDA) for autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 4, no. 4, pp. 219–225, 2003.

[35] K. Kluge, "Extracting road curvature and orientation from image edge points without perceptual grouping into features," in *Intelligent Vehicles Symposium*. IEEE, 1994, pp. 109–114.

[36] E. D. Dickmanns, R. Behringer, D. Dickmanns, T. Hildebrandt, M. Maurer, F. Thomanek, and J. Schiehlen, "The seeing passenger car 'VaMoRs-P'," in *Intelligent Vehicles Symposium*, 1994.

[37] U. Franke, S. Mehring, A. Suissa, and S. Hahn, "The Daimler-Benz steering assistant: a spin-off from autonomous driving," in *Intelligent Vehicles Symposium*, 1994.

[38] S. Bohrer, T. Zielke, and V. Freiburg, "An integrated obstacle detection framework for intelligent cruise control on motorways," in *Intelligent Vehicles Symposium*, 1995.

[39] T. Hong, M. Abrams, T. Chang, and M. Shneier, "An intelligent world model for autonomous off-road driving," *Computer Vision and Image Understanding*, vol. 80, no. 11, pp. 1–16, 2000.

[40] R. Behringer, S. Sundareswaran, B. Gregory, R. Elsley, B. Addison, W. Guthmiller, R. Daily, and D. Bevly, "The DARPA grand challenge-development of an autonomous vehicle," in *Intelligent Vehicles Symposium*, 2004.

[41] T. Dang, S. Kammel, C. Duchow, B. Hummel, and C. Stiller, "Path planning for autonomous driving based on stereoscopic and monoscopic vision cues," in *International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2006.

[42] S. Thrun, M. Montemerlo, H. Dahlkamp, D. Stavens, A. Aron, J. Diebel, P. Fong, J. Gale, M. Halpenny, G. Hoffmann *et al.*, "Stanley: The robot that won the DARPA grand challenge," *Journal of Field Robotics*, vol. 23, no. 9, pp. 661–692, 2006.

[43] S. Thrun, M. Montemerlo, and A. Aron, "Probabilistic terrain analysis for high-speed desert driving." in *Robotics Science and Systems*, 2006.

[44] G. M. Hoffmann, C. J. Tomlin, M. Montemerlo, and S. Thrun, "Autonomous automobile trajectory tracking for off-road driving: Controller design, experimental validation and racing," in *American Control Conference*, 2007.

[45] D. Dolgov, S. Thrun, M. Montemerlo, and J. Diebel, "Practical search techniques in path planning for autonomous driving," *AAAI Workshop*, 2008.

[46] C. Urmson, J. Anhalt, D. Bagnell, C. Baker, R. Bittner, M. Clark, J. Dolan, D. Duggins, T. Galatali, C. Geyer *et al.*, "Autonomous driving in urban environments: Boss and the urban challenge," *Journal of Field Robotics*, vol. 25, no. 8, pp. 425–466, 2008.

[47] D. Dolgov and S. Thrun, "Autonomous driving in semi-structured environments: Mapping and planning," in *International Conference on Robotics and Automation*, 2009.

[48] R. Kummerle, D. Hahnel, D. Dolgov, S. Thrun, and W. Burgard, "Autonomous driving in a multi-level parking structure," in *International Conference on Robotics and Automation*, 2009.

[49] J. Wei and J. M. Dolan, "A robust autonomous freeway driving algorithm," in *Intelligent Vehicles Symposium*, 2009.

[50] J. Wei, J. M. Snider, J. Kim, J. M. Dolan, R. Rajkumar, and B. Litkouhi, "Towards a viable autonomous driving research platform," in *Intelligent Vehicles Symposium*, 2013.

[51] S. Brechtel, T. Gindele, and R. Dillmann, "Probabilistic decision-making under uncertainty for autonomous driving using continuous pomdps," in *Intelligent Transportation Systems*, 2014.

[52] M. Bertozzi, L. Bombini, A. Broggi, M. Buzzoni, E. Cardarelli, S. Cattani, P. Cerri, S. Debattisti, R. Fedriga, M. Felisa *et al.*, "The VISLAB intercontinental autonomous challenge: 13,000 km, 3 months, no driver," in *World Congress on Intelligent Transport Systems*, 2010.

[53] C. Squatriglia, "Audi's robotic car climbs pikes peak," Online, 2017-05-28. [Online]. Available: https://www.wired.com/2010/11/audis-robotic-car-climbs-pikes-peak/

[54] M. Matousek, "Waymo is telling customers it will start offering rides in its autonomous cars without safety drivers," Online, 2019-02-02. [Online]. Available: https://www.businessinsider.com/waymo-says-it-will-start-giving-rides-without-safety-drivers-2019-10

[55] S. Millward, "Baidu's driverless cars on chinas roads by 2020," Online, 2017-05-30. [Online]. Available: https://www.techinasia.com/baidu-autonomous-car-sales-2020

[56] A. English, "Toyota's driveless car," Online, 2017-05-30. [Online]. Available: http://www.telegraph.co.uk/motoring/car-manufacturers/toyota/10404575/Toyotas-driverless-car.html

[57] "44 corporations working on autonomous vehicles," Online, 2017-05-30. [Online]. Available: https://www.cbinsights.com/blog/autonomous-driverless-vehicles-corporations-list/

[58] "Full self-driving hardware on all cars," Online, 2017-05-30. [Online]. Available: https://www.tesla.com/en_CA/autopilot?redirect=no

[59] G. Nica, "BMW CEO wants autonomous driving cars within five years," Online, 2017-05-28. [Online]. Available: http://www.bmwblog.com/2016/08/02/bmw-ceo-wants-autonomous-driving-cars-within-five-years/

[60] J. Ziegler, P. Bender, M. Schreiber, H. Lategahn, T. Strauss, C. Stiller, T. Dang, U. Franke, N. Appenrodt, C. G. Keller *et al.*, "Making Bertha drive-An autonomous journey on a historic route," *Intelligent Transportation Systems Magazine*, vol. 6, no. 2, pp. 8–20, 2014.

[61] "Waymo," Online, 2017-05-30. [Online]. Available: https://waymo.com/

[62] A. Davieg, "Ubers self-driving truck makes its first delivery: 50,000 beers," Online, 2017-05-30. [Online]. Available: https://www.wired.com/2016/10/ubers-self-driving-truck-makes-first-delivery-50000-beers/

[63] A. Marshal, "Dont look now, but even buses are going autonomous," Online, 2017-05-30. [Online]. Available: https://www.wired.com/2017/05/reno-nevada-autonomous-bus/

[64] O. Levander, "Forget autonomous cars-autonomous ships are almost here," Online, 2017-05-30. [Online]. Available: https://www.wired.com/2016/10/ubers-self-driving-truck-makes-first-delivery-50000-beers/

[65] F. Lambert, "Elon Musk clarifies teslas plan for level 5 fully autonomous driving: 2 years away from sleeping in the car," Online, 2017-05-30. [Online]. Available: https://electrek.co/2017/04/29/elon-musk-tesla-plan-level-5-full-autonomous-driving/

[66] E. Ackerman, "Toyota's Gill Pratt on self-driving cars and the reality of full autonomy," Online, 2017-05-30. [Online]. Available: http://spectrum.ieee.org/cars-that-think/transportation/self-driving/toyota-gill-pratt-on-the-reality-of-full-autonomy

[67] B. Friedrich, "The effect of autonomous vehicles on traffic," *Autonomous Driving*, pp. 317–334, 2016.

[68] T. M. Gasser, "Fundamental and special legal questions for autonomous vehicles," *Autonomous Driving*, pp. 523–551, 2016.

[69] D. Muoio, "6 scenarios self-driving cars still can't handle," Online, 2017-05-30. [Online]. Available: http://www.businessinsider.com/autonomous-car-limitations-2016-8/#1-driverless-cars-struggle-going-over-bridges-1

[70] R. Tussy, "The challenges facing autonomous vehicles," Online, 2017-05-30. [Online]. Available: http://auto-sens.com/the-challenges-facing-autonomous-vehicles/

[71] F. Lambert, "Tesla Model S driver crashes into a van while on autopilot [video]," Online, 2017-05-30. [Online]. Available: https://electrek.co/2016/05/26/tesla-model-s-crash-autopilot-video/

[72] "Tesla on autopilot hits police motorcycle," Online, 2017-05-30. [Online]. Available: http://www.government-fleet.com/channel/safety-accident-management/news/story/2017/03/tesla-on-autopilot-hits-police-motorcycle.aspx

[73] "Uber suspends self-driving fleet after Ariz. crash," Online, 2017-05-30. [Online]. Available: http://www.automotive-fleet.com/news/story/2017/03/uber-self-driving-car-struck-in-ariz-crash.aspx

[74] "Tesla driver dies in first fatal crash while using autopilot mode," Online, 2017-05-30. [Online]. Available: https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk

[75] "Another fatal tesla crash reportedly on autopilot emerges, Model S hits a streetsweeper truck caught on dashcam," Online, 2017-05-30. [Online]. Available: https://electrek.co/2016/09/14/another-fatal-tesla-autopilot-crash-emerges-model-s-hits-a-streetsweeper-truck-caught-on-dashcam/

[76] S. Levin and J. C. Wong, "Self-driving uber kills arizona woman in first fatal crash involving pedestrian," Online, 2020-02-08. [Online]. Available: https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe

[77] I. Wolf, "The interaction between humans and autonomous agents," in *Autonomous Driving*, 2016, pp. 103–124.

