# Towards pedestrian-aware autonomous cars

*Citation for published version (APA):*
Dolatabadi, M. (2021). *Towards pedestrian-aware autonomous cars*. [Phd Thesis 1 (Research TU/e / Graduation TU/e), Mechanical Engineering]. Eindhoven University of Technology.

*Document status and date:*
Published: 01/12/2021

*Document Version:*
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

# Towards Pedestrian-aware Autonomous Cars

Marzieh Dolatabadi

# Towards pedestrian-aware autonomous cars

Marzieh Dolatabadi

Department of Mechanical Engineering
EINDHOVEN UNIVERSITY OF TECHNOLOGY
Eindhoven, The Netherlands. 2021

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

European Commission

AUTOPILOT

The work described in this thesis was carried out at the Eindhoven University of Technology and is part of the research program “Autopilot” funded from the European Union’s Horizon 2020 Framework Program for Research and Innovation under grant agreement No. 731993.

A catalogue record is available from the Eindhoven University of Technology Library.
ISBN: 978-90-386-5425-6

Reproduction: ProefschriftMaken|| www.proefschriftmaken.nl, the Netherlands

Cover photo: photo by Marzieh Dolatabadi, photo editing by Malihe Dolatabadi

# Towards pedestrian-aware autonomous cars

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus prof.dr.ir. F.P.T. Baaijens, voor een commissie aangewezen door het College voor Promoties, in het openbaar te verdedigen op
woensdag 01 december 2021 om 11 uur

door

Marzieh Dolatabadi

geboren te Tehran, Iran

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

| | |
|---|---|
| voorzitter: | prof.dr.ir A. van Steenhoven |
| 1e promotor: | dr.ir. M.J.G. van de Molengraft |
| 2e promotor: | prof.dr.ir M. Steinbuch |
| Copromotor: | dr.ir Elfring. |
| leden: | prof.dr. D. Gavrila (Technische Universiteit Delft) |
| | prof.dr. H.P.J. Bruyninckx |
| | dr.ir. I.J.M. Besselink |

*Het onderzoek dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.*

*Dedicated to my parents.*

# Summary

Autonomous cars are anticipated to gain significant attention in the market over the following decades. Despite the considerable progress in autonomous cars, foreseeable challenges persist, including pedestrian awareness systems. Based on the world health organization (WHO), more than one-fifth of road traffic deaths worldwide are pedestrians. Therefore, one of the core requirements underlying many of the possible tasks that autonomous cars could perform is a description of the environment in terms of pedestrians. Thus, this research focuses on improving pedestrian-aware systems for autonomous cars in an urban environment.

The tracking of multiple pedestrians is one of the vital tasks of autonomous vehicles. This includes estimating the positions and velocities of pedestrians surrounding a vehicle. This thesis proposes a tracker that receives the ankle, knee, and hip positions as measurements to track pedestrians based on human motion patterns. Then, based on the legs' reflection and extension angles, the tracker estimates pedestrians' position and velocity. To overcome this critical issue of existing pedestrian detection, we can take advantage of Internet-of-Things (IoT) technologies. We use both IoT technology and a camera to track pedestrians in this work.

Even the most accurate pedestrian trackers are affected by measurement noise, background clutter, and occlusion. Such uncertainties can cause deviations in sensors' data association, thereby leading to challenging situations from a tracking perspective and potentially even the failure of a tracker. To improve data association's accuracy and reduce the number of false tracks, we propose steps to find a trade-off between the parameters of a probabilistic data association model. The results show that the tracking precision and accuracy increase up to 3.6% with the proposed initialization compared to the state-of-the-art algorithms in tracking multiple pedestrians. After detecting and localizing pedestrians related to the vehicle, intention prediction and action recognition are two critical tasks to drive safely and smoothly. In particular, knowing the intention of a pedestrian to cross on a piece of road that is used by the vehicle in the near future, before the pedestrian has entered the road, would allow the vehicle to perform smoother maneuvers. Intention can be predicted using previous actions of pedestrians. Examples of such actions are walking, starting, standing, and stopping . As a result of pedestrians' impending motion uncertainties, the pedestrians' intention prediction and action recognition are not trivial tasks. To recognize the current action of pedestrians, we utilize a unique set of body features that are distinctive among pedestrian actions. To predict intention, we tackle intention prediction by observing pedestrians' distance to the vehicle, action, and spatio-temporal context information. Spatio-temporal context information includes traffic signs, environmental factors, zebra-crossings, pedestrians' occlusions with elements in the scene, and pedestrians' gaze information.

# Contents

**Summary** **vii**

**List of Figures** **xiii**

**List of Tables** **xvii**

**1 Introduction** **1**

1.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . 1

1.2 Contributions . . . . . . . . . . . . . . . . . . . . . . . 3

1.3 State-of-the-art . . . . . . . . . . . . . . . . . . . . . . 4

1.3.1 State estimation . . . . . . . . . . . . . . . . . . 4

1.3.2 Data association . . . . . . . . . . . . . . . . . . 7

1.3.3 Environment description . . . . . . . . . . . . 8

1.4 Structure of the thesis . . . . . . . . . . . . . . . . . . 10

1.4.1 Chapter 2: Multiple joints pedestrian tracker 11

1.4.2 Chapter 3: Data association improvement 11

1.4.3 Chapter 4: Vision-IOT based tracker . . . . 12

1.4.4 Chapter 5: Pedestrian behavior predictor . 12

**2 Multiple-Joint Pedestrian Tracking Using Periodic Models** **13**

2.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . 14

2.2 Related work . . . . . . . . . . . . . . . . . . . . . . . . 17

2.3 Pedestrian tracker . . . . . . . . . . . . . . . . . . . . . 19

2.4 Proposed models . . . . . . . . . . . . . . . . . . . . . . 22

2.4.1 Process model . . . . . . . . . . . . . . . . . . . 22

2.4.2 Measurement model . . . . . . . . . . . . . . . 25

2.5 Data association . . . . . . . . . . . . . . . . . . . . 27
2.6 Performance . . . . . . . . . . . . . . . . . . . . . . . 28
2.7 Experimental evaluation . . . . . . . . . . . . . . . . 29
2.7.1 HuGaDB dataset . . . . . . . . . . . . . . . . . 30
2.7.2 Comparison . . . . . . . . . . . . . . . . . . . 31
2.7.3 KITTI dataset . . . . . . . . . . . . . . . . . . 33
2.8 Conclusions . . . . . . . . . . . . . . . . . . . . . . . 35

**3 Improved Data Association of Hypothesis-Based Trackers Using Object Initialization 43**
3.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . 45
3.2 Related work . . . . . . . . . . . . . . . . . . . . . . 48
3.3 Multiple hypothesis tracker . . . . . . . . . . . . . . 51
3.3.1 Basic idea . . . . . . . . . . . . . . . . . . . . 51
3.3.2 Reason . . . . . . . . . . . . . . . . . . . . . . 53
3.4 VRU tracker . . . . . . . . . . . . . . . . . . . . . . . 54
3.4.1 State estimator . . . . . . . . . . . . . . . . . 54
3.4.2 Performance . . . . . . . . . . . . . . . . . . . 55
3.5 Results . . . . . . . . . . . . . . . . . . . . . . . . . . 56
3.5.1 Simulation . . . . . . . . . . . . . . . . . . . . 58
3.5.2 Experimental setup . . . . . . . . . . . . . . . 62
3.5.3 KITTI raw data . . . . . . . . . . . . . . . . . 67
3.5.4 Comparison . . . . . . . . . . . . . . . . . . . 69
3.6 Conclusions . . . . . . . . . . . . . . . . . . . . . . . 71

**4 Multiple pedestrian tracking using vision-based sensors and IoT technology 73**
4.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . 74
4.2 Related work . . . . . . . . . . . . . . . . . . . . . . 75
4.3 Input data . . . . . . . . . . . . . . . . . . . . . . . . 78
4.4 Pedestrian tracker . . . . . . . . . . . . . . . . . . . 79
4.4.1 IoT state estimation . . . . . . . . . . . . . . . 80
4.4.2 Data association . . . . . . . . . . . . . . . . . 81
4.5 Performance evaluation . . . . . . . . . . . . . . . . 83
4.5.1 Simulation . . . . . . . . . . . . . . . . . . . . 83
4.5.2 Scenario 1 . . . . . . . . . . . . . . . . . . . . 85
4.5.3 Scenario 2 . . . . . . . . . . . . . . . . . . . . 87
4.5.4 Experiment . . . . . . . . . . . . . . . . . . . . 90

4.6 Conclusion . . . . . . . . . . . . . . . . . . . . . . . 93

**5 Intention Prediction and Action Recognition of Pedestrians 95**

5.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . 97

5.2 Related work . . . . . . . . . . . . . . . . . . . . . . . 99

5.3 Framework architecture . . . . . . . . . . . . . . . . . 104

5.4 Action . . . . . . . . . . . . . . . . . . . . . . . . . . . 106

5.4.1 Action recognition . . . . . . . . . . . . . . . . 107

5.5 Intention . . . . . . . . . . . . . . . . . . . . . . . . . . 108

5.6 HMM . . . . . . . . . . . . . . . . . . . . . . . . . . . . 110

5.6.1 Measurement and state . . . . . . . . . . . . . 110

5.6.2 Prediction . . . . . . . . . . . . . . . . . . . . . 111

5.6.3 Training HMM . . . . . . . . . . . . . . . . . . 112

5.7 Evaluation . . . . . . . . . . . . . . . . . . . . . . . . . 113

5.7.1 Action Experimental Evaluation . . . . . . 114

5.7.2 Intention Experimental Evaluation . . . . . 117

5.8 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . 123

**6 Conclusions and Future Work 125**

6.1 Conclusions and Recommendation . . . . . . . . . 125

6.2 Future work . . . . . . . . . . . . . . . . . . . . . . . . 128

**Acknowledgments 131**

**Curriculum Vitae 135**

**Bibliography 137**

# List of Figures

1.1 Distribution of Deaths by Road User Type by WHO Region in 2018 [Organization et al. 2018]. . . . . 2

2.1 The joints of interest to detect and use in this tracker are at the ankles, knees, and hips. . . . . . 20
2.2 General framework of a tracker. . . . . . . . . . . . 21
2.3 The right leg from the side view in a schematic way. $a_{hk}$ corresponds to a length between the hip and the knee. $a_{ka}$ is a length between the knee and ankle. Both of the angles are defined as positive in counterclockwise direction. . . . . . . . 22
2.4 A campus pedestrians tracking.The training sequences number 17 of the Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) tracking benchmark. . . . . . . . . . . . . . . . . . . . 29
2.5 Result of OpenPose joints detection.The training sequences number 16 of the KITTI tracking benchmark. . . . . . . . . . . . . . . . . . . . . . . . . . . . 29
2.8 A pedestrian is crossing in front of a camera with constant velocity. . . . . . . . . . . . . . . . . . . . . 32
2.6 Measured and estimated results of this tracker for the right leg of one participant in HuGaDB dataset who was a 24-year old male with 177 cm stature and 75 kg body mass. (**a**) shows the angle between thigh and hip ($\theta_{H_1}$). (**b**) shows ($\theta_{K_1}$) measured using the accelerometers and estimated using this tracker. . . . . . . . . . . . . . . 37