[78] B. Färber, "Communication and communication problems between autonomous vehicles and human drivers," in *Autonomous Driving*, 2016, pp. 125–144.

[79] N. McAlone, "Google's self-driving cars are really confused by 'hipster bicyclists' and their fixies," Online, 2020-02-08. [Online]. Available: https://www.businessinsider.com/google-self-driving-cars-get-confused-by-hipster-bicycles-2015-8

[80] A. Efrati, "Money pit: Self-driving cars $ 16 billion cash burn," Online, 2020-02-08. [Online]. Available: https://www.theinformation.com/articles/money-pit-self-driving-cars-16-billion-cash-burn

[81] M. Gough, "Machine smarts: how will pedestrians negotiate with driverless cars?" Online, 2017-05-30. [Online]. Available: https://www.theguardian.com/sustainable-business/2016/sep/09/machine-smarts-how-will-pedestrians-negotiate-with-driverless-cars

[82] M. Richtel, "Google's driverless cars run into problem: Cars with drivers," Online, 2017-05-30. [Online]. Available: https://www.nytimes.com/2015/09/02/technology/personaltech/google-says-its-not-the-driverless-cars-fault-its-other-drivers.html?_r=2

[83] S. E. Anthony, "The trollable self-driving car," Online, 2017-05-30. [Online]. Available: http://www.slate.com/articles/technology/future_tense/2016/03/google_self_driving_cars_lack_a_human_s_intuition_for_what_other_drivers.html

[84] M. McFarland, "Robots hit the streets – and the streets hit back," Online, 2017-05-30. [Online]. Available: http://money.cnn.com/2017/04/28/technology/robot-bullying/

[85] T. Lee, "New report highlights limitations of cruise self-driving cars," Online, 2020-02-08. [Online]. Available: https://arstechnica.com/cars/2018/03/new-report-highlights-limitations-of-cruise-self-driving-cars/

[86] S. Gibbs, "This article is more than 4 years old google self-driving car gets pulled over for driving too slowly," Online, 2020-02-08. [Online]. Available: https://www.theguardian.com/technology/2015/nov/13/google-self-driving-car-pulled-over-driving-too-slowly

[87] A. Efrati, "Waymos backseat drivers: Confidential data reveals self-driving taxi hurdles," Online, 2020-02-08. [Online]. Available: https://www.theinformation.com/articles/waymos-backseat-drivers-confidential-data-reveals-self-driving-taxi-hurdles

[88] J. Kahn, "Why not just retrain pedestrians to make self-driving vehicles safer?" Online, 2020-02-08. [Online]. Available: https://www.insurancejournal.com/news/national/2018/08/17/498414.htm

[89] R. Brooks, "Bothersome bystanders and self driving cars," Online, 2020-02-08. [Online]. Available: https://rodneybrooks.com/bothersome-bystanders-and-self-driving-cars/

[90] M. Kumashiro, H. Ishibashi, Y. Uchiyama, S. Itakura, A. Murata, and A. Iriki, "Natural imitation induced by joint attention in Japanese monkeys," *International Journal of Psychophysiology*, vol. 50, no. 1, pp. 81–99, 2003.

[91] M. Kidwell and D. H. Zimmerman, "Joint attention as action," *Journal of Pragmatics*, vol. 39, no. 3, pp. 592–611, 2007.

[92] G. Butterworth and E. Cochran, "Towards a mechanism of joint visual attention in human infancy," *International Journal of Behavioral Development*, vol. 3, no. 3, pp. 253–272, 1980.

[93] M. Scaife and J. S. Bruner, "The capacity for joint visual attention in the infant." *Nature*, 1975.

[94] M. Tomasello and J. Todd, "Joint attention and lexical acquisition style," *First Language*, vol. 4, no. 12, pp. 197–211, 1983.

[95] M. Botero, "Tactless scientists: Ignoring touch in the study of joint attention," *Philosophical Psychology*, vol. 29, no. 8, pp. 1200–1214, 2016.

[96] R. Sheldrake and A. Beeharee, "Is joint attention detectable at a distance? Three automated, internet-based tests," *Explore: The Journal of Science and Healing*, vol. 12, no. 1, pp. 34–41, 2016.

[97] C. Moore, M. Angelopoulos, and P. Bennett, "The role of movement in the development of joint visual attention," *Infant Behavior and Development*, vol. 20, no. 1, pp. 83–92, 1997.

[98] P. Mundy and M. Crowson, "Joint attention and early social communication: Implications for research on intervention with autism," *Journal of Autism and Developmental Disorders*, vol. 27, no. 6, pp. 653–676, 1997.

[99] W. V. Dube, R. P. MacDonald, R. C. Mansfield, W. L. Holcomb, and W. H. Ahearn, "Toward a behavioral analysis of joint attention," *The Behavior Analyst*, vol. 27, no. 2, p. 197, 2004.

[100] P. Holth, "An operant analysis of joint attention skills." *Journal of Early and Intensive Behavior Intervention*, vol. 2, no. 3, p. 160, 2005.

[101] M. Tomasello and M. Carpenter, "Shared intentionality," *Developmental Science*, vol. 10, no. 1, pp. 121–125, 2007.

[102] T. Charman, S. Baron-Cohen, J. Swettenham, G. Baird, A. Cox, and A. Drew, "Testing joint attention, imitation, and play as infancy precursors to language and theory of mind," *Cognitive Development*, vol. 15, no. 4, pp. 481–498, 2000.

[103] M. Carpenter, K. Nagell, M. Tomasello, G. Butterworth, and C. Moore, "Social cognition, joint attention, and communicative competence from 9 to 15 months of age," *Monographs of the Society for Research in Child Development*, pp. i–174, 1998.

[104] P. Mundy and A. Gomes, "Individual differences in joint attention skill development in the second year," *Infant Behavior and Development*, vol. 21, no. 3, pp. 469–482, 1998.

[105] R. MacDonald, J. Anderson, W. V. Dube, A. Geckeler, G. Green, W. Holcomb, R. Mansfield, and J. Sanchez, "Behavioral assessment of joint attention: A methodological report," *Research in Developmental Disabilities*, vol. 27, no. 2, pp. 138–150, 2006.

[106] T. Deroche, C. Castanier, A. Perrot, and A. Hartley, "Joint attention is slowed in older adults," *Experimental Aging Research*, vol. 42, no. 2, pp. 144–150, 2016.

[107] N. Sebanz, H. Bekkering, and G. Knoblich, "Joint action: Bodies and minds moving together," *Trends in Cognitive Sciences*, vol. 10, no. 2, pp. 70–76, 2006.

[108] E. Goffman *et al.*, *The presentation of self in everyday life.* Harmondsworth, 1978.

[109] P. D. Bardis, "Social interaction and social processes," *Social Science*, vol. 54, no. 3, pp. 147–167, 1979.

[110] A. Fiebich and S. Gallagher, "Joint attention in joint action," *Philosophical Psychology*, vol. 26, no. 4, pp. 571–587, 2013.

[111] P. Nuku and H. Bekkering, "Joint attention: Inferring what others perceive (and dont perceive)," *Consciousness and Cognition*, vol. 17, no. 1, pp. 339–349, 2008.

[112] M. Sucha, D. Dostal, and R. Risser, "Pedestrian-driver communication and decision strategies at marked crossings," *Accident Analysis & Prevention*, vol. 102, pp. 41–50, 2017.

[113] L. Fogassi, P. F. Ferrari, B. Gesierich, S. Rozzi, F. Chersi, and G. Rizzolatti, "Parietal lobe: from action organization to intention understanding," *Science*, vol. 308, no. 5722, pp. 662–667, 2005.

[114] M. A. Umilta, E. Kohler, V. Gallese, L. Fogassi, L. Fadiga, C. Keysers, and G. Rizzolatti, "I know what you are doing: A neurophysiological study," *Neuron*, vol. 31, no. 1, pp. 155–165, 2001.

[115] J. K. Tsotsos, "Motion understanding: Task-directed attention and representations that link perception with action," *International Journal of Computer Vision*, vol. 45, no. 3, pp. 265–280, 2001.

[116] K. Verfaillie and A. Daems, "Representing and anticipating human actions in vision," *Visual Cognition*, vol. 9, no. 1-2, pp. 217–232, 2002.

[117] J. R. Flanagan and R. S. Johansson, "Action plans used in action observation," *Nature*, vol. 424, no. 6950, p. 769, 2003.

[118] D. C. Dennett, *Brainstorms: Philosophical essays on mind and psychology.* MIT press, 1981.

[119] S. Baron-Cohen, *Mindblindness: An essay on autism and theory of mind. Cambridge, MA: Bradford.* MIT press, 1995.

[120] N. Humphrey, *Consciousness regained: Chapters in the development of mind.* Nicholas Humphrey, 1984.

[121] D. Sperber and D. Wilson, "Precis of relevance: Communication and cognition," *Behavioral and Brain Sciences*, vol. 10, no. 4, pp. 697–710, 1987.

[122] G. Wilde, "Immediate and delayed social interaction in road user behaviour," *Applied Psychology*, vol. 29, no. 4, pp. 439–460, 1980.

[123] J. M. Price and S. J. Glynn, "The relationship between crash rates and drivers' hazard assessments using the connecticut photolog," in *The HFES Annual Meeting*, vol. 44, no. 20, 2000, pp. 3–263.

[124] D. Crundall, "Driving experience and the acquisition of visual information," Ph.D. dissertation, University of Nottingham, 1999.

[125] S. Deb, L. Strawderman, D. W. Carruth, J. DuBien, B. Smith, and T. M. Garrison, "Development and validation of a questionnaire to assess pedestrian receptivity toward fully autonomous vehicles," *Transportation Research Part C: Emerging Technologies*, vol. 84, pp. 178–195, 2017.

[126] J. M. Sullivan and M. J. Flannagan, "Differences in geometry of pedestrian crashes in daylight and darkness," *Journal of Safety Research*, vol. 42, no. 1, pp. 33–37, 2011.

[127] R. Risser, "Behavior in traffic conflict situations," *Accident Analysis & Prevention*, vol. 17, no. 2, pp. 179–197, 1985.

[128] A. Tom and M.-A. Granié, "Gender differences in pedestrian rule compliance and visual search at signalized and unsignalized crossroads," *Accident Analysis & Prevention*, vol. 43, no. 5, pp. 1794–1801, 2011.

[129] M. Lefkowitz, R. R. Blake, and J. S. Mouton, "Status factors in pedestrian violation of traffic signals." *The Journal of Abnormal and Social Psychology*, vol. 51, no. 3, p. 704, 1955.

[130] S. Schmidt and B. Färber, "Pedestrians at the kerb–recognising the action intentions of humans," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 12, no. 4, pp. 300–310, 2009.

[131] D. Dey, M. Martens, B. Eggen, and J. Terken, "The impact of vehicle appearance and vehicle behavior on pedestrian interaction with autonomous vehicles," in *Automotive UI*, 2017, pp. 158–162.

[132] D. Clay, "Driver attitude and attribution: implications for accident prevention," Ph.D. dissertation, Cranfield University, 1995.

[133] D. R. Geruschat, S. E. Hassan, and K. A. Turano, "Gaze behavior while crossing complex intersections," *Optometry and Vision Science*, vol. 80, no. 7, pp. 515–528, 2003.

[134] M. Reed, "Intersection kinematics: a pilot study of driver turning behavior with application to pedestrian obscuration by a-pillars," University of Michigan, Tech. Rep., 2008.

[135] V. L. Neale, T. A. Dingus, S. G. Klauer, J. Sudweeks, and M. Goodman, "An overview of the 100-car naturalistic study and findings," *National Highway Traffic Safety Administration*, no. 05-0400, 2005.

[136] C. M. DiPietro and L. E. King, "Pedestrian gap-acceptance," *Highway Research Record*, no. 308, 1970.

[137] N. W. Heimstra, J. Nichols, and G. Martin, "An experimental methodology for analysis of child pedestrian behavior," *Pediatrics*, vol. 44, no. 5, pp. 832–838, 1969.

[138] R. Eenink, Y. Barnard, M. Baumann, X. Augros, and F. Utesch, "UDRIVE: The European naturalistic driving study," in *Transport Research Arena*, 2014.

[139] D. Sun, S. Ukkusuri, R. F. Benekohal, and S. T. Waller, "Modeling of motorist-pedestrian interaction at uncontrolled mid-block crosswalks," *Urbana*, vol. 51, p. 61801, 2002.

[140] T. Wang, J. Wu, P. Zheng, and M. McDonald, "Study of pedestrians' gap acceptance behavior when they jaywalk outside crossing facilities," in *Intelligent Transportation Systems Conference*, 2010, pp. 1295–1300.

[141] T. Lagstrom and V. M. Lundgren, "AVIP-autonomous vehicles interaction with pedestrians," Master's thesis, Chalmers University of Technology, Gothenborg, Sweden, 2015.

[142] B. Herwig, "Verhalten von kraftfahrern und fussgangern an zebrastreifen," *Zeitschrift für Verkehrssicherheit*, vol. 11, pp. 189–202, 1965.

[143] P. Schioldborg, "Children, traffic and traffic training: analysis of the childrens traffic club," *The Voice of the Pedestrian*, vol. 6, pp. 12–19, 1976.

[144] W. A. Harrell, "Factors influencing pedestrian cautiousness in crossing streets," *The Journal of Social Psychology*, vol. 131, no. 3, pp. 367–372, 1991.

[145] T. Rosenbloom, "Crossing at a red light: Behaviour of individuals and groups," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 12, no. 5, pp. 389–394, 2009.

[146] R. Wiedemann, *Simulation des Straßenverkehrsflusses*. Institute for Transportation Science, University of Karlsruhe, Germany, 1974.

[147] D. Evans and P. Norman, "Understanding pedestrians' road crossing decisions: an application of the theory of planned behaviour," *Health Education Research*, vol. 13, no. 4, pp. 481–489, 1998.

[148] D. Johnston, "Road accident casuality: A critique of the literature and an illustrative case," *Ontario: Grand Rounds. Department of Psychiatry, Hotel Dieu Hospital*, 1973.

[149] M. Gheri, "Über das blickverhalten von kraftfahrern an kreuzungen," *Kuratorium für Verkehrssicherheit, Kleine Fachbuchreihe Bd*, vol. 5, 1963.

[150] M. Šucha, "Road users strategies and communication: driver-pedestrian interaction," *Transport Research Arena*, 2014.

[151] D. Yagil, "Beliefs, motives and situational factors related to pedestrians self-reported behavior at signal-controlled crossings," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 3, no. 1, pp. 1–13, 2000.

[152] J. Dolphin, L. Kennedy, S. O'Donnell, and G. Wilde, "Factors influencing pedestrian violations," *Queens University, Kingston, Ontario*, 1970.

[153] C. Holland and R. Hill, "The effect of age, gender and driver status on pedestrians intentions to cross the road in risky situations," *Accident Analysis & Prevention*, vol. 39, no. 2, pp. 224–237, 2007.

[154] R. L. Moore, "Pedestrian choice and judgment," *Journal of the Operational Research Society*, vol. 4, no. 1, pp. 3–10, 1953.

[155] G. Jacobs and D. G. Wilson, "A study of pedestrian risk in crossing busy roads in four towns," *Road Research Laboratory Reports*, 1967.

[156] M. M. Ishaque and R. B. Noland, "Behavioural issues in pedestrian speed choice and street crossing behaviour: A review," *Transport Reviews*, vol. 28, no. 1, pp. 61–85, 2008.

[157] 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 *Intelligent Transportation Systems Conference*, 2014, pp. 1758–1763.

[158] W. A. Harrell, "Precautionary street crossing by elderly pedestrians," *The International Journal of Aging and Human Development*, vol. 32, no. 1, pp. 65–80, 1991.

[159] R. R. Oudejans, C. F. Michaels, B. van Dort, and E. J. Frissen, "To cross or not to cross: The effect of locomotion on street-crossing behavior," *Ecological Psychology*, vol. 8, no. 3, pp. 259–267, 1996.

[160] R. Tian, E. Y. Du, K. Yang, P. Jiang, F. Jiang, Y. Chen, R. Sherony, and H. Takahashi, "Pilot study on pedestrian step frequency in naturalistic driving environment," in *Intelligent Vehicles Symposium*, 2013, pp. 1215–1220.

[161] D. Crompton, "Pedestrian delay, annoyance and risk: preliminary results from a 2 years study," in *PTRC Summer Annual Meeting*, 1979, pp. 275–299.

[162] C. OFlaherty and M. Parkinson, "Movement on a city centre footway," *Traffic Engineering and Control*, vol. 13, no. 10, pp. 434–438, 1972.

[163] A. Willis, N. Gjersoe, C. Havard, J. Kerridge, and R. Kukla, "Human movement behaviour in urban spaces: Implications for the design and modelling of effective pedestrian environments," *Environment and Planning B: Planning and Design*, vol. 31, no. 6, pp. 805–828, 2004.

[164] L. Sjöstedt, "Behaviour of pedestrians at pedestrian crossings," *Human Factors in Traffic Safety Research*, 1969.

[165] G. Underwood, P. Chapman, N. Brocklehurst, J. Underwood, and D. Crundall, "Visual attention while driving: Sequences of eye fixations made by experienced and novice drivers," *Ergonomics*, vol. 46, no. 6, pp. 629–646, 2003.

[166] S. G. Klauer, V. L. Neale, T. A. Dingus, D. Ramsey, and J. Sudweeks, "Driver inattention: A contributing factor to crashes and near-crashes," in *The HFES Annual Meeting*, vol. 49, no. 22, 2005, pp. 1922–1926.

[167] G. Underwood, "Visual attention and the transition from novice to advanced driver," *Ergonomics*, vol. 50, no. 8, pp. 1235–1249, 2007.

[168] Y. Barnard, F. Utesch, N. Nes, R. Eenink, and M. Baumann, "The study design of UDRIVE: The naturalistic driving study across Europe for cars, trucks and scooters," *European Transport Research Review*, vol. 8, no. 2, pp. 1–10, 2016.

[169] Z. Ren, X. Jiang, and W. Wang, "Analysis of the influence of pedestrians eye contact on drivers comfort boundary during the crossing conflict," *Procedia Engineering*, vol. 137, pp. 399–406, 2016.

[170] I. E. Hyman, S. M. Boss, B. M. Wise, K. E. McKenzie, and J. M. Caggiano, "Did you see the unicycling clown? Inattentional blindness while walking and talking on a cell phone," *Applied Cognitive Psychology*, vol. 24, no. 5, pp. 597–607, 2010.