2.7 Measured and estimated results of the tracker for the right leg of one participant in HuGaDB dataset who was a 24-year old male with 177 cm stature and 75 kg body mass. (**a**) indicates $\omega_{H_1}$ and (**b**) compares the $\omega_{K_1}$ between the tracker and the dataset. 38

2.9 Measured and estimated results of the trackers for a person who was crossing in front of the camera. (**a**) shows the position of the hip joint in the left leg. (**b**) indicates the position of the knee joint in the left leg, and (**c**) compares the two trackers with each other based on the position of the ankle of the left leg. . . . . . . . . . . . . . . . . 39

2.10 In (**a**), one pedestrian occludes another one, and the legs of two pedestrians are occluded by a car. In (**b**), one of the occluded pedestrians is in the field of view of the camera. In (**c**), OpenPose detects their joints. In situations such as (**a**) and (**b**), OpenPose can not detect pedestrians, affecting the results of Multiple Object Tracking Precision (MOTP). . . . . . . . . . . . . . . . . . . . . . 40

2.11 The result of the tracking two pedestrians in the sequences number 15 of the KITTI tracking benchmark. The vertical axes of the figure is the displacement of the pedestrians to a standing car. In the figure, *ND* means no detection and *PD* indicates partial detection. . . . . . . . . . . . . . . . 41

3.1 A framework to represent the place of the data association's initialization. . . . . . . . . . . . . . . 47

3.2 Tree representation of the formed multiple hypotheses. Black dot represents a hypothesis, and a red dot represents a hypothesis that pruned. . . . 53

3.3 The car frame coordinate system. $x_c$ denotes the direction of driving, and $y_c$ indicates the side direction of the car. . . . . . . . . . . . . . . . . . . . . 55

3.4 Sequence '5' from the KITTI raw dataset. . . . . . 58

3.5 Simulate the effect of different sets of probabilities on *MOTP* with static VRUs and without covariance noises. . . . . . . . . . . . . . . . . . . . . . . 60
3.6 Simulate the effect of different sets of probabilities on *MOTP* with dynamic VRUs and without covariance noises. . . . . . . . . . . . . . . . . . . . . 60
3.7 Simulate the effect of different sets of probabilities on *MOTP* with static VRUs and with one meter measurement noise. . . . . . . . . . . . . . . . . 61
3.8 Simulate the effect of different sets of probabilities on *MOTP* with dynamic VRUs and with one meter measurement noise. . . . . . . . . . . . . . . 62
3.9 Illustration of our experimental test. The car receives the center of each rectangle as the position of pedestrians. . . . . . . . . . . . . . . . . . . . . . 63
3.10 Different sets of probabilities affect MOTP of an experiment test. . . . . . . . . . . . . . . . . . . . 65
3.11 This detection contains a false positive since our detection method detected a bicycle as a VRU. Therefore, our tracker should select a hypothesis that assumes those bicycles are false-positives. . . 66
3.12 *SE* of the data association based on the procedure and the Poisson. . . . . . . . . . . . . . . . . . . 70

4.1 In the situation number 1,the vehicle has no vision data only IoT data. In the situation number 2, the vehicle receives IoT data and a partial measurement from its camera. The arrow shows vehicle's moving direction. . . . . . . . . . . . . . 84
4.2 Average MOTA for each test case . . . . . . . . . . 87
4.3 A schematic of experimental tests, a vehicle should start driving autonomously from start point to the blue flag. . . . . . . . . . . . . . . . . . . . . . . . . 90
4.4 One pedestrian occludes another pedestrian. Therefore camera is not able to detect one of them, while both of them have IoT devices. . . . . . . . 91
4.5 One of the environmental noises that our vehicle faced during tests. . . . . . . . . . . . . . . . . . . . 92

5.1 Pedestrian is walking in place and is preparing herself to walk. The number below each image indicates its sequence order [Rasouli, Kotseruba, et al. 2017b]. . . . . . . . . . . . . . . . . . . . . . . . 102
5.2 Block diagram of this framework. . . . . . . . . . . 105
5.3 Due to occlusion, there is wrong information regarding the frequency [Rasouli, Kotseruba, et al. 2017b]. . . . . . . . . . . . . . . . . . . . . . . . . . . 108
5.4 State diagram for modeling pedestrian actions. . . 108
5.5 Information regarding traffic scene [Rasouli, Kotseruba, et al. 2017b]. . . . . . . . . . . . . . . . . . . 109
5.6 Information regarding pedestrians annotation in the JAAD dataset [Rasouli, Kotseruba, et al. 2017b].109
5.7 Results of our action recognizer for one of the videos in the JAAD dataset. GT means ground truth actions. . . . . . . . . . . . . . . . . . . . . . . . 116
5.8 Results of our action recognizer for one of the videos in the JAAD dataset. GT means ground truth actions. . . . . . . . . . . . . . . . . . . . . . . . 116
5.9 A sequence of the JAAD dataset that a pedestrian crosses the main road at a random location and breaks the traffic rules [Rasouli, Kotseruba, et al. 2017b]. . . . . . . . . . . . . . . . . . . . . . . . . . . 119
5.10 Average stride frequency of walking pedestrians in JAAD dataset during rainy days and clear days. 122
5.11 There is no measurements regarding the context information [Rasouli, Kotseruba, et al. 2017b]. . . 122

# List of Tables

2.1 Evaluation metrics for tracking all 40 participates in HuGaDB dataset. . . . . . . . . . . . . . . . . . . 31

2.2 Evaluation metrics for a person tracking based on the sequence of images. . . . . . . . . . . . . . . . . 33

2.3 Multiple Target tracking evaluation metrics for KITTI Pedestrian tracking benchmark. . . . . . . 33

3.1 Probabilities sets that cause the minimum MOTP for each scenario. . . . . . . . . . . . . . . . . . . 63

3.2 Minimum and maximum of MOTP for each experiment . . . . . . . . . . . . . . . . . . . . . . . . 65

3.3 Average of the minimum and maximum MOTP and MOTA for all the experiments . . . . . . . . . . 66

3.4 Average MOTP and MOTA in a set of probabilities 67

3.5 Minimum *SE* for each sequences . . . . . . . . . . 67

3.6 Compare our work with a state-of-the-art algorithm on KITTI raw dataset . . . . . . . . . . . . . . 68

3.7 *SE* of two different initialization procedures. . . . 70

4.1 A fixed range of noise for each test case . . . . . . 86

4.2 Average MOTA(%) in the proposed tracker with both camera images and IoT devices in the simulation environment. . . . . . . . . . . . . . . . . . . 89

4.3 Difference between average MOTA in the two different cases. . . . . . . . . . . . . . . . . . . . . . . 92

5.1 Action recognizing accuracy for the JAAD dataset 116

5.2 Intention recognizing accuracy for the JAAD dataset. (N.A means not available) . . . . . . . . . . . . . . . 121

Chapter 1

# Introduction

## 1.1 Introduction

Road transportation plays an essential role in daily human life. Using transportation, people can access services and activities, such as education, employment, shopping, and social events. Therefore, the quality of road transportation affects people's ability to participate in social and economic activities. One of the most critical issues in road transportation is the safety of driving. Safe transport solutions allow achieving reliable technologies in logistics services and smart cities.

Automation technologies such as autonomous vehicles can improve the safety of transportation systems and bring a wide range of global environmental and economic impacts. Moreover, autonomous vehicles are anticipated to be a key technology for addressing societal problems caused by the proliferation of automobiles worldwide [Abuelsamid et al. 2017]. These problems include traffic congestion, injuries, and fatalities caused by collisions. Based on the world health organization (WHO), road traffic crashes lead to the fatality of 1.3 million people each year [Organization et al. 2018]. Figure.1.1 shows the distribution of the road traffic fatalities by road user types. Based on a report by the European commission in 2017 [Commission 2017], replacing human drivers with autonomous vehicles can eliminate 90% of traffic fatalities. This is due to the fact that these vehicles use multiple sensors to observe their surrounding environment. Therefore, they can have a more complete perception than human drivers. Besides, these vehicles can communicate with each other and with infrastructure. Hence, they are anticipated to be safer than human drivers.

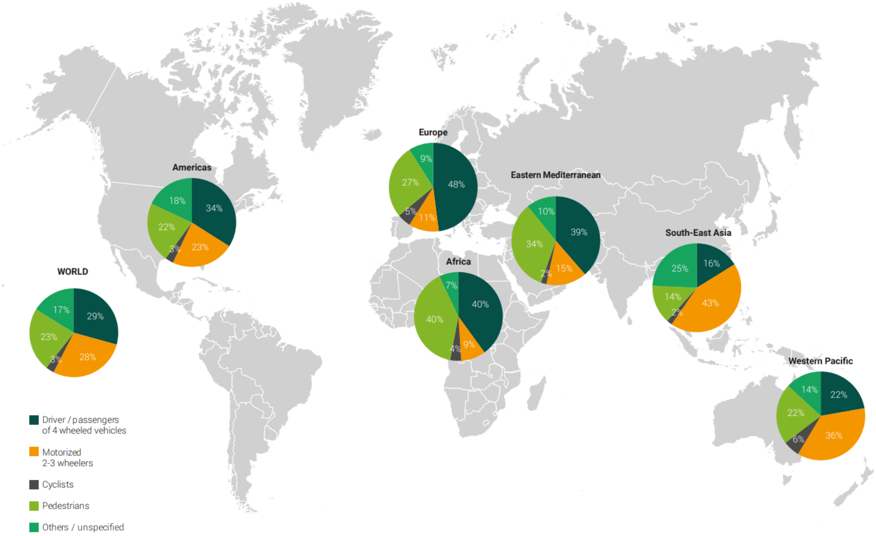


Figure 1.1: Distribution of Deaths by Road User Type by WHO Region in 2018 [Organization et al. 2018].

The transition from manual driving (by humans) to fully autonomous driving is expected to involve several semi-automated features such as awareness systems. One of the core aspects underlying the many possible tasks that the autonomous vehicle may perform is the awareness of the surrounding environment. An awareness system can be a combination of sensors, devices, software, and infrastructure. Its tasks are defined as identifying the road users, analyzing their behavior, communicating with them, predicting their future actions, and choosing an appropriate vehicle response. The vehicle response can include changing a route, increasing or decreasing acceleration, and braking.

In Europe, amongst all road users, pedestrians are known to be the most vulnerable [Commission 2017]. This thesis aims to evaluate, explore and, contribute various directions to describe the surrounding environment in terms of pedestrians. The proposed a pedestrian-aware system contains positions, velocities, and properties of pedestrians that are relevant to an approaching vehicle. Various algorithms are proposed to create and maintain the pedestrian-aware system.

## 1.2 Contributions

Developing an environment perception system is a broad topic in autonomous vehicles. The research presented in this thesis contributes to developing a methodology for creating and maintaining a pedestrian-aware system. Based on embedded sensor data, models, and information from external sources, a pedestrian-aware system estimates the actual state of pedestrians as well as its short-term prediction. Reflecting on the complicating factors while constructing a pedestrians awareness system, the core question is formulated to be answered in this thesis as follows:

How can a pedestrian-aware system estimate the state and describe the behavior of multiple pedestrians?

To answer the core question, the following key questions are to be addressed individually.