[171] A. Lindgren, F. Chen, P. W. Jordan, and H. Zhang, "Requirements for the design of advanced driver assistance systems-the differences between Swedish and Chinese drivers," *International Journal of Design*, vol. 2, no. 2, 2008.

[172] G. M. Björklund and L. Åberg, "Driver behaviour in intersections: Formal and informal traffic rules," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 8, no. 3, pp. 239–253, 2005.

[173] T. Rosenbloom, H. Barkan, and D. Nemrodov, "For heavens sake keep the rules: Pedestrians behavior at intersections in ultra-orthodox and secular cities," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 7, pp. 395–404, 2004.

[174] V. P. Sisiopiku and D. Akin, "Pedestrian behaviors at and perceptions towards various pedestrian facilities: an examination based on observation and survey data," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 6, no. 4, pp. 249–274, 2003.

[175] R. Sun, X. Zhuang, C. Wu, G. Zhao, and K. Zhang, "The estimation of vehicle speed and stopping distance by pedestrians crossing streets in a naturalistic traffic environment," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 30, pp. 97–106, 2015.

[176] R. Mortimer, "Behavioral evaluation of pedestrian signals," *Traffic Engineering*, vol. 44, no. 2, pp. 22–26, 1973.

[177] W. H. Lam, J. F. Morrall, and H. Ho, "Pedestrian flow characteristics in hong kong," *Transportation Research Record*, no. 1487, 1995.

[178] B. C. de Lavalette, C. Tijus, S. Poitrenaud, C. Leproux, J. Bergeron, and J.-P. Thouez, "Pedestrian crossing decision-making: A situational and behavioral approach," *Safety Science*, vol. 47, no. 9, pp. 1248–1253, 2009.

[179] X. Chu, M. Guttenplan, and M. Baltes, "Why people cross where they do: the role of street environment," *Transportation Research Record: Journal of the Transportation Research Board*, no. 1878, pp. 3–10, 2004.

[180] P.-S. Lin, Z. Wang, and R. Guo, "Impact of connected vehicles and autonomous vehicles on future transportation," *Bridging the East and West*, pp. 46–53, 2016.

[181] E. CYingzi Du, K. Yang, F. Jiang, P. Jiang, R. Tian, M. Luzetski, Y. Chen, R. Sherony, and H. Takahashi, "Pedestrian behavior analysis using 110-car naturalistic driving data in USA," Online, 2017-06-3. [Online]. Available: https://www-nrd.nhtsa.dot.gov/pdf/Esv/esv23/23ESV-000291.pdf

[182] S. Das, C. F. Manski, and M. D. Manuszak, "Walk or wait? an empirical analysis of street crossing decisions," *Journal of Applied Econometrics*, vol. 20, no. 4, pp. 529–548, 2005.

[183] W. A. Harrell and T. Bereska, "Gap acceptance by pedestrians," *Perceptual and Motor Skills*, vol. 75, no. 2, pp. 432–434, 1992.

[184] J. Cohen, E. Dearnaley, and C. Hansel, "The risk taken in crossing a road," *Journal of the Operational Research Society*, vol. 6, no. 3, pp. 120–128, 1955.

[185] M. M. Hamed, "Analysis of pedestrians behavior at pedestrian crossings," *Safety Science*, vol. 38, no. 1, pp. 63–82, 2001.

[186] A. Varhelyi, "Drivers' speed behaviour at a zebra crossing: a case study," *Accident Analysis & Prevention*, vol. 30, no. 6, pp. 731–743, 1998.

[187] D. Dey and J. Terken, "Pedestrian interaction with vehicles: roles of explicit and implicit communication," in *Automotive UI*, 2017, pp. 109–113.

[188] I. Walker, "Drivers overtaking bicyclists: Objective data on the effects of riding position, helmet use, vehicle type and apparent gender," *Accident Analysis & Prevention*, vol. 39, no. 2, pp. 417–425, 2007.

[189] N. Guéguen, S. Meineri, and C. Eyssartier, "A pedestrians stare and drivers stopping behavior: A field experiment at the pedestrian crossing," *Safety Science*, vol. 75, pp. 87–89, 2015.

[190] S. Gupta, M. Vasardani, and S. Winter, "Conventionalized gestures for the interaction of people in traffic with autonomous vehicles," in *International Workshop on Computational Transportation Science*, 2016, pp. 55–60.

[191] J. Caird and P. Hancock, "The perception of arrival time for different oncoming vehicles at an intersection," *Ecological Psychology*, vol. 6, no. 2, pp. 83–109, 1994.

[192] M. Matthews, G. Chowdhary, and E. Kieson, "Intent communication between autonomous vehicles and pedestrians," *arXiv:1708.07123*, 2017.

[193] R. Zimmermann and R. Wettach, "First step into visceral interaction with autonomous vehicles," in *Automotive UI*, 2017, pp. 58–64.

[194] S. Yang, "Driver behavior impact on pedestrians' crossing experience in the conditionally autonomous driving context," Master's thesis, KTH Royal Institute of Technology, 2017.

[195] M. Clamann, M. Aubert, and M. L. Cummings, "Evaluation of vehicle-to-pedestrian communication displays for autonomous vehicles," in *Transportation Research Board 96th Annual Meeting*, 2017.

[196] A. Pillai, "Virtual reality based study to analyse pedestrian attitude towards autonomous vehicles," Master's thesis, KTH Royal Institute of Technology, 2017.

[197] C.-M. Chang, K. Toda, D. Sakamoto, and T. Igarashi, "Eyes on a car: An interface design for communication between an autonomous car and a pedestrian," in *Automotive UI*, 2017, pp. 65–73.

[198] K. Mahadevan, S. Somanath, and E. Sharlin, "Communicating awareness and intent in autonomous vehicle-pedestrian interaction," University of Calgary, Tech. Rep., 2017.

[199] M. Beggiato, C. Witzlack, S. Springer, and J. Krems, "The right moment for braking as informal communication signal between automated vehicles and pedestrians in crossing situations," in *International Conference on Applied Human Factors and Ergonomics*, 2017, pp. 1072–1081.

[200] L. M. Hulse, H. Xie, and E. R. Galea, "Perceptions of autonomous vehicles: Relationships with road users, risk, gender and age," *Safety Science*, vol. 102, pp. 1–13, 2018.

[201] S. Jayaraman, C. Creech, L. Robert, D. Tilbury, J. Yang, A. Pradhan, K. Tsui *et al.*, "Trust in av: An uncertainty reduction model of av-pedestrian interactions," in *Human Robot Interaction*, 2018.

[202] S. Deb, M. M. Rahman, L. J. Strawderman, and T. M. Garrison, "Pedestrians receptivity toward fully automated vehicles: Research review and roadmap for future research," *IEEE Transaction on Human-Machine Systems*, vol. 48, no. 3, pp. 279–290, 2018.

[203] M. Risto, C. Emmenegger, E. Vinkhuyzen, M. Cefkin, and J. Hollan, "Human-vehicle interfaces: The power of vehicle movement gestures in human road user coordination," in *Human Factors in Driver Assessment, Training, and Vehicle Design*, 2017, pp. 186–192.

[204] A. Millard-Ball, "Pedestrians, autonomous vehicles, and cities," *Journal of Planning Education and Research*, pp. 6–12, 2016.

[205] D. Rothenbücher, J. Li, D. Sirkin, B. Mok, and W. Ju, "Ghost driver: A field study investigating the interaction between pedestrians and driverless vehicles," in *International Symposium on Robot and Human Interactive Communication*, 2016, pp. 795–802.

[206] L. Müller, M. Risto, and C. Emmenegger, "The social behavior of autonomous vehicles," in *International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, 2016, pp. 686–689.

[207] M. Meeder, E. Bosina, and U. Weidmann, "Autonomous vehicles: Pedestrian heaven or pedes-trian hell?" in *Swiss Transport Research Conference*, 2017.

[208] H. Prakken, "On the problem of making autonomous vehicles conform to traffic law," *Artificial Intelligence and Law*, vol. 25, no. 3, pp. 341–363, 2017.

[209] J. Wang, J. Lu, F. You, and Y. Wang, "Act like a human: Teach an autonomous vehicle to deal with traffic encounters," in *Intelligent Human Systems Integration*, 2018, pp. 537–542.

[210] O. Juhlin, "Traffic behaviour as social interaction-implications for the design of artificial drivers," in *World Congress on Intelligent Transport Systems*, 1999.

[211] Bikeleauge, "Autonomous and connected vehicles: Implications for bicyclists and pedestrians," Online, 2014, 2017-06-3. [Online]. Available: http://bikeleague.org/sites/default/files/Bike_Ped_Connected_Vehicles.pdf

[212] N. J. Briton and J. A. Hall, "Beliefs about female and male nonverbal communication," *Sex Roles*, vol. 32, no. 1, pp. 79–90, 1995.

[213] M. A. Hecht and N. Ambady, "Nonverbal communication and psychology: Past and future," *Atlantic Journal of Communication*, vol. 7, no. 2, pp. 156–170, 1999.

[214] R. Buck and C. A. VanLear, "Verbal and nonverbal communication: Distinguishing symbolic, spontaneous, and pseudo-spontaneous nonverbal behavior," *Journal of Communication*, vol. 52, no. 3, pp. 522–541, 2002.

[215] C. Darwin, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[216] R. M. Krauss, Y. Chen, and P. Chawla, "Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us?" *Advances in Experimental Social Psychology*, vol. 28, pp. 389–450, 1996.