1. How can a motion model improve the accuracy and precision of a pedestrian tracker during occlusions?

2. How can a hypothesis tree be initialized faster compared to the ones existing in the current state-of-the-art?

3. How can additional information be combined to add tracking robustness?

4. How can a pedestrian-aware system predict the relevant behavior of pedestrians?

The following section explains challenges in constructing and maintaining pedestrian-aware systems for autonomous vehicles. Moreover, the state-of-the-art are explained by this chapter.

## 1.3 State-of-the-art

The awareness system is one of the most critical components in autonomous vehicles. Through the awareness system, vehicles discover their environment and can adapt their decisions to the current state of the world. The awareness systems' output can be used by various parts of a vehicle, such as navigation systems and vehicle control [Song et al. 2016]. The awareness system and road information are utilized in navigation systems to plan the path and find a route to reach a destination. Vehicle control systems utilize information from the awareness system and navigation system to perform kinematic commands. The kinematic commands can be defined as changing velocity, braking, parking the vehicle, and steering. Therefore, constructing an awareness system capable of describing the surrounding environment is critical for developing autonomous vehicles [Ojala et al. 2002].

The awareness system generally combines a motion model, a data association method, and an environment description. The environment description can be built from prior knowledge, onboard sensors measurements, and information obtained through communication with other vehicles or infrastructures. A motion model is defined as an algorithm to estimate and predict object states. The state can contain variables such as position, velocity, and orientation of an object. Objects are what the vehicle is surrounded by, including pedestrians, cyclists, other vehicles, obstacles, and buildings. Data association is the process of relating sensor data to the vehicle's model of the world. The awareness system is faced with various challenges. In the following subsections, these challenges are explained.

### 1.3.1 State estimation

Autonomous vehicles can be equipped with a suite of sensors to collect comprehensive input for an awareness system. However, such data are not helpful without data association and a motion model [Wongthongtham et al. 2017]. The awareness system uses available measurements to estimate and predict the states (such as position and velocity) of the objects that are relevant to an approaching vehicle. The prediction of pedestrian motion has been addressed from various perspectives.

- Physics-based methods : In these models, motion is predicted by simulating a set of dynamics equations that follow a physics-based model [Baxter et al. 2014; Corbetta et al. 2018; Ess et al. 2010; Kooij et al. 2019]. In these methods, different filtering algorithms are proposed, such as Kalman filtering [Spincemaille et al. 2008], extended Kalman filtering [Hsu et al. 2017] and complementary filtering [Abbasi-Kesbi and Nikfarjam 2018].

- Pattern-based methods : These approaches can learn statistical behavioral patterns in the observed motion trajectories [Bruneliere et al. 2019; Jeung et al. 2007; Laursen et al. 2012; Mathew et al. 2012; Nielsen et al. 2013]. Hidden Markov Models (HMM)[Vasquez et al. 2009], support vector machines (SVM)[Bilal 2017], deep learning methods [Chen, Zhao, et al. 2020] such as convolutional neural networks (CNN) [Pfeiffer et al. 2018] are some of the approaches that are widely used in pattern-based methods.

In this work tracking defines as the process of fusing input data, estimating states, and making associations between them. Despite the availability of many models and trackers, the following issues complicate state estimation and multiple pedestrians tracking:

- Pedestrians can change their position, posture, and direction instantly and at any time. Besides, the time between two consecutive measurements of the same pedestrian can vary. Therefore, there is a probability that a motion model could not estimate and update the state sufficiently well.

- Pedestrians are often associated with self-occlusion. They may have pets or accessories such as backpacks, hats, suitcases, and walking assistance devices. Moreover, other objects such as cars, cyclists, infrastructures, and other pedestrians can partially or entirely occlude pedestrians.

During the occlusion, the sensory information is typically incomplete and the model used may be unable to keep track of the pedestrians.

- The input information may contain false positive and false negative. It means sensors detect an object which is similar to pedestrians in shape, position, and structure or sensors do not detect present pedestrians.
- Changing environmental conditions can disturb sensing such as moving backgrounds, changing weather conditions, and lighting variations. As a result, drifts and poor detections may cause a failure in estimating the pedestrian states.
- A tracker considers a state vector for each pedestrian. Therefore, the size of the state-space will increase with the number of pedestrians.
- Most of the time, the input information is unlabeled, which means that it is not clear which information belongs to which pedestrian. Using unlabeled data without or with incorrect data association can lead to failure of a tracker.
- The number of pedestrians to track is unknown, and based on the situation, the number can vary.

These issues indicate that a motion model of a pedestrian-aware system should deal with different challenges. Furthermore, a motion model system should link measured properties to pedestrians and estimate their state over time. As a result of the mentioned issues, inputs are not always valid [Wasik et al. 2020]. To deal with these challenges, various directions available for multiple pedestrian tracking and human motion patterns are analyzed. From stick figures and single point trackers to more complicated models, a wide variety of models are presented in the literature [Führ and Jung 2014; Iqbal et al. 2017; Sidenbladh et al. 2000; Sundaresan et al. 2004; Waddell and Amazeen 2017]. One of the structures to track multiple pedestrians is using multiple joints of the body. As a result of occlusions or false joint detections, tracking joints of multiple pedestrians is not a trivial task. To make a relation between multiple joints, researchers usually use linear models. There is a high probability that linear models are unable to estimate and predict the position of joints during occlusions.

In Chapter 2, this thesis proposes non-linear models based on human anatomy to find a relation between each joint of pedestrians and track them related to an approaching vehicle. To achieve this goal, the proposed tracker takes advantage of human kinematic constraints and Fourier series approximations. In the proposed motion model, different sources of information and multiple joints of the human body are used to cover the mentioned issues and improve the accuracy of a pedestrian-aware system compared to the state-of-the-art.

### 1.3.2 Data association

Another challenge of an awareness system is that it must accommodate the uncertainty inherent to sensor data, vehicle state, and motion models. A data association algorithm should confirm or refuse a track within a short time frame (in the order of milliseconds). Therefore, different data association algorithms are applied to reduce this challenge and maintain them over time [He, Luo, et al. 2019]. The algorithms include the Multiple Hypothesis Tracker (MHT) [Bhuvaneswari and Subashini 2014], the Joint Probabilistic Data Association Filter (JPDAF) [Bar-Shalom, Willett, et al. 2011], the probabilistic multiple hypothesis tracker (PMHT)[Streit and Luginbuhl 1995], and probability hypothesis density (PHD) [Mahler 2007]. One approach to solve the data association problem is using the multiple data association hypotheses [Bar-Shalom, Daum, et al. 2009; Blackman 2004; Rasmussen and Hager 1998]. In a hypotheses-based approach, data association decisions can be deferred until uncertainties on data association are resolved. Therefore, to solve the data association problem, a hypotheses-based approach is used in this thesis.

The first step in a hypotheses-based data association is initialization and its primary aim is to provide a guess to decide whether a new filter must be created. Quick data association can allow any tracker to dive into tracking with a lower error, fewer false positives, and a minimal time delay between the first detection and the first track. Besides, the initialization of a data association method can improve the performance of a pedestrian-aware system. A typical approach to initialize a hypothesis-based data association algorithm is waiting to collect a fixed number of measurements. The fixed number of measurements are usually obtained by an approximate algorithm such as the Lagrangian relaxation approach [Deb et al. 1997], the m-best assignment [Blackman and Popoli 1999], linear programming [Areta et al. 2006], and Murty's ranking algorithm [Murthy 1968]. Postponing initialization too long may lead to late or worse response. Simultaneously, initializing a data association algorithm after a single measurement increases the risk of introducing false positives, which could lead to an uncomfortable driving experience. In Chapter 3, this thesis offers a step to find a trade-off between the parameters in a probabilistic model and initialize a hypothesis tree.

### *1.3.3 Environment description*

The awareness system must be adapted to the environment, including objects such as highways, urban areas, the parking lots. To do this, automated vehicles widely rely on on-board sensors to perceive the environment in an urban area. On-board sensors can include light detection and ranging (LIDAR), different types of cameras, and radio detection and ranging (RADAR) [Asvadi et al. 2018; Banerjee et al. 2018; Zhao, Sun, et al. 2020]. The utilization of the these sensors have become more and more popular in intelligent transportation systems (ITS). However, as a result of low-illumination conditions and being positioned in blind spots of a vehicle, on-board sensors are unable to provide the required information all the time [Rawashdeh and Wang 2018a]. Based on a research in 2019 regarding traffic safety [Reed et al. 2019], more than 70% of pedestrian car crashes are due to the low-visibility of pedestrians.

A typical approach to increasing the performance of an awareness system during low-visible situations is using multiple sensors simultaneously. Therefore, several sensor fusion methods were published with the aim of better pedestrian-aware systems. Camera has been combined with LIDAR [Schlosser et al. 2016]. A fused system of camera and RADAR is introduced in [Streubel and Yang 2016]. LIDAR, camera, and RADAR were fused in [Chavez-Garcia and Aycard 2015] to detect moving objects. In [Kwon, Hyun, et al. 2017], a fusion of LIDAR and RADAR was used to detect pedestrians in occlusion. However, multiple sensors cannot always solve occlusion, partial detection, false detection, and environmental variation. This means that even by applying several sensors, there is a high probability that a pedestrian-aware system would be unaware of the pedestrians in the blind spots. Therefore, the recognition is not a trivial task.

Another challenge in environment description is an understanding of pedestrians' behaviors. Pedestrians strongly influence each other's behavior [Yi et al. 2015], which means that unpredictable behaviors increase when a tracker faces multiple pedestrians. Therefore, non-verbal communication with pedestrians can improve environment description' performance. Predicting intention of crossing the road can be used to predict human behavior and may help to have safe and comfortable driving [Fang and López 2018]. Most of the existing approaches tackle pedestrian intention prediction using trajectories or poses [Bai et al. 2015; Muscholl et al. 2020; Quan et al. 2021; Saleh et al. 2019b]. They do not offer a deeper interpretation of a pedestrian's action or how intention influences a pedestrian's action to cross in the near future.

In Chapter 4 and 5, external information are used to overcome these critical issues of existing pedestrian awareness and environment definition. External information include status of traffic lights, definition of traffic signs, and collecting data using the Internet-of-Things (IoT) [Privat 2012; Soldatos et al. 2015; Vermesan, Friess, et al. 2014]. The advantage is that all of the information from different sources detect details from the same scene. By fusing the strengths of all available inputs, an accurate and robust awareness system can be achieved. Therefore, the external information and on-board sensors potentially add detection robustness and lead to new ways of designing automated vehicles [Kwon, Park, et al. 2018]. Moreover, Chapter 5 proposes a framework to recognize the action and predict intention. This work shows that combining action, distance, external information, and interaction with elements in the scene can improve prediction results.

## 1.4 Structure of the thesis

The present thesis is divided into four contributions, where each chapter elaborates on one of the contributions of this research. Each of the chapters can be read without reading the prior chapters. The chapters follow from published manuscripts.

- Chapter 2 is based on

  Dolatabadi, M., Elfring, J., van de Molengraft, R. (2020). Multiple-joint pedestrian tracking using periodic models. Sensors, 20(23), 6917

- Chapter 3 is based on

  Dolatabadi, M., Elfring, J., van de Molengraft, R. (2021). Improved Data Association of Hypothesis-Based Trackers Using Fast and Robust Object Initialization. Sensors, 21(9), 3146.

- Chapter 4 is based on

  This chapter is based on:
  Dolatabadi, M., Elfring, J., van de Molengraft, R. Multiple pedestrian tracking using vision-based sensors and IoT technology. Internet of Things. Submitted

- Chapter 5 is based on

  Dolatabadi, M., Elfring, J., Aboutalebian.B, van de Molengraft, R. (2021). Intention Prediction and Action Recognition of Pedestrians Using Body Features and Contextual

Information In Automotive Applications . Robotics and Autonomous Systems, Submitted.

Chapter 6 restates the most important conclusions and the answer of the core research question of the thesis. Recommendation for the future work are outlined in Chapter 6. The following subsections summarise the key contribution per chapter.

### 1.4.1 Chapter 2: Multiple joints pedestrian tracker

**Contribution 1**. *Propose a motion model to track multiple joints of pedestrians. This model considers human kinematic constraints, gait models, and hypothesis-based data association to track multiple pedestrians during partial occlusions.*

The scenes involved in autonomous driving scenarios in an urban area rarely feature a single individual pedestrian. Most commonly, multiple pedestrians must be tracked concurrently, some of which may be in motion relative to the vehicle and each other. Therefore, it can be complicated to perform robust tracking of multiple pedestrians without using a accurate motion model. This research aims to answer this research question: *How can a motion model improve the accuracy and precision of a pedestrian tracker during occlusions?*

### 1.4.2 Chapter 3: Data association improvement

**Contribution 2**. *Finding a trade-off between the parameters to initialize a probabilistic data association model.*

This research aims to answer the following research question : *How can a hypothesis tree be initialized faster compared to the ones existing in the current state-of-the-art?* In order to answer this question, this thesis proposes a framework to initialize a hypothesis-based data association. This framework finds a trade-off between the parameters in a probabilistic model. Using the trade-off, the probability of choosing the correct hypothesis increases. Therefore, after entering a new place without prior knowledge, the data association can increase the accuracy of a pedestrian-aware system.

### 1.4.3 Chapter 4: Vision-IOT based tracker

**Contribution 3**. *Fuse vision and IoT data to track pedestrians during occlusion and before a camera detection.*

Occlusions frequently occur when pedestrians walk past each other or are not in the sensors' field on view. In these situations, a detector cannot detect pedestrians continuously. To deal with occlusion, this thesis uses internal and external sources of information. This research aims to answer the research question: *How can internal and external sources of information be combined to add tracking robustness?*

To answer this question, a hypothesis-based pedestrian tracker is developed to fuse both internal and external sources. Each source provides different types of measurements. Therefore, to deal with data fusion, this tracker considers different attributes. Besides, this tracker uses external data to improve data association.

### 1.4.4 Chapter 5: Pedestrian behavior predictor

**Contribution 4**. *Propose an approach to predict the intention and recognize the action of pedestrians.*

Accurate prediction of pedestrians crossing a road that is shared with cars can significantly improve traffic safety. This research aims to answer the research question : *How can a pedestrian-aware system predict the relevant behavior of pedestrians?*

This thesis uses a body feature and a learning approach to recognize the probabilities of the current action of a pedestrian. Then, it shows that pedestrians' current and previous actions can affect their final decision to cross the street. Moreover, this thesis shows that using a combination of multi-class actions (walking, starting, standing, stopping) and context information (such as traffic signs, gazing) can improve intention prediction.

# Chapter 2

# Multiple-Joint Pedestrian Tracking Using Periodic Models

**Abstract:** Estimating accurate positions of multiple pedestrians is a critical task in robotics and autonomous cars. This chapter proposes a tracker based on typical human motion patterns to track multiple pedestrians. This work assumes that the legs' reflection and extension angles are approximately changing periodically during human motion. A Fourier series is fitted in order to describe the moving, such as describing the position and velocity of the hip, knee, and ankle. The tracker receives the position of the ankle, knee, and hip as measurements. As a proof of concept, this chapter compares the tracker with state-of-the-art methods. The proposed models have been validated by experimental data, the Human Gait Database (HuGaDB), and the Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) tracking benchmark. The results indicate that the tracker is able to estimate the reflection and extension angles with a precision of 90.97%. Moreover, the comparison shows that the tracking precision increases up to 1.3% with the proposed tracker when compared to a constant velocity based tracker.

This chapter is based on:
Dolatabadi, M., Elfring, J., van de Molengraft, R. (2020). Multiple-joint pedestrian tracking using periodic models. Sensors, 20(23), 6917

## 2.1 Introduction

Pedestrian deaths account for more than one-fifth of road traffic deaths around the world [Organization et al. 2018]. Therefore, transportation systems, including vehicles and infrastructures, use various approaches to track pedestrians, due to the high number of fatalities. The tracking here is defined as estimating a pedestrian's position and velocity. A tracker helps cars to plan their driving path and navigate safely. For example, suppose that a pedestrian is walking and does not notice a car near him/her. A tracker estimates the position and the velocity of the pedestrian. Subsequently, based on the tracker's output, the car can alert the pedestrian or change its speed or path.

Therefore, tracking pedestrians is one of the critical tasks in robotics, non-autonomous, and autonomous cars. A tracker faces challenges, such as occlusion, noisy measurements, and a limited field of view. Moreover, tracking is not a trivial task when a tracker faces multiple pedestrians. Although some research has focused on this topic [Bao et al. 2017; Dimitrievski et al. 2019], tracking multiple pedestrians is still a challenge in urban areas [Nguyen et al. 2019].

A tracker must estimate the position and velocity of a pedestrian. To do this, trackers utilize a measurement model and a process model. A measurement model describes the relation between the pedestrian position and velocity that are estimated by the tracker and joint position measurements that are received from sensors. A process model describes how the pedestrian position and velocity are assumed to change over time. Earlier trackers [Fang, Vázquez, et al. 2017; Ho et al. 2016; Liu and Wu 2017; Yang, Lu, et al. 2013] use the linear process or measurement models. The measurement model and process model are usually nonlinear in nonideal situations due to the occlusion, noisy measurements, and human moving patterns. Therefore, linear models should make assumptions, such as the routes are linear, pedestrians have linear movements, or pedestrians have movements with simple variations of direction [Chau et al. 2013]. These assumptions have negative consequences on tracking, and there is a probability that trackers with linear measurement models or process models are prone to fail during the tracking. Therefore, they are not always sufficient for tracking multiple pedestrians in an urban area [Zhuang et al. 2014].

The current state-of-the-art algorithms that track multiple pedestrians can be roughly divided into combined detection-tracking algorithms and tracking-by-detection paradigms. In the combined detection-tracking algorithm, the typical approach in the literature is to use deep learning algorithms in order to track pedestrians while detecting [Feichtenhofer et al. 2017; Ren et al. 2015; Zhou et al. 2020]. Although these kinds of trackers can match pedestrians anywhere in their sensors' field of view, they likely produce more false positives [Bao et al. 2017]. Moreover, as a result of pedestrians' nonlinear kinematic, this approach requires large datasets in practice. Because training on smaller datasets might lead to inaccurate tracks [Ghori et al. 2018].

In the tracking-by-detection paradigm, there is an assumption that the detections are provided independently of a tracker. It means that the tracking-by-detection paradigm can draw a sharp distinction between the detection and tracking of pedestrians. Therefore, trackers of this paradigm can work with any detector. In this paradigm, after receiving detections, most of the trackers first define a bounding box (BB) around the pedestrian and localize the BB in a frame. The tracker associates the center of BB to pedestrians who were previously tracked [Zhou et al. 2020]. As a result of the association, they can identify new pedestrians [Feng et al. 2019]. It means that detections that cannot be associated with tracked pedestrians can represent false detections or newly appeared pedestrians.

However, tracking a single point in a pedestrian's body may produce more false-positives than a multiple point tracker due to noisy measurement and occlusion [Xie et al. 2012]. Tracking multiple joints of a body can offer a more attractive alternative than tracking a single position of each pedestrian. Suppose that a tracker receives several joints in a BB that overlap with each other. Subsequently, one pedestrian is tracked and the other joints are considered as a new pedestrian or false positives. The pre-requisite for these approaches is the ability to detect multiple joints from the sensor data.

A goal of this chapter is to track multiple pedestrians surrounding a car, even when there are occlusions. Therefore, this chapter proposes a pedestrian tracker that tracks pedestrians while using multiple joints instead of a single point. The tracker belongs to the tracking-by-detection paradigm and the main goal is improving the measurement model and process model of a pedestrian tracker. In this tracker, a camera will be used for detecting pedestrians since cameras are typically available in automated vehicles. The tracker should satisfy the following requirements:

- Require an algorithm to associate noisy measurements with the position and velocity of pedestrians.
- Contain models to predict and describe the movements of each pedestrian.
- The tracker should use images that it receives from a camera.
- The tracker should estimate the position of a pedestrian at a joint level.

The tracker comprises a process model and a measurement model. The process model defines how the state vector is expected to change over time. The measurement model describes how to make a connection between the state vector and detected joints. For each pedestrian, the measurement vector is the positions of joints in pixel coordinates. The contributions of this work are as follows:

- This chapter proposes a pedestrian tracker that can track multiple joints of pedestrians. This work considers human kinematic constraints and a physical model to make a relation between joints. In the process model, time-varying Fourier series approximations and constant velocity assumptions are utilized.

- the state vector includes the position, the hip velocity of pedestrians, reflection, and extension angles between hip-knee and knee-ankle of each leg, and a pedestrian's step frequency.
- This chapter validates the tracker's performance by evaluating it on experimental data, one gait dataset, and one tracking benchmark.

The rest of this Chapter is arranged, as follows: 3.2 discusses related work. Section 3.3 describes the general framework of the proposed tracker. Section 3.4 introduces the proposed models. Section 3.5 describes how the issue of data association is handled. Section 3.6 contains the evaluation procedure, and, Section 2.7 validates the tracker. Section 2.8 presents conclusions and outlines future directions in this research.

## 2.2 Related work

The first group of related works represents pedestrians by a single point. In order to track pedestrians, [Nguyen et al. 2019] tracks a single point in the center of the body. In [Bajracharya et al. 2009], based on the detection, the authors define the BB around each pedestrian in each image. Subsequently, they track the center of the BBs and estimate the position and velocity of the pedestrian. In [He, Zhang, et al. 2016], the researchers track each pedestrian as a point. In [Linder and Arras 2014], the authors address the problem of detecting and tracking groups of people in RGB-D data. They consider each group to be a point. Therefore, they do not track each person individually.

Pedestrians can continuously change their position and direction. Therefore, the position of the BB varies with time. In a crowded area, there is a probability that a single point is occluding another point during tracking. Therefore, a tracker cannot receive any measurement regarding the occluded point [Masoud and Papanikolopoulos 2001]. Having more details about the detected pedestrian can decrease the false tracks. Pedestrians can be represented by more complicated models, including multiple joints, as an alternative to single-point trackers.

In [Moon et al. 2016], they develop a human skeleton tracking system. They use a constant velocity KF in order to track the positions of body joints. [Troje 2002] computes the displacements of 15 joints in the body relative to each other. They define the position of people based on the displacements of their joints. [Steinbring et al. 2016] proposes a real-time method for tracking a pedestrian's entire body and motion using unlabeled marker measurements. They track each joint based on the sensor attached to the body. To track, they use their measurement in a Kalman filter (KF). All of these trackers use a linear process model and motion model to track a human skeleton. A linear tracker can be used in a static camera [Swalaganata, Affriyenni, et al. 2018]. Moreover, the linear models may produce more errors in their estimations when compared to the nonlinear model [Wieser et al. 2004].