[217] A. Mehrabian, *Public places and private spaces: the psychology of work, play, and living environments*. Basic Books New York, 1976.

[218] R. L. Birdwhistell, *Kinesics and context: Essays on body motion communication*. University of Pennsylvania press, 2010.

[219] M. R. DiMatteo, A. Taranta, H. S. Friedman, and L. M. Prince, "Predicting patient satisfaction from physicians' nonverbal communication skills," *Medical Care*, pp. 376–387, 1980.

[220] S. Nowicki and M. P. Duke, "Individual differences in the nonverbal communication of affect: The diagnostic analysis of nonverbal accuracy scale," *Journal of Nonverbal Behavior*, vol. 18, no. 1, pp. 9–35, 1994.

[221] M. Argyle and J. Dean, "Eye-contact, distance and affiliation," *Sociometry*, pp. 289–304, 1965.

[222] A. Senju and M. H. Johnson, "The eye contact effect: mechanisms and development," *Trends in Cognitive Sciences*, vol. 13, no. 3, pp. 127–134, 2009.

[223] A. E. Scheflen, "The significance of posture in communication systems," *Psychiatry*, vol. 27, no. 4, pp. 316–331, 1964.

[224] P. Dollár, "Piotr's Computer Vision Matlab Toolbox (PMT)," https://github.com/pdollar/toolbox.

[225] O. Friard and M. Gamba, "BORIS: a free, versatile open-source event-logging software for video/audio coding and live observations," *Methods in Ecology and Evolution*, vol. 7, no. 11, pp. 1325–1330, 2016.

[226] Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian detection aided by deep learning semantic tasks," in *Conference on Computer Vision and Pattern Recognition*, 2015.

[227] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532–1545, 2014.

[228] S. Zhang, R. Benenson, and B. Schiele, "Filtered channel features for pedestrian detection," in *Conference on Computer Vision and Pattern Recognition*, 2015.

[229] X. Du, M. El-Khamy, J. Lee, and L. Davis, "Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection," in *Winter Conference on Applications of Computer Vision*, 2017.

[230] G. Brazil, X. Yin, and X. Liu, "Illuminating pedestrians via simultaneous detection & segmentation," in *International Conference on Computer Vision*, 2017.

[231] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Conference on Computer Vision and Pattern Recognition*, 2009.

[232] 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.

[233] E. Ohn-Bar and M. M. Trivedi, "To boost or not to boost? On the limits of boosted trees for object detection," in *International Conference on Pattern Recognition*, 2016.

[234] L. Zhang, L. Lin, X. Liang, and K. He, "Is Faster R-CNN doing well for pedestrian detection?" in *European Conference on Computer Vision*, 2016.

[235] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Convolutional channel features," in *International Conference on Pattern Recognition*, 2015.

[236] W. Ouyang and X. Wang, "Joint deep learning for pedestrian detection," in *International Conference on Pattern Recognition*, 2013.

[237] Y. Tian, P. Luo, X. Wang, and X. Tang, "Deep learning strong parts for pedestrian detection," in *International Conference on Pattern Recognition*, 2015.

[238] J. Noh, S. Lee, B. Kim, and G. Kim, "Improving occlusion and hard negative handling for single-stage pedestrian detectors," in *Conference on Computer Vision and Pattern Recognition*, 2018.

[239] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in cnns," in *Conference on Computer Vision and Pattern Recognition*, 2018.

[240] C. Zhou and J. Yuan, "Bi-box regression for pedestrian detection and occlusion estimation," in *European Conference on Computer Vision*, 2018.

[241] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Occlusion-aware r-cnn: Detecting pedestrians in a crowd," in *European Conference on Computer Vision*, 2018.

[242] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion loss: Detecting pedestrians in a crowd," in *Conference on Computer Vision and Pattern Recognition*, 2018.

[243] Y. Pang, J. Xie, M. H. Khan, R. M. Anwer, F. S. Khan, and L. Shao, "Mask-guided attention network for occluded pedestrian detection," in *International Conference on Computer Vision*, 2019.

[244] C. Zhou, M. Yang, and J. Yuan, "Discriminative feature transformation for occluded pedestrian detection," in *International Conference on Computer Vision*, 2019.

[245] T. Song, L. Sun, D. Xie, H. Sun, and S. Pu, "Small-scale pedestrian detection based on topological line localization and temporal feature aggregation," in *European Conference on Computer Vision*, 2018.

[246] S. Liu, D. Huang, and Y. Wang, "Adaptive nms: Refining pedestrian detection in a crowd," in *Conference on Computer Vision and Pattern Recognition*, 2019.

[247] C. Lin, J. Lu, G. Wang, and J. Zhou, "Graininess-aware deep feature learning for pedestrian detection," in *European Conference on Computer Vision*, 2018.

[248] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-level semantic feature detection: A new perspective for pedestrian detection," in *Conference on Computer Vision and Pattern Recognition*, 2019.

[249] G. Brazil and X. Liu, "Pedestrian detection with autoregressive network phases," in *Conference on Computer Vision and Pattern Recognition*, 2019.

[250] W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen, "Learning efficient single-stage pedestrian detectors by asymptotic localization fitting," in *European Conference on Computer Vision*, 2018.

[251] S. Wu, S. Lin, W. Wu, M. Azzam, and H.-S. Wong, "Semi-supervised pedestrian instance synthesis and detection with mutual reinforcement," in *International Conference on Computer Vision*, 2019.

[252] S. Zhang, R. Benenson, and B. Schiele, "CityPersons: A diverse dataset for pedestrian detection," in *Conference on Computer Vision and Pattern Recognition*, 2017.

[253] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," *arXiv:1912.04838*, 2019.

[254] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," *arXiv:1903.11027*, 2019.

[255] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, and J. Hays, "Argoverse: 3d tracking and forecasting with rich maps," in *Conference on Computer Vision and Pattern Recognition*, 2019.

[256] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, "Pedestrian detection using wavelet templates," in *Conference on Computer Vision and Pattern Recognition*, 1997.

[257] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Conference on Computer Vision and Pattern Recognition*, 2005.

[258] S. Munder and D. M. Gavrila, "An experimental study on pedestrian classification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 11, pp. 1863–1868, 2006.

[259] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2179–2195, 2009.

[260] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila, "Multi-cue pedestrian classification with partial occlusion handling," in *Conference on Computer Vision and Pattern Recognition*.

[261] C. G. Keller, M. Enzweiler, and D. M. Gavrila, "A new benchmark for stereo-based pedestrian detection," in *Intelligent Vehicles Symposium*, 2011.

[262] L. Wang, J. Shi, G. Song, and I.-F. Shen, "Object detection combining recognition and segmentation," in *Asian Conference on Computer Vision*, 2007.

[263] A. Ess, B. Leibe, K. Schindler, and L. Van Gool, "A mobile vision system for robust multi-person tracking," in *Conference on Computer Vision and Pattern Recognition*, 2008.

[264] C. Wojek, S. Walk, and B. Schiele, "Multi-cue onboard pedestrian detection," in *Conference on Computer Vision and Pattern Recognition*, 2009.

[265] K. Fragkiadaki, W. Zhang, G. Zhang, and J. Shi, "Two-granularity tracking: Mediating trajectory and detection graphs for tracking under occlusions," in *European Conference on Computer Vision*, 2012.

[266] W. Ouyang and X. Wang, "A discriminative deep model for pedestrian detection with occlusion handling," in *Conference on Computer Vision and Pattern Recognition*, 2012.

[267] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *International Conference on Multimedia*, 2014.

[268] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Conference on Computer Vision and Pattern Recognition*, 2015.

[269] M. Braun, S. Krebs, F. B. Flohr, and D. M. Gavrila, "Eurocity persons: A novel benchmark for person detection in traffic scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1844–1861, 2019.

[270] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *International Conference on Pattern Recognition*, 2017.

[271] R. N. Rajaram, E. Ohn-Bar, and M. M. Trivedi, "An exploration of why and when pedestrian detection fails," in *International Conference on Intelligent Transportation Systems*, 2015.

[272] S.-I. Jung and K.-S. Hong, "Deep network aided by guiding network for pedestrian detection," *Pattern Recognition Letters*, vol. 90, pp. 43–49, 2017.

[273] M. Taiana, J. Nascimento, and A. Bernardino, "On the purity of training and testing data for learning: The case of pedestrian detection," *Neurocomputing*, vol. 150, pp. 214–226, 2015.

[274] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?" in *Conference on Computer Vision and Pattern Recognition*, 2016.

[275] D. Li, Z. Zhang, X. Chen, H. Ling, and K. Huang, "A richly annotated dataset for pedestrian attribute recognition," *arXiv:1603.07054*, 2016.

[276] A. Rangesh, E. Ohn-Bar, K. Yuen, and M. M. Trivedi, "Pedestrians and their phones-Detecting phone-based activities of pedestrians for autonomous vehicles," in *International Conference on Intelligent Transportation Systems*, 2016.

[277] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *European Conference on Computer Vision*, 2016.

[278] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila, "Context-based pedestrian path prediction," in *European Conference on Computer Vision*, 2014.

[279] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human trajectory prediction in crowded spaces," in *Conference on Computer Vision and Pattern Recognition*, 2016.

[280] A. Bhattacharyya, M. Fritz, and B. Schiele, "Long-term on-board prediction of people in traffic scenes under uncertainty," in *Conference on Computer Vision and Pattern Recognition*, 2018.