Among the various studies that track the pedestrian's entire body, researchers have tracked them based on specific parts of the body. [Kong et al. 2013] develops an eight joint skeleton model in order to track a person in a given video. They track each joint individually while using a KF. They assume that all joints move independently. Therefore, they define no relation between joints. One side effect of this assumption is that there is a probability that they use the joints of other pedestrians during occlusion. In [Zhao and Shibasaki 2005], the authors propose a system for tracking both feet of pedestrians as they walked, based on multiple single-row laser scanners. In a crowded area, it is challenging for a detector to detect the feet. Therefore, a tracker requires more information regarding a pedestrian.

Several researchers have assessed the kinematic coupling between the hip and knee and ankle of a person walking in recent years [Baghdadi et al. 2018; Bennett et al. 2013; Nwaizu et al. 2016]. [Fod et al. 2002] models the pedestrian leg as a pendulum with an EKF in order to estimate the displacement of a pedestrian. They attach two sensors on the right leg of a pedestrian to extract accelerations. [Baghdadi et al. 2018] considers the periodic nature of walking, and they modify a bio-mechanical model with a first-order Taylor series expansion. Their state vector contains the angular position of the trunk relative to the vertical axis in a 2D plane, the angular position of the ankle relative to the hip joint, linear acceleration of the hip and the ankle. Using an IMU that was attached to the ankle joint, they measure the acceleration of the ankle and the angular position. Based on the measurements that were received from sensors, they calculate the coefficients of the Taylor series. Their process model has constant coefficients, whereas the coefficients should be varied based on age, weight, height, and gender. In [Nwaizu et al. 2016], they use an accelerometer to measure movement angle, velocity, acceleration ,and displacement of knees.

This chapter proposes a tracker that can be used for each age, weight, and gender. This tracker uses pedestrians' legs because of the simplicity of the shape instead of the whole body. Legs are detectable, even from a low-resolution camera [Fod et al. 2002]. Moreover, this chapter focuses on tracking the position of six joints of pedestrians as they walk. The joints that will be used throughout this work are at the ankle, knee, and hip. Figure 2.1 shows those body joints.

This tracker uses the Fourier series and EKF in the process model. With the Fourier series approximation, this chapter computes the angles between each of the detected joints. On the measurement model, a two-link pendulum is utilized to make a relation between the joints. Moreover,in this work, a state vector are used that facilitates using this process model and measurement model. At the same time, the state vector can be updated while using the measurements that are just explained. More details will be given in the following sections.

## 2.3 Pedestrian tracker

This section introduces the proposed pedestrian tracker. Figure 2.2 shows its conceptual composition. The joint measurements is input to the pedestrian tracker. There are libraries to extract joints [Cao et al. 2017a; Fang, Xie, et al. 2017; Geiger, Lenz, and Urtasun 2012; Goodfellow 2016]; one of the most popular ones is OpenPose [Cao et al. 2017a]. In this work, OpenPose is utilized in order to detect the joints. OpenPose provides a position vector for each joint in pixel coordinates.

In this work, after receiving the data about the joints, a pixel-to-Cartesian coordinate frame transformation is implemented. To perform this transformation, the tracker requires knowledge of the camera's orientation with respect to a pedestrian's joints. Each joint has a frame with an x parallel to the ground and y pointing upwards. Moreover,information such as the camera's focal length and each joint position in the pixel coordinates are required. Based on the Dutch population, this works assumes an average height of 177 cm for pedestrians. Having no depth information regarding a pedestrian was the only reason to make this assumption. Afterwards, this chapter solve a backward perspective projection model equation [Riley 2006]. The length of a pedestrian leg is computed when the tracker receives the positions of a pedestrian's joints in the Cartesian coordinate frame

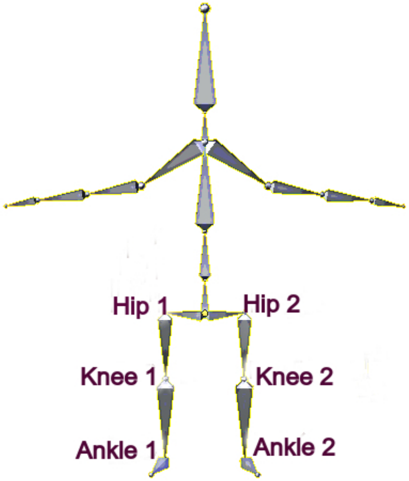


Figure 2.1: The joints of interest to detect and use in this tracker are at the ankles, knees, and hips.

for the first time. Subsequently, there is an assumption that this length is constant and equal for the two legs of a pedestrian.

In the data association block, a multiple-hypothesis tree is used, as implemented in [Elfring et al. 2013], to match each leg of a detected pedestrian with pedestrians that the tracker is already tracking. This chapter used an EKF in order to track and predict the position of the joints of pedestrians based on detections of individual joints and nonlinear models. The EKF comprises a measurement and a process model.

As mentioned, this tracker should track a pedestrian, even if a detector does not detect a joint. For example, it should estimate the ankle's position based on the hip where a detector cannot detect the ankle, but it detects the hip. To meet the requirement, this chapter uses a two-link pendulum to define each joint's position with respect to the other joints. To do this, this tracker requires angles between joints. $\theta_{H_1}$ and $\theta_{K_1}$, represent the hip and knee flexion and extension angles in the right leg.

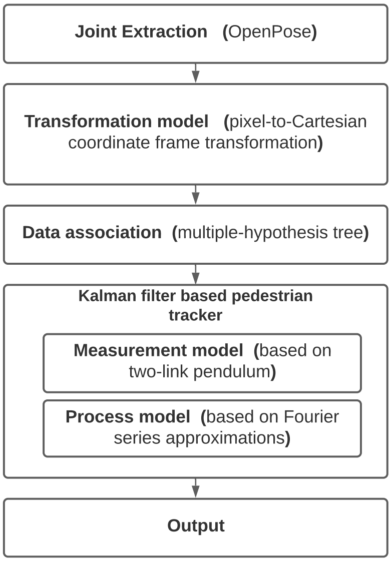

Joint Extraction (OpenPose)
Transformation model (pixel-to-Cartesian coordinate frame transformation)
Data association (multiple-hypothesis tree)
Kalman filter based pedestrian tracker
Measurement model (based on two-link pendulum)
Process model (based on Fourier series approximations)
Output

Figure 2.2: General framework of a tracker.

Figure 2.3 shows these angles. The angles $\theta_{H_2}$ and $\theta_{K_2}$ have the same definition in the left leg. Based on the output of the blocks, the tracker delivers each pedestrian's hip, knee, and ankle position, the velocity of the hip with respect to the camera, the angles of joints, and the step frequency of the pedestrian.

## 2.4 Proposed models

In the process model, the periodic nature of walking and the constant velocity model are used to describe how the state changes over time. This chapter assumes that, during walking, the hip, knee, and ankle lie on a 2D plane. In the measurement model, this chapter exploits the relations shown in Figure 2.3.

### 2.4.1 Process model

To define the process model, the assumptions are as follows:

1. In gait analysis, walking is assumed to be periodic [Elfring et al. 2013].
2. In between two frames, we assume that the frequency of the angles is constant.
3. There is a linear relation between walking velocity and

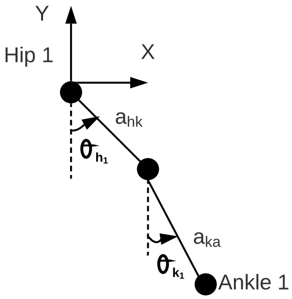


Figure 2.3: The right leg from the side view in a schematic way. $a_{hk}$ corresponds to a length between the hip and the knee. $a_{ka}$ is a length between the knee and ankle. Both of the angles are defined as positive in counterclockwise direction.

frequency.

4. Both of the legs move with the same frequency during one continuous walking.
5. The hip velocity of a pedestrian in the Y direction is zero.
6. The two joints of the hip have the same linear velocity in the X direction.
7. In each leg, the frequency of the angles is equal. It means that the rate of completing a stride is equal in the joints of a leg.

Based on the assumptions, each angle could be modeled as a periodic signal. The Fourier series can approximate such a periodic function as a function of time. Hence, it is possible to use a Fourier series to propagate each angle [Kurz and Stergiou 2007]. Additionally, based on the assumptions, a frequency-velocity model is used to estimate the motion of pedestrians. The process model has been structured to be represented while using the following equation:

$$x(t) = f(x(t-1)) + w(t), \qquad w(t) \sim N(0,Q) \tag{2.1}$$

where $x$ is a state vector, $f$ is a non-linear state transition function that computes the predicted state from the previous estimate, and $w$ is process noise. This chapter assumes that it is zero-mean white noise with a known covariance matrix. $Q$ is the covariance matrix and it is constant, because the upper value of $Q$ can obtain an acceptable estimating precision [Wang, Deng, et al. 2017].The state vector for each pedestrian is defined as:

$$\begin{aligned} x(t) = [&X_{h_1}(t), Y_{h_1}(t), V_{x_h}(t), SF(t), \theta_{H_1}(t), \omega_{H_1}(t), \theta_{K_1}(t), \\ &\omega_{K_1}(t), X_{h_2}(t), Y_{h_2}(t), \theta_{H_2}(t), \omega_{H_2}(t), \theta_{K_2}(t), \omega_{K_2}(t)]^T \end{aligned}$$

where:

- $X_{h_1}$ and $Y_{h_1}$ are the hip position of the right leg in two directions at time $t$ with respect to the measurement sensor.
- $V_{x_h}(t)$ is the linear velocity of the hip at time $t$.

- $SF(t)$ is the frequency of the joints at time $t$.
- $\omega_{H_1}(t)$ is a time derivative of $\theta_{H_1}(t)$ and $\omega_{K_1}(t)$ is a time derivative of $\theta_{K_1}(t)$ in the right leg.
- $X_{h_2}$ and $Y_{h_2}$ are the hip position of the left leg in two directions at time $t$ with respect to the measurement sensor.
- $\omega_{H_2}(t)$ is a time derivative of $\theta_{H_2}(t)$ and $\omega_{K_2}(t)$ is a time derivative of $\theta_{K_2}(t)$ in the left leg.

Based on the third assumption, the tracker can use a linear model in order to propagate the velocity of the hip joints. Based on [Tsang et al. 2019], a first-order Fourier series can cover hip, knee, and ankle position with an accuracy of 96% ,93%, and 89%. Therefore, this chapter utilizes the first order of the Fourier series. It means that the maximum amplitude of angles and the initial phase angles are constant. The non-linear state transition function for each state can be defined, as follows:

$$\begin{aligned}
X_{h_1}(t+1) &= V_{x_h}(t)dt + X_{h_1}(t)\\
Y_{h_1}(t+1) &= Y_{h_1}(t)\\
V_{x_h}(t+1) &= V_{x_h}(t)\\
SF(t+1) &= SF(t)\\
\theta_{H_1}(t+1) &= A_{H_1}\sin(SF(t+1)(t+1)+\phi_{H_1})\\
\omega_{H_1}(t+1) &= A_{H_1}SF(t+1)\cos(SF(t+1)(t+1)+\phi_{H_1})\\
\theta_{K_1}(t+1) &= A_{K_1}\sin(SF(t+1)(t+1)+\phi_{K_1})\\
\omega_{K_1}(t+1) &= A_{K_1}SF(t+1)\cos(SF(t+1)(t+1)+\phi_{K_1})\\
\\
X_{h_2}(t+1) &= V_{x_h}(t)dt + X_{h_2}(t)\\
Y_{h_2}(t+1) &= Y_{h_2}(t)\\
\theta_{H_2}(t+1) &= A_{H_2}\sin(SF(t+1)(t+1)+\phi_{H_2})\\
\omega_{H_2}(t+1) &= A_{H_2}SF(t+1)\cos(SF(t+1)(t+1)+\phi_{H_2})\\
\theta_{K_2}(t+1) &= A_{K_2}\sin(SF(t+1)+\phi_{K_2})\\
\omega_{K_2}(t+1) &= A_{K_2}SF(t+1)\cos(SF(t+1)(t+1)+\phi_{K_2})
\end{aligned} \tag{2.2}$$

where $dt$ is a time difference between discrete time steps $t$ and $(t+1)$. $\phi_{H_1}$,$\phi_{K_1}$,$\phi_{H_2}$, and $\phi_{K_2}$ are the initial phase angles of hip and knee in both legs . $A_{H_1}$,$A_{K_1}$,$A_{H_2}$, and $A_{K_2}$ are the maximum amplitude of angles in both legs. The maximum amplitudes and the initial phase angles are different in males and females [Bertram and Ruina 2001]. Therefore, this tracker estimates the angles and their rate independence of them. To do it, first, the cosine and sine expansion are used. According to the constant frequency assumption and the expansions, the $\omega_{H_1}$ and $\theta_{H_1}$ from (Equation (2.2)) are written, as follows:

$$\begin{aligned}
&\theta_{H_1}(t+1) = A_{H_1}[\sin(SF(t)t+\phi_{H_1})\cos(SF(t)) \\
&+\cos(SF(t)t+\phi_{H_1})\sin(SF(t))] \\
&= A_{H_1}(C1.C2) + A_{H_1}(C3.C4)
\end{aligned} \tag{2.3}$$

$$\begin{aligned}
C1 &= \sin(SF(t)t+\phi_{H_1}) \\
C2 &= \cos(SF(t)) \\
C3 &= \cos(SF(t)t+\phi_{H_1}) \\
C4 &= \sin(SF(t))
\end{aligned}$$

As can been seen, this chapter can make a relation between $C1$ and $\theta_{H_1}(t)$ and between $C3$ and $\omega_{H_1}(t)$. Therefore, we have:

$$\theta_{H_1}(t+1) = \theta_{H_1}(t)C2 + \frac{\omega_{H_1}(t)}{SF(t)}C4 \tag{2.4}$$

Similar to (2.4), the $\omega_{H_1}(t+1)$ is rewritten, as:

$$\omega_{H_1}(t+1) = \frac{d\theta_{H_1}}{dt} = \omega_{H_1}(t)tC2 - \theta_{H_1}(t)SF(t)C4 \tag{2.5}$$

(Equation (2.4)) is repeated and (Equation (2.5)) for the right knee and for the left leg.

### 2.4.2 Measurement model

The measurement model has been structured to be represented using the following equation:

$$z(t) = h(x(t)) + v(t), \qquad v(t) \sim N(0,R) \tag{2.6}$$

$h$ is used to compute the predicted measurement position from the predicted state. $v$ is measurement noise. This chapter assumes that it is zero-mean white noise with a known covariance matrix. $R$ is the covariance matrix of measurements.

The structure of the human lower limb acts as a kinetic chain during walking. Therefore, the position of the hip joint interacts with the knee and ankle position. This work uses homogeneous transformation matrices to transform the position of knee and ankle joints to the hip joint. The matrices are computed, as follows:

$$T_K^H = \begin{bmatrix} \cos(\theta_{H_1}(t)) & -\sin(\theta_{H_1}(t)) & a_{hk}\sin(\theta_{H_1}(t)) \\ \sin(\theta_{H_1}(t)) & \cos(\theta_{H_1}(t)) & -a_{hk}\cos(\theta_{H_1}(t)) \\ 0 & 0 & 1 \end{bmatrix} \tag{2.7}$$

$$T_A^K = \begin{bmatrix} \cos(\theta_{K_1}(t)) & -\sin(\theta_{K_1}(t)) & a_{ka}\sin(\theta_{K_1}(t)) \\ \sin(\theta_{K_1}(t)) & \cos(\theta_{K_1}(t)) & -a_{ka}\cos(\theta_{K_1}(t)) \\ 0 & 0 & 1 \end{bmatrix}$$

where $T_K^H$ is a transformation matrix of the knee position to the hip joint, and $T_A^K$ transforms the ankle joint to the knee joint. $a_{hk}$ corresponds to a length between the hip and the knee. $a_{ka}$ is a length between the knee and ankle. This chapter assumes these two lengths are equal for two legs. $T_A^H$ is a transformation of the ankle joint to the hip joint. $T_A^H$ is computed by multiplying $T_K^H$ and $T_A^K$.

Figure 2.3 illustrates the right leg from the side view in a schematic way. This part repeats the same matrices for the left leg and, then, based on the transformation matrices, the following equations are extracted, which are the joints' positions with respect to the camera frame.

$$
\begin{aligned}
X_{h_1}(t) &= X_{h_1}(t) + a_{hk}\sin(\theta_{H_1}(t)) \\
Y_{h_1}(t) &= Y_{h_1}(t) - a_{hk}\cos(\theta_{H_1}(t)) \\
X_{a_1}(t) &= X_{h_1}(t) + a_{hk}\sin(\theta_{H_1}(t)) \\
&\quad + a_{ka}\sin(\theta_{H_1}(t) + \theta_{K_1}(t)) \\
Y_{a_1}(t) &= Y_{h_1}(t) - a_{hk}\cos(\theta_{H_1}(t)) \\
&\quad - a_{ka}\cos(\theta_{H_1}(t) + \theta_{K_1}(t)) \\
X_{k_2}(t) &= X_{h_2}(t) + a_{hk}\sin(\theta_{H_2}(t)) \\
Y_{k_2}(t) &= Y_{h_2}(t) - a_{hk}\cos(\theta_{H_2}(t)) \\
X_{a_2}(t) &= Y_{h_2}(t) + a_{hk}\sin(\theta_{H_2}(t)) \\
&\quad + a_{ka}\sin(\theta_{L_1}(t) + \theta_{L_2}(t)) \\
Y_{a_2}(t) &= Y_{h_2}(t) - a_{hk}\cos(\theta_{H_2}(t)) \\
&\quad - a_{ka}\cos(\theta_{H_1}(t) + \theta_{K_2}(t)) \\
h(x(t)) &= [X_{h_1}, Y_{h_1}, X_{k_1}, Y_{k_1}, X_{a_1}, Y_{a_1}, \\
&\quad X_{h_2}, Y_{h_2}, X_{h_2}, Y_{h_2}, X_{h_2}, Y_{h_1}]^T
\end{aligned}
\tag{2.8}
$$

where $X_{k_1}$ and $Y_{k_1}$ are the knee positions and $X_{a_1}$ and $Y_{a_1}$ are the ankle position of the right leg, which are computed using the hip positions and a double pendulum model. $X_{h_2}$,$Y_{h_2}$,$X_{h_2}$, and $Y_{h_2}$ have the same definition for the left leg. $z$ is the measurement vector. This chapter can use a linear model to track the hip because the hip's angular displacement is insignificant. In contrast, the knee and ankle have angular displacement; therefore, this work utilizes angles to compute other joints' positions. For other joints, this chapter considers the effect of nonlinear motion.

## 2.5 Data association

This section describes how the tracker solves the association problem while using a multiple-hypothesis tree (MHT) [Murthy 1968]. Data association is the process of matching newly detected pedestrians with pedestrians that were already being tracked. Moreover, data association determines which of the detected legs is the right leg and which one is the left leg. To associate data, MHT generates a hypothesis tree with several branches.

Each measurement can be associated with an existing pedestrian, clutter, or a pedestrian that was not tracked before. Therefore, each branch is a collection with hypotheses. For each measurement, each branch can be formed with different possible associations. Every hypothesis contains a list of pedestrians and the estimation of their state vector. Hypotheses are considered in parallel. Therefore, data association decisions can be deferred until uncertainties on data association are resolved. The tree expands by receiving a new measurement at a time of $t+1$. The probability of each hypothesis is computed in order to pick the most probable hypothesis and keep the tree size bound.

## 2.6 Performance

In this work, Multiple Object Tracking Precision (MOTP) are utilized to have a clear and understandable evaluation [Geiger, Lenz, and Urtasun 2012]. MOTP quantifies the tracker's ability to determine a pedestrian's exact position.

$$MOTP = \frac{\sum_{i,t} d_t^i}{\sum_t c_t} \tag{2.9}$$

where

- $c_t$ is the total number of pedestrians; and,
- $d_t$ is the total position error for matched pedestrians.

To evaluate, the human gait dataset (HuGaDB) [Chereshnev and Kertész-Farkas 2017] and the Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) tracking benchmark are used [Geiger, Lenz, and Urtasun 2012]. HuGaDB collects data from a body sensor network of six wearable accelerometers that were located on the right and left legs. The KITTI benchmark consists of 21 training sequences and 29 test sequences. They collect data at 10 Hz with a camera mounted on a moving car in a city, residential area, campus, and road. Figure 2.4 shows one of the individual benchmarks of KITTI. An output of OpenPose in the KITTI benchmark is presented in

Figure 2.5. OpenPose links all joints that belong to the person and assigns them different colors. Figure 2.5 shows these links.

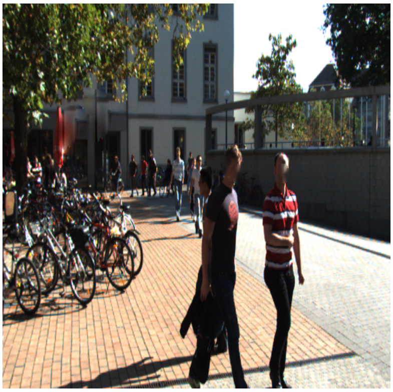


Figure 2.4: A campus pedestrians tracking.The training sequences number 17 of the Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) tracking benchmark.

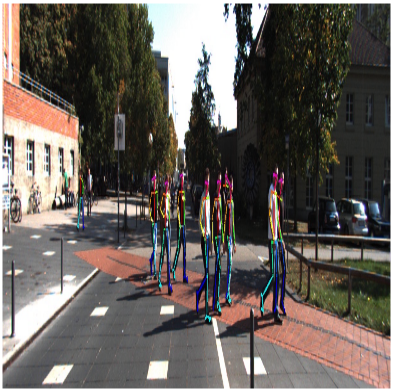


Figure 2.5: Result of OpenPose joints detection.The training sequences number 16 of the KITTI tracking benchmark.

## 2.7 Experimental evaluation

This section evaluates the tracker with the HuGaDB dataset, experimental data, and the KITTI tracking benchmark. The first part shows that the tracker can determine an acceptable MOTP.

The tracker is compared with another tracker that is used in [Moon et al. 2016]. For both trackers, the same data association and measurements are implemented. The last part compares the results with the state-of-the-art [Geiger, Lenz, and Urtasun 2012; He, Zhang, et al. 2016; Nguyen et al. 2019; Zhou et al. 2020].