[281] F. Schneemann and P. Heinemann, "Context-based detection of pedestrian crossing intention for autonomous driving in urban environments," in *International Conference on Intelligent Robots and Systems*, 2016.

[282] J. F. Carvalho, M. Vejdemo-Johansson, F. T. Pokorny, and D. Kragic, "Long-term prediction of motion trajectories using path homology clusters," in *International Conference on Intelligent Robots and Systems*, 2019.

[283] Y. Yoo, K. Yun, S. Yun, J. Hong, H. Jeong, and J. Young Choi, "Visual path prediction in complex scenes with crowded moving objects," in *Conference on Computer Vision and Pattern Recognition*, 2016.

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

[285] Chenghui Zhou, B. Balle, and J. Pineau, "Learning time series models for pedestrian motion prediction," in *International Conference on Robotics and Automation*, 2016.

[286] A. Rudenko, L. Palmieri, and K. O. Arras, "Joint long-term prediction of human motion using a planning-based social force approach," in *International Conference on Robotics and Automation*, 2018.

[287] A. Rudenko, L. Palmieri, A. J. Lilienthal, and K. O. Arras, "Human motion prediction under social grouping constraints," in *International Conference on Intelligent Robots and Systems*, 2018.

[288] J. Schulz, C. Hubmann, J. Lchner, and D. Burschka, "Interaction-aware probabilistic behavior prediction in urban environments," in *International Conference on Intelligent Robots and Systems*, 2018.

[289] M. Shen, G. Habibi, and J. P. How, "Transferable pedestrian motion prediction models at intersections," in *International Conference on Intelligent Robots and Systems*, 2018.

[290] F. Shkurti and G. Dudek, "Topologically distinct trajectory predictions for probabilistic pursuit," in *International Conference on Intelligent Robots and Systems*, 2017.

[291] D. Vasquez, "Novel planning-based algorithms for human motion prediction," in *International Conference on Robotics and Automation*, 2016.

[292] V. Karasev, A. Ayvaci, B. Heisele, and S. Soatto, "Intent-aware long-term prediction of pedestrian motion," in *International Conference on Robotics and Automation*, 2016.

[293] N. Lee and K. M. Kitani, "Predicting wide receiver trajectories in american football," in *Winter Conference on Applications of Computer Vision*, 2016.

[294] H. Bai, S. Cai, N. Ye, D. Hsu, and W. S. Lee, "Intention-aware online pomdp planning for autonomous driving in a crowd," in *International Conference on Robotics and Automation*, 2015.

[295] S. Solaimanpour and P. Doshi, "A layered hmm for predicting motion of a leader in multi-robot settings," in *International Conference on Robotics and Automation*, 2017.

[296] Y. F. Chen, M. Liu, and J. P. How, "Augmented dictionary learning for motion prediction," in *International Conference on Robotics and Automation*, 2016.

[297] R. Sanchez-Matilla and A. Cavallaro, "A predictor of moving objects for first-person vision," in *International Conference on Image Processing*, 2019.

[298] B. Lee, J. Choi, C. Baek, and B. Zhang, "Robust human following by deep bayesian trajectory prediction for home service robots," in *International Conference on Robotics and Automation*, 2018.

[299] I. Hasan, F. Setti, T. Tsesmelis, A. Del Bue, M. Cristani, and F. Galasso, ""seeing is believing": Pedestrian trajectory forecasting using visual frustum of attention," in *Winter Conference on Applications of Computer Vision*, 2018.

[300] L. Ballan, F. Castaldo, A. Alahi, F. Palmieri, and S. Savarese, "Knowledge transfer for scene-specific motion prediction," in *European Conference on Computer Vision*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., 2016, pp. 697–713.

[301] M. Pfeiffer, U. Schwesinger, H. Sommer, E. Galceran, and R. Siegwart, "Predicting actions to act predictably: Cooperative partial motion planning with maximum entropy models," in *International Conference on Intelligent Robots and Systems*, 2016.

[302] N. N. Vo and A. F. Bobick, "Augmenting physical state prediction through structured activity inference," in *International Conference on Robotics and Automation*, 2015.

[303] V. Akbarzadeh, C. Gagn, and M. Parizeau, "Kernel density estimation for target trajectory prediction," in *International Conference on Intelligent Robots and Systems*, 2015.

[304] A. Schulz and R. Stiefelhagen, "A controlled interactive multiple model filter for combined pedestrian intention recognition and path prediction," in *International Conference on Intelligent Transportation Systems*, 2015.

[305] J. Liang, L. Jiang, J. C. Niebles, A. G. Hauptmann, and L. Fei-Fei, "Peeking into the future: Predicting future person activities and locations in videos," in *Conference on Computer Vision and Pattern Recognition*, 2019.

[306] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese, "Sophie: An attentive gan for predicting paths compliant to social and physical constraints," in *Conference on Computer Vision and Pattern Recognition*, 2019.

[307] P. Zhang, W. Ouyang, P. Zhang, J. Xue, and N. Zheng, "Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction," in *Conference on Computer Vision and Pattern Recognition*, 2019.

[308] T. Zhao, Y. Xu, M. Monfort, W. Choi, C. Baker, Y. Zhao, Y. Wang, and Y. N. Wu, "Multi-agent tensor fusion for contextual trajectory prediction," in *Conference on Computer Vision and Pattern Recognition*, 2019.

[309] H. Bi, Z. Fang, T. Mao, Z. Wang, and Z. Deng, "Joint prediction for kinematic trajectories in vehicle-pedestrian-mixed scenes," in *International Conference on Computer Vision*, 2019.

[310] C. Choi and B. Dariush, "Looking to relations for future trajectory forecast," in *International Conference on Computer Vision*, 2019.

[311] Y. Huang, H. Bi, Z. Li, T. Mao, and Z. Wang, "Stgat: Modeling spatial-temporal interactions for human trajectory prediction," in *International Conference on Computer Vision*, 2019.

[312] L. A. Thiede and P. P. Brahma, "Analyzing the variety loss in the context of probabilistic trajectory prediction," in *International Conference on Computer Vision*, 2019.

[313] V. Kosaraju, A. Sadeghian, R. Martín-Martín, I. Reid, H. Rezatofighi, and S. Savarese, "Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks," in *Advances in Neural Information Processing Systems*, 2019.

[314] W. Ding and S. Shen, "Online vehicle trajectory prediction using policy anticipation network and optimization-based context reasoning," in *International Conference on Robotics and Automation*, 2019.

[315] J. Li, H. Ma, and M. Tomizuka, "Interaction-aware multi-agent tracking and probabilistic behavior prediction via adversarial learning," in *International Conference on Robotics and Automation*, 2019.

[316] C. Anderson, X. Du, R. Vasudevan, and M. Johnson-Roberson, "Stochastic sampling simulation for pedestrian trajectory prediction," in *International Conference on Intelligent Robots and Systems*, 2019.

[317] S. Srikanth, J. A. Ansari, S. Sharma *et al.*, "Infer: Intermediate representations for future prediction," in *International Conference on Intelligent Robots and Systems*, 2019.

[318] Y. Zhu, D. Qian, D. Ren, and H. Xia, "Starnet: Pedestrian trajectory prediction using deep neural network in star topology," in *International Conference on Intelligent Robots and Systems*, 2019.

[319] H. Xue, D. Huynh, and M. Reynolds, "Location-velocity attention for pedestrian trajectory prediction," in *Winter Conference on Applications of Computer Vision*, 2019.

[320] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *Conference on Computer Vision and Pattern Recognition*, 2018.

[321] I. Hasan, F. Setti, T. Tsesmelis, A. Del Bue, F. Galasso, and M. Cristani, "Mx-lstm: Mixing tracklets and vislets to jointly forecast trajectories and head poses," in *Conference on Computer Vision and Pattern Recognition*, 2018.

[322] Y. Xu, Z. Piao, and S. Gao, "Encoding crowd interaction with deep neural network for pedestrian trajectory prediction," in *Conference on Computer Vision and Pattern Recognition*, 2018.

[323] T. Yao, M. Wang, B. Ni, H. Wei, and X. Yang, "Multiple granularity group interaction prediction," in *Conference on Computer Vision and Pattern Recognition*, 2018.

[324] M. Pfeiffer, G. Paolo, H. Sommer, J. Nieto, R. Siegwart, and C. Cadena, "A data-driven model for interaction-aware pedestrian motion prediction in object cluttered environments," in *International Conference on Robotics and Automation*, 2018.

[325] E. Rehder, F. Wirth, M. Lauer, and C. Stiller, "Pedestrian prediction by planning using deep neural networks," in *International Conference on Robotics and Automation*, 2018.

[326] H. Xue, D. Q. Huynh, and M. Reynolds, "Ss-lstm: A hierarchical lstm model for pedestrian trajectory prediction," in *Winter Conference on Applications of Computer Vision*, 2018.

[327] F. Bartoli, G. Lisanti, L. Ballan, and A. Del Bimbo, "Context-aware trajectory prediction," in *International Conference on Pattern Recognition*, 2018.

[328] J. Hong, B. Sapp, and J. Philbin, "Rules of the road: Predicting driving behavior with a convolutional model of semantic interactions," in *Conference on Computer Vision and Pattern Recognition*, 2019.

[329] N. Rhinehart, R. McAllister, K. Kitani, and S. Levine, "Precog: Prediction conditioned on goals in visual multi-agent settings," in *International Conference on Computer Vision*, 2019.