### 2.7.1 HuGaDB dataset

Objective of this part is validating the models that are implemented in EKF. This chapter compare the result of $\theta_{H_1}$, $\theta_{K_1}$, $\omega_{H_1}$, $\omega_{K_1}$ in the tracker and the HuGaDB dataset for validation.

In HuGaDB, they placed six inertial sensors and electromyography (EMG) sensors on the right and the left pedestrians' thigh, shin, and foot. This dataset provides detailed gait data of the legs during walking and running [Chereshnev and Kertész-Farkas 2017]. The dataset contains the measurements that this work needs. Therefore, the ground truth (GT) data is available for all the joints while using the HuGaDB dataset. GT is calculated from the acceleration of the sensors attached to the leg [Chereshnev and Kertész-Farkas 2017]. The sensors send their output to this tracker. Therefore, Openpose is not uses to detect the joints of a pedestrian. Hence, this part can prove that the output of the state vector has a high MOTP.

Figure 2.6 draws comparisons for the right leg between the GT and the tracker. The ground truth angles come from two gyroscope sensors. The GT values of $\theta_{H_1}$ and $\theta_{K_1}$ were calculated as [Nwaizu et al. 2016]. The results in the part (a) and (b) of Figure 2.6 show that this participant completes thirteen cycles during his walking in 20 $(s)$. Each maximum peak shows a swing phase of his right leg, and the minimum peaks indicate the stance phases. Based on Figure 2.6, the tracker estimates a consistent angle pattern. This consistent pattern means that the first order of the Fourier series can cover the walking pattern. Figure 2.7 provides a visual representation of $\omega_{H_1}$ and $\omega_{K_1}$ in this tracker and the dataset.

In Figure 2.7a,b , the zero values indicate no angular movement at that time, and the peaks occur during stance and swing

Table 2.1: Evaluation metrics for tracking all 40 participates in HuGaDB dataset.

| Measurement | MOTP |
|---|---|
| Angle | 90.97% |
| Angular velocity | 84.53% |

phases.There are clear errors in Figure 2.7b. These errors may arise from facts, such as the sensor having a vibration, the attached sensors being mounted in slightly different positions, or the knee's process model should be different. The probability of the third fact is low, since, Figure 2.6b estimates the knee angle close to the GT.Unlike deep learning trackers, the tracker is explainable and can explain the variation of the state vector and the measurement.

Table 2.1 gives the MOTP of the tracker for both legs of all participants. The result of MOTP indicates that the tracker can compute the angles and their rate close to the ground truth data. Based on (4.1), MOTP is a function of the estimation total error. Therefore, for all of the participants, the mean error in angle is 3.6°.

### 2.7.2 Comparison

In order to compare the advantages of the process model and measurement model with another tracker, this chapter replaced them with models that are used in [Moon et al. 2016]. It means that the data association part of the two trackers is the same. In this scenario, a person was crossing a line at a constant speed for a given time. Then proceeded a curve to turn back to the starting point. The camera was fixed during this test, and the camera's distance to the joints and the crossing distance was known. Figure 3.12 shows an illustrative camera image with the detections that were used by both trackers.

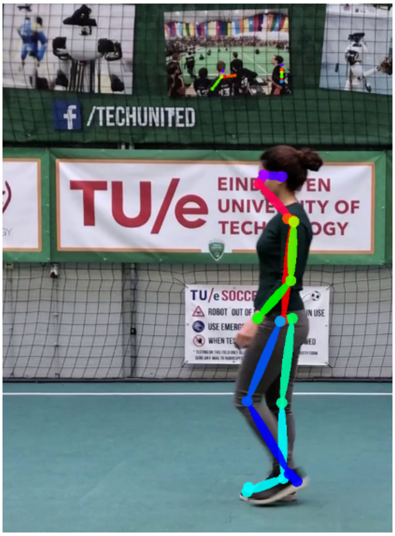


Figure 2.8: A pedestrian is crossing in front of a camera with constant velocity.

In [Moon et al. 2016], they use a constant velocity KF; their measurement vector contains the positions of each joint. Figure 2.9 compares GT with the two trackers. Figure 2.9 shows that the joints move roughly with a constant velocity during the swing phase, they are constant during the stance phase, and then they move again. Although [Moon et al. 2016] and the tracker used the same data association and measurements, there is a difference between GT and [Moon et al. 2016] during turning. This difference is due to the use of a linear measurement model and a process model in [Moon et al. 2016]. Similar to [Moon et al. 2016], this tracker also use a constant velocity model to compute the hip position. For other joints, we consider the effect of nonlinear motion.

For a quantitative comparison, Table 2.2 gives MOTP. Table 2.2 shows that the tracker estimates the positions of the hip with more precision since it has a higher MOTP than the tracker used in [Moon et al. 2016]. This work achieves a relatively high MOTP, because the pedestrian walks with both linear and nonlinear patterns. It indicates that the process model and measurement model help to improve pedestrian tracking.

Table 2.2: Evaluation metrics for a person tracking based on the sequence of images.

| Method | MOTP |
|---|---|
| This tracker | 98.60% |
| KF [Moon et al. 2016] | 97.37% |

### 2.7.3 KITTI dataset

As GT data, this dataset provides the center of a BB around the pedestrian in an urban environment. This chapter assumes that the center of a BB is equal to the center of the body. In order to compare the tracker with state-of-the-art, this part calculates the center of the pedestrian in relation to the hip joints. Therefore, this chapter only compares the center with the GT data. Table 2.3 compares the results of the tracker to the state-of-the-art algorithms for pedestrian tracking of the KITTI benchmark. The results of the algorithms presented in Table 2.3 are published on the KITTI website.

Table 2.3: Multiple Target tracking evaluation metrics for KITTI Pedestrian tracking benchmark.

| Method | MOTP |
|---|---|
| This tracker | $74.03 \pm 2.95\%$ |
| SRK-ODESA [Geiger, Lenz, and Urtasun 2012] | 75.07% |
| HWFD [Geiger, Lenz, and Urtasun 2012] | 74% |
| Quasi-Dense [Geiger, Lenz, and Urtasun 2012] | 73.99% |
| CenterTrack+MTFF [Geiger, Lenz, and Urtasun 2012] | 75.02% |
| TuSimple [He, Zhang, et al. 2016] | 71.93% |
| VVteam [Zhou et al. 2020] | 72.29% |
| MDP [Geiger, Lenz, and Urtasun 2012] | 70.36% |

There are multiple reasons for having a MOTP with a margin of

2.95%. The reasons are as follows:

- Lack of GT for the joints.
- OpenPose.

As mentioned before, this work assumed that all of the pedestrians have the same height. This assumption can produce an error. For example, when the sensors detect a pedestrian with 160 (cm) height and the tracker assumes an average height of 177 (cm), MOTP would be different from reality.

Moreover, it was assumed that the center of a BB is in the center of the body. This assumption affects all properties, such as size, location, orientation, and even pose of a pedestrian. For example, when a pedestrian has no symmetry pose from a detector point of view, this assumption can produce an error. This part revalidated the tracker with new heights in order to explore the effects of these assumptions on the tracker. Once, it is assumed that the average height is 150 cm. Subsequently, the average height is 190 cm. the benchmark was repeated with these two new values and computed MOTP. The results proved that the MOTP of KITTI is different based on the height value. These two assumptions can change the MOTP of KITTI for 2.95%. The variation of 2.95% in Table 2.3 shows that the comparison is not entirely fair. Other methods did not require the depth information at a joint level. Therefore, they only used the center of BBs and did not require estimating the height. Figure 2.10a,b show a situation that as result of occlusion OpenPose cannot detect the legs of pedestrians. After one frame, OpenPose detects all of them, as shown in Figure 2.10c. In these kinds of situations, that OpenPose does not perform well, this tracker can be negatively affected.

The pedestrian's height assumption is strong. It can be mitigated while using the information from the stereo cameras or point cloud data. Therefore, Velodyne point data were used to decrease the height assumption's effect. For one experiment in the testing part of KITTI, this chapter matched the Velodyne point cloud data's timestamps with the camera data. Therefore, the pedestrian's distance to the car was available. Subsequently, this work estimated the MOTP for that specific experiment. The result shows that this chapter can increase the MOTP up to 0.72% for that experiment. Other methods in Table 2.3 compute MOTP without stereo cameras or point cloud data. Therefore, it is not fair to compute MOTP while using these data.

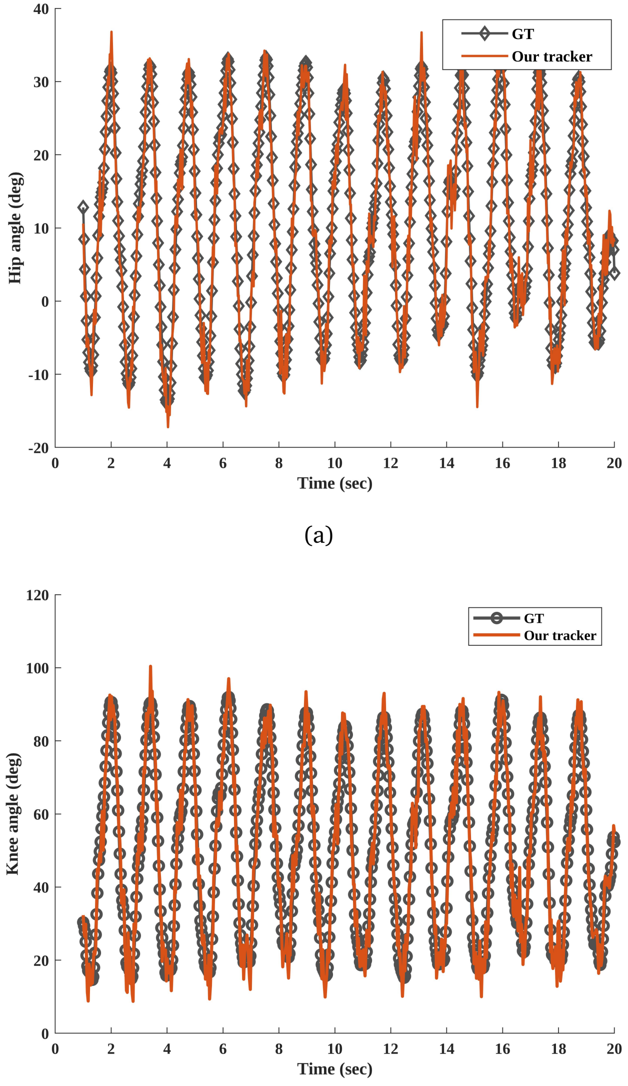
However, KITTI does not provide the position at multiple joints levels. The benchmark was recorded in crowded areas, and pedestrians often occlude each other. Therefore, this work uses the dataset in order to show the performance of the tracker in challenging situations that are representative for the application domain. Figure 2.11 shows the tracking results of two pedestrians, pedestrian 1 and pedestrian 25, in Figure 2.10c. The vertical axis in Figure 2.11 is the distance of the two pedestrians' left knees relative to the car, and the horizontal axis indicates time. The two pedestrians crossed the road in seven seconds. It should be noted that, as mentioned before, the goal of Figure 2.11 is indicating that the tracker can estimate a distance of a pedestrian continuously, even during occlusion. Therefore, the peaks in the figure do not mean swing or stance phases.

In Figure 2.11, there is a period that the tracker receives partial detection ($PD$) for the pedestrian with ID 1 and no detection ($ND$) for the pedestrian with ID 25. The tracker estimates the distance of the knee to the standing car during $PD$ and $ND$. Figure 2.11 shows that the tracker is able to track a pedestrian, even during an occlusion. The left knees are chosen, because, based on Figure 2.10, the left sides of these two pedestrians are not always visible. Therefore, estimating the position of the left knee was more difficult.

## 2.8 Conclusions

This chapter introduced a pedestrian tracker in order to track pedestrians' position as a two-link pendulum with an Extended Kalman Filter. The tracker is an explainable tracker, it receives skeleton data of each pedestrian. Subsequently, based on the human anatomy, this research models the relation between skeleton data. The tracker can track six different joints of each pedestrian. Tracking with multiple joints helps the tracker to achieve more information regarding a pedestrian. The evaluations show that this tracker can track pedestrians in urban areas during occlusion and turning.

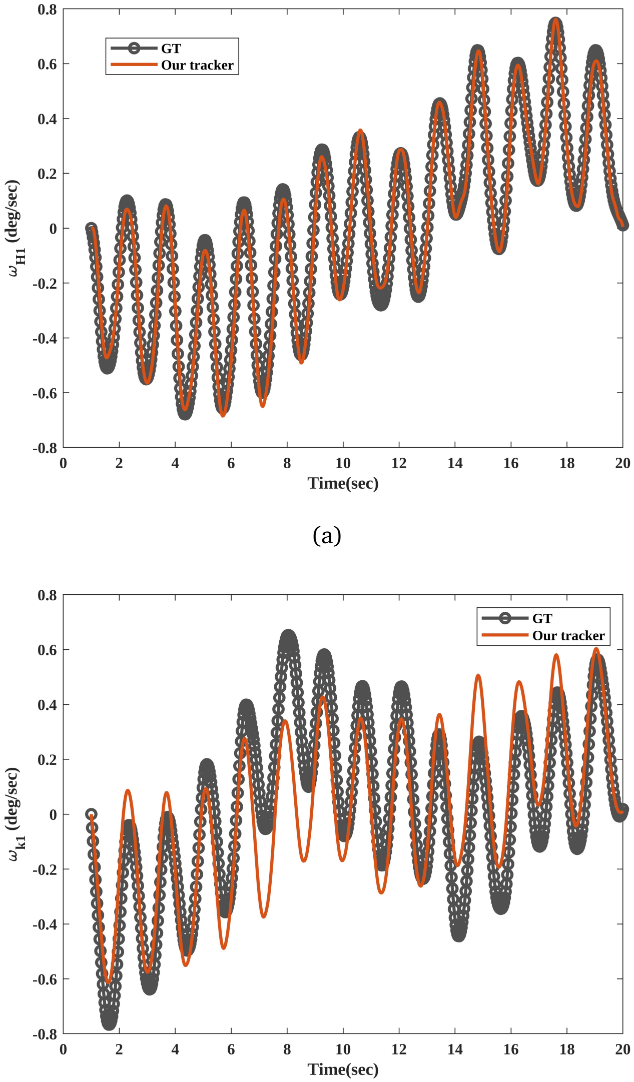
In future work, the proposed method will be extended to support joints along the entire body, such that partial occlusions are expected to be handled even better.



(a)

(b)

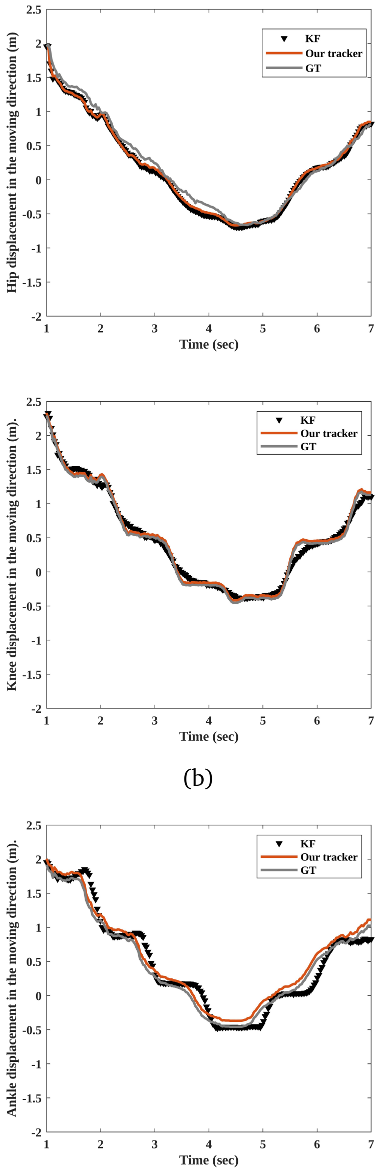
Figure 2.6: Measured and estimated results of this tracker for the right leg of one participant in HuGaDB dataset who was a 24-year old male with 177 cm stature and 75 kg body mass. (**a**) shows the angle between thigh and hip ($\theta_{H_1}$). (**b**) shows ($\theta_{K_1}$) measured using the accelerometers and estimated using this tracker.



(a)

(b)

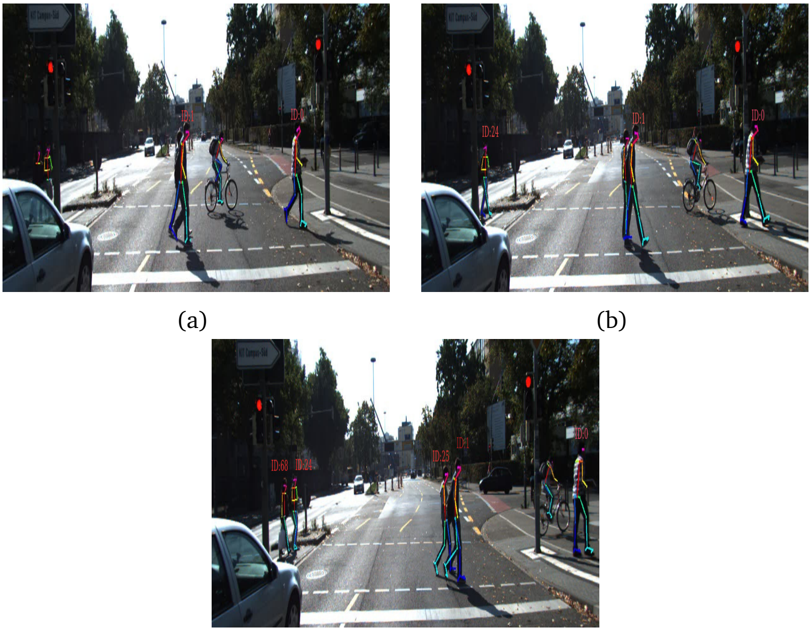
Figure 2.7: Measured and estimated results of the tracker for the right leg of one participant in HuGaDB dataset who was a 24-year old male with 177 cm stature and 75 kg body mass. (**a**) indicates $\omega_{H_1}$ and (**b**) compares the $\omega_{K_1}$ between the tracker and the dataset.



(b)

(c)

Figure 2.9: Measured and estimated results of the trackers for a person who was crossing in front of the camera. (**a**) shows the position of the hip joint in the left leg. (**b**) indicates the position of the knee joint in the left leg, and (**c**) compares the two trackers with each other based on the position of the ankle of the left leg.



(a) (b)

(c)

Figure 2.10: In (**a**), one pedestrian occludes another one, and the legs of two pedestrians are occluded by a car. In (**b**), one of the occluded pedestrians is in the field of view of the camera. In (**c**), OpenPose detects their joints. In situations such as (**a**) and (**b**), OpenPose can not detect pedestrians, affecting the results of Multiple Object Tracking Precision (MOTP).

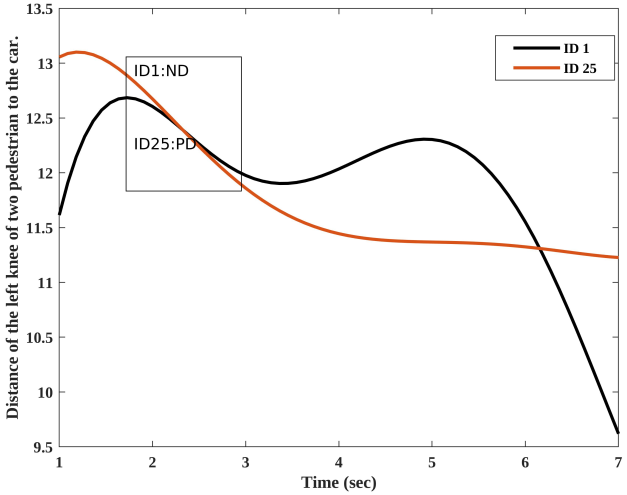


Figure 2.11: The result of the tracking two pedestrians in the sequences number 15 of the KITTI tracking benchmark. The vertical axes of the figure is the displacement of the pedestrians to a standing car. In the figure, *ND* means no detection and *PD* indicates partial detection.

# Chapter 3

# Improved Data Association of Hypothesis-Based Trackers Using Object Initialization

**Abstract:** The tracking of Vulnerable Road Users (VRU) is one of the vital tasks of autonomous cars. This includes estimating the positions and velocities of VRUs surrounding a car. To do this, VRU trackers must utilize measurements that are received from sensors. However, even the most accurate VRU trackers are affected by measurement noise, background clutter, and VRUs' interaction and occlusion. Such uncertainties can cause deviations in sensors' data association, thereby leading to dangerous situations and potentially even the failure of a tracker. The initialization of a data association depends on various parameters. This paper proposes steps to reveal the trade-offs between stochastic model parameters to improve data association's accuracy in autonomous cars. The proposed steps can reduce the number of false tracks; besides, it is independent of variations in measurement noise and the number of VRUs. Our initialization can reduce the lag between the first detection and initialization of the VRU trackers. As a proof of concept, the procedure is validated using experiments, simulation data, and the publicly available KITTI dataset. Moreover, we compared our initialization method with the most popular approaches that were found in the literature. The results showed that the tracking precision and accuracy increase to 3.6% with the proposed initialization

as compared to the state-of-the-art algorithms in tracking VRU.

This chapter is based on:
Dolatabadi, M., Elfring, J., van de Molengraft, R. (2021). Improved Data Association of Hypothesis-Based Trackers Using Fast and Robust Object Initialization. Sensors, 21(9), 3146.

## 3.1 Introduction

The possibility of driving autonomously through an urban environment has been a vision for many years. One of the many challenges of an autonomous vehicle is its safe operation through urban traffic. Therefore, the vehicle needs an accurate description of the environment. Although environment description is a broad topic in autonomous cars, this chapter focuses on one part of an environment descriptor. This work proposes a probabilistic step to initialize the data association of a hypothesis-based Vulnerable Road User (VRU) tracker.

This chapter considers a situation where multiple pedestrians are crossing the road to illustrate the effect of initialization of the data association. When pedestrians appear in front of a car for the first time, the following steps will happen:

- Records data from its surroundings.
- The detection algorithm detects the positions of pedestrians within data.
- The car estimates the pedestrians' position and velocity. Then it decides to decrease its speed or brake. At the same time, this detection could be a false positive, leading to unnecessary braking.

Several VRU trackers have been developed in recent years to track VRUs. Some recent studies on this topic