[330] J. Li, H. Ma, and M. Tomizuka, "Conditional generative neural system for probabilistic trajectory prediction," in *International Conference on Intelligent Robots and Systems*, 2019.

[331] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Gd-gan: Generative adversarial networks for trajectory prediction and group detection in crowds," in *Asian Conference on Computer Vision*, C. V. Jawahar, H. Li, G. Mori, and K. Schindler, Eds., 2019.

[332] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. S. Torr, and M. Chandraker, "Desire: Distant future prediction in dynamic scenes with interacting agents," in *Conference on Computer Vision and Pattern Recognition*, 2017.

[333] N. Rhinehart, K. M. Kitani, and P. Vernaza, "R2p2: A reparameterized pushforward policy for diverse, precise generative path forecasting," in *European Conference on Computer Vision*, 2018.

[334] K.-R. Kim, W. Choi, Y. J. Koh, S.-G. Jeong, and C.-S. Kim, "Instance-level future motion estimation in a single image based on ordinal regression," in *International Conference on Computer Vision*, 2019.

[335] Y. Chai, B. Sapp, M. Bansal, and D. Anguelov, "Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction," in *Conference on Robot Learning*, 2019.

[336] H. Cui, V. Radosavljevic, F. Chou, T. Lin, T. Nguyen, T. Huang, J. Schneider, and N. Djuric, "Multimodal trajectory predictions for autonomous driving using deep convolutional networks," in *International Conference on Robotics and Automation*, 2019.

[337] X. Huang, S. G. McGill, B. C. Williams, L. Fletcher, and G. Rosman, "Uncertainty-aware driver trajectory prediction at urban intersections," in *International Conference on Robotics and Automation*, 2019.

[338] S. Zhou, M. J. Phielipp, J. A. Sefair, S. I. Walker, and H. B. Amor, "Clone swarms: Learning to predict and control multi-robot systems by imitation," in *International Conference on Intelligent Robots and Systems*, 2019.

[339] A. Jain, S. Casas, R. Liao, Y. Xiong, S. Feng, S. Segal, and R. Urtasun, "Discrete residual flow for probabilistic pedestrian behavior prediction," in *Conference on Robot Learning*, 2019.

[340] U. Baumann, C. Guiser, M. Herman, and J. M. Zollner, "Predicting ego-vehicle paths from environmental observations with a deep neural network," in *International Conference on Robotics and Automation*, 2018.

[341] S. Casas, W. Luo, and R. Urtasun, "Intentnet: Learning to predict intention from raw sensor data," in *Conference on Robot Learning*, 2018.

[342] Y. Zhang, W. Wang, R. Bonatti, D. Maturana, and S. Scherer, "Integrating kinematics and environment context into deep inverse reinforcement learning for predicting off-road vehicle trajectories," in *Conference on Robot Learning*, 2018.

[343] W.-C. Ma, D.-A. Huang, N. Lee, and K. M. Kitani, "Forecasting interactive dynamics of pedestrians with fictitious play," in *Conference on Computer Vision and Pattern Recognition*, 2017.

[344] S. Yi, H. Li, and X. Wang, "Pedestrian behavior understanding and prediction with deep neural networks," in *European Conference on Computer Vision*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., 2016.

[345] R. Chandra, U. Bhattacharya, A. Bera, and D. Manocha, "Traphic: Trajectory prediction in dense and heterogeneous traffic using weighted interactions," in *Conference on Computer Vision and Pattern Recognition*, 2019.

[346] Y. Li, "Which way are you going? imitative decision learning for path forecasting in dynamic scenes," in *Conference on Computer Vision and Pattern Recognition*, 2019.

[347] W. Zhi, L. Ott, and F. Ramos, "Kernel trajectory maps for multi-modal probabilistic motion prediction," in *Conference on Robot Learning*, 2019.

[348] A. Vemula, K. Muelling, and J. Oh, "Social attention: Modeling attention in human crowds," in *International Conference on Robotics and Automation*, 2018.

[349] C. Tang, J. Chen, and M. Tomizuka, "Adaptive probabilistic vehicle trajectory prediction through physically feasible bayesian recurrent neural network," in *International Conference on Robotics and Automation*, 2019.

[350] K. Cho, T. Ha, G. Lee, and S. Oh, "Deep predictive autonomous driving using multi-agent joint trajectory prediction and traffic rules," in *International Conference on Intelligent Robots and Systems*, 2019.

[351] Z. Fang, D. Vázquez, and A. López, "On-board detection of pedestrian intentions," *Sensors*, vol. 17, no. 10, p. 2193, 2017.

[352] T. Bandyopadhyay, K. S. Won, E. Frazzoli, D. Hsu, W. S. Lee, and D. Rus, "Intention-aware motion planning," in *Algorithmic Foundations of Robotics X*, 2013, pp. 475–491.

[353] E. Rehder and H. Kloeden, "Goal-directed pedestrian prediction," in *International Conference on Computer Vision Workshops*, 2015.

[354] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," *Computer Graphics Forum*, vol. 26, no. 3, pp. 655–664, 2007.

[355] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *International Conference on Computer Vision*, 2009.

[356] B. Majecka, "Statistical models of pedestrian behaviour in the forum," Master's thesis, School of Informatics, University of Edinburgh, 2009.

[357] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *European conference on computer vision*, 2016.

[358] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis *et al.*, "A large-scale benchmark dataset for event recognition in surveillance video," in *Conference on Computer Vision and Pattern Recognition*, 2011.

[359] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *Conference on Computer Vision and Pattern Recognition*, 2011.

[360] B. Zhou, X. Wang, and X. Tang, "Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents," in *Conference on Computer Vision and Pattern Recognition*, 2012.

[361] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: uses in assessing rater reliability." *Psychological Bulletin*, vol. 86, no. 2, p. 420, 1979.

[362] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014.

[363] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Coputer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[364] T. Tieleman and G. Hinton, "Lecture 6.5-RMSProp, COURSERA: Neural networks for machine learning," *University of Toronto, Technical Report*, 2012.

[365] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[366] Q. Ke, M. Fritz, and B. Schiele, "Time-conditioned action anticipation in one shot," in *Conference on Computer Vision and Pattern Recognition*, 2019.

[367] A. Furnari and G. M. Farinella, "What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention," in *International Conference on Computer Vision*, 2019.

[368] F. Sener and A. Yao, "Zero-shot anticipation for instructional activities," in *International Conference on Computer Vision*, 2019.

[369] E. Alati, L. Mauro, V. Ntouskos, and F. Pirri, "Help by predicting what to do," in *International Conference on Image Processing*, 2019.

[370] A. Furnari and G. M. Farinella, "Egocentric action anticipation by disentangling encoding and inference," in *International Conference on Image Processing*, 2019.

[371] Y. Abu Farha, A. Richard, and J. Gall, "When will you do what? - anticipating temporal occurrences of activities," in *Conference on Computer Vision and Pattern Recognition*, 2018.

[372] C. Sun, A. Shrivastava, C. Vondrick, R. Sukthankar, K. Murphy, and C. Schmid, "Relational action forecasting," in *Conference on Computer Vision and Pattern Recognition*, 2019.

[373] X. Wang, J.-F. Hu, J.-H. Lai, J. Zhang, and W.-S. Zheng, "Progressive teacher-student learning for early action prediction," in *Conference on Computer Vision and Pattern Recognition*, 2019.

[374] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Predicting the future: A jointly learnt model for action anticipation," in *International Conference on Computer Vision*, 2019.

[375] H. Zhao and R. P. Wildes, "Spatiotemporal feature residual propagation for action prediction," in *International Conference on Computer Vision*, 2019.

[376] Y. Shi, B. Fernando, and R. Hartley, "Action anticipation with rbf kernelized feature mapping rnn," in *The European Conference on Computer Vision*, 2018.

[377] J. Btepage, H. Kjellstrm, and D. Kragic, "Anticipating many futures: Online human motion prediction and generation for human-robot interaction," in *International Conference on Robotics and Automation*, 2018.

[378] S. Cho and H. Foroosh, "A temporal sequence learning for action recognition and prediction," in *Winter Conference on Applications of Computer Vision*, 2018.

[379] M. Sadegh Aliakbarian, F. Sadat Saleh, M. Salzmann, B. Fernando, L. Petersson, and L. Andersson, "Encouraging lstms to anticipate actions very early," in *International Conference on Computer Vision*, 2017.

[380] W. Li and M. Fritz, "Recognition of ongoing complex activities by sequence prediction over a hierarchical label space," in *Winter Conference on Applications of Computer Vision*, 2016.

[381] H. Joo, T. Simon, M. Cikara, and Y. Sheikh, "Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction," in *Conference on Computer Vision and Pattern Recognition*, 2019.

[382] M. Wu, T. Louw, M. Lahijanian, W. Ruan, X. Huang, N. Merat, and M. Kwiatkowska, "Gaze-based intention anticipation over driving manoeuvres in semi-autonomous vehicles," in *International Conference on Intelligent Robots and Systems*, 2019.

[383] S. Qi, S. Huang, P. Wei, and S.-C. Zhu, "Predicting human activities using stochastic grammar," in *International Conference on Computer Vision*, 2017.

[384] N. Rhinehart and K. M. Kitani, "First-person activity forecasting with online inverse reinforcement learning," in *International Conference on Computer Vision*, 2017.

[385] 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 *International Conference on Computer Vision*, 2015.

[386] Y. Zhou and T. L. Berg, "Temporal perception and prediction in ego-centric video," in *International Conference on Computer Vision*, 2015.

[387] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Forecasting future action sequences with neural memory networks," in *British Machine Vision Conference*, 2019.

[388] T. Mahmud, M. Hasan, and A. K. Roy-Chowdhury, "Joint prediction of activity labels and starting times in untrimmed videos," in *International Conference on Computer Vision*, 2017.

[389] W. Ding, J. Chen, and S. Shen, "Predicting vehicle behaviors over an extended horizon using behavior interaction network," in *International Conference on Robotics and Automation*, 2019.

[390] P. Gujjar and R. Vaughan, "Classifying pedestrian actions in advance using predicted video of urban driving scenes," in *International Conference on Robotics and Automation*, 2019.

[391] K. Saleh, M. Hossny, and S. Nahavandi, "Real-time intent prediction of pedestrians for autonomous ground vehicles via spatio-temporal densenet," in *International Conference on Robotics and Automation*, 2019.

[392] O. Scheel, N. S. Nagaraja, L. Schwarz, N. Navab, and F. Tombari, "Attention-based lane change prediction," in *International Conference on Robotics and Automation*, 2019.

[393] M. S. Aliakbarian, F. S. Saleh, M. Salzmann, B. Fernando, L. Petersson, and L. Andersson, "Viena: A driving anticipation dataset," in *Asian Conference on Computer Vision*, C. V. Jawahar, H. Li, G. Mori, and K. Schindler, Eds., 2019.

[394] M. Strickland, G. Fainekos, and H. B. Amor, "Deep predictive models for collision risk assessment in autonomous driving," in *International Conference on Robotics and Automation*, 2018.

[395] F.-H. Chan, Y.-T. Chen, Y. Xiang, and M. Sun, "Anticipating accidents in dashcam videos," in *Asian Conference on Computer Vision*, S.-H. Lai, V. Lepetit, K. Nishino, and Y. Sato, Eds., 2017.

[396] A. Jain, A. Singh, H. S. Koppula, S. Soh, and A. Saxena, "Recurrent neural networks for driver activity anticipation via sensory-fusion architecture," in *International Conference on Robotics and Automation*, 2016.

[397] A. Manglik, X. Weng, E. Ohn-Bar, and K. M. Kitani, "Forecasting time-to-collision from monocular video: Feasibility, dataset, and challenges," in *International Conference on Intelligent Robots and Systems*, 2019.

[398] P. Wang, S. Lien, and M. Lee, "A learning-based prediction model for baby accidents," in *International Conference on Image Processing*, 2019.

[399] T. Suzuki, H. Kataoka, Y. Aoki, and Y. Satoh, "Anticipating traffic accidents with adaptive loss and large-scale incident db," in *The Conference on Computer Vision and Pattern Recognition*, 2018.

[400] K.-H. Zeng, S.-H. Chou, F.-H. Chan, J. Carlos Niebles, and M. Sun, "Agent-centric risk assessment: Accident anticipation and risky region localization," in *Conference on Computer Vision and Pattern Recognition*, 2017.

[401] S. Su, J. Pyo Hong, J. Shi, and H. Soo Park, "Predicting behaviors of basketball players from first person videos," in *Conference on Computer Vision and Pattern Recognition*, 2017.

[402] P. Felsen, P. Agrawal, and J. Malik, "What will happen next? forecasting player moves in sports videos," in *International Conference on Computer Vision*, 2017.

[403] Y. Shen, B. Ni, Z. Li, and N. Zhuang, "Egocentric activity prediction via event modulated attention," in *European Conference on Computer Vision*, 2018.

[404] P. Schydlo, M. Rakovic, L. Jamone, and J. Santos-Victor, "Anticipation in human-robot cooperation: A recurrent neural network approach for multiple action sequences prediction," in *International Conference on Robotics and Automation*, 2018.

[405] Y. Zhong and W. Zheng, "Unsupervised learning for forecasting action representations," in *International Conference on Image Processing*, 2018.

[406] K.-H. Zeng, W. B. Shen, D.-A. Huang, M. Sun, and J. Carlos Niebles, "Visual forecasting by imitating dynamics in natural sequences," in *International Conference on Computer Vision*, 2017.

[407] J. Gao, Z. Yang, and R. Nevatia, "Red: Reinforced encoder-decoder networks for action anticipation," in *British Machine Vision Conference*, 2017.

[408] C. Vondrick, H. Pirsiavash, and A. Torralba, "Anticipating visual representations from unlabeled video," in *Conference on Computer Vision and Pattern Recognition*, 2016.

[409] H. Kataoka, Y. Miyashita, M. Hayashi, K. Iwata, and Y. Satoh, "Recognition of transitional action for short-term action prediction using discriminative temporal cnn feature," in *British Machine Vision Conference*, 2016.

[410] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-rnn: Deep learning on spatio-temporal graphs," in *Conference on Computer Vision and Pattern Recognition*, 2016.

[411] O. Scheel, L. Schwarz, N. Navab, and F. Tombari, "Situation assessment for planning lane changes: Combining recurrent models and prediction," in *International Conference on Robotics and Automation*, 2018.

[412] A. S. Yoav Levine, Or Sharir, "Benefits of depth for long-term memory of recurrent networks," in *International Conference on Learning Representations*, 2018.

[413] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Conference on Computer Vision and Pattern Recognition*, 2015.

[414] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Conference on Computer Vision and Pattern Recognition*, 2015.

[415] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv:1412.3555*, 2014.

[416] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Conference on Computer Vision and Pattern Recognition*, 2017.

[417] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, 2014.

[418] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 2015.

[419] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Conference on Computer Vision and Pattern Recognition*, 2015.

9.4 Appendix A: Chapters, Corresponding Publications and Contribution

Below, we list the contribution of other authors' work and corresponding publications for each chapter.

Chapter 1: Introduction

Contribution

100% of the work presented in this chapter was done by the author of this dissertation.

Paper(s)

A. Rasouli and J. K. Tsotsos, "Joint attention in driver-pedestrian interaction: From theory to practice," arXiv, 2018

Chapter 2: Social Interaction, Coordination and Behavior Prediction

Contribution

100% of the work presented in this chapter was done by the author of this dissertation.

Paper(s)

A. Rasouli and J. K. Tsotsos, "Joint attention in driver-pedestrian interaction: From theory to practice," arXiv, 2018

Chapter 3: Traffic Context and its Influence on Pedestrian Behavior

Contribution

100% of the work presented in this chapter was done by the author of this dissertation.

Paper(s)

A. Rasouli and J. K. Tsotsos, "Autonomous vehicles that interact with pedestrians: A survey of theory and practice," Transactions on Intelligent Transportation Systems, 2019

Chapter 4: Pedestrian Communication in Traffic

Contribution

80% of the work presented in this chapter was done by the author of this dissertation. I. Kotseruba contributed on collecting and annotating the dataset that was used in this chapter.

Paper(s)

I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Joint attention in autonomous driving (JAAD)," arXiv, 2016.

A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior," International Conference on

Computer Vision (ICCV) Workshops 2017.
A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Towards Social Autonomous Vehicles: Understanding Pedestrian-Driver Interactions," Intelligent Transportation Systems Conference (ITSC), 2018.
A. Rasouli and J. K. Tsotsos, "Autonomous vehicles that interact with pedestrians: A survey of theory and practice," Transactions on Intelligent Transportation Systems, 2019

Chapter 5: Understanding Pedestrian Crossing Behavior: An Empirical Study

Contribution

65% of the work presented in this chapter was done by the author of this dissertation. I. Kotseruba contributed on collecting and annotating the dataset that was used in this chapter. She also performed a part of analysis on pedestrian walking patterns.

Paper(s)

A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Agreeing to cross: How drivers and pedestrians communicate," Intelligent Vehicle Symposium (IV), 2017.
A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Understanding pedestrian behavior in complex traffic scenes," Transactions on Intelligent Vehicles, 2018.

Chapter 6: Detecting Pedestrians in Cluttered Traffic Scenes

Contribution

60% of the work presented in this chapter was done by the author of this dissertation. I. Kotseruba contributed on collecting and annotating the dataset that was used in this chapter. She also performed a part of experiments for evaluating detection algorithms and illustration of the results.

Paper(s)

A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "It's Not All About Size: On the Role of Data Properties in Pedestrian Detection," European Conference on Computer Vision (ECCV), 2018

Chapter 7: Understanding Pedestrians Intentions and Their Role in Predicting Trajectories

Contribution

50% of the work presented in this chapter was done by the author of this dissertation. I. Kotseruba contributed on collecting and 30 % of annotating the dataset that was used in this chapter. In addition, she conducted human experiments, compiled its results and developed and tested the intention component of the proposed model. T.

Kunic helped with designing the interface for conducting human experiments.

Paper(s)

A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, "PIE: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction", International Conference on Computer Vision (ICCV), 2019.

Chapter 8: Understanding Pedestrians Intentions and Their Role in Predicting Trajectories

Contribution

85% of the work presented in this chapter was done by the author of this dissertation. I. Kotseruba generated the pose information for pedestrian samples.

Paper(s)

A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Pedestrian Action Anticipation using Contextual Feature Fusion in Stacked RNNs", British Machine Vision Conference, 2019.

I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Do They Want to Cross? Understanding Pedestrian Intention for Behavior Prediction", Intelligent Vehicle Symposium (IV), 2020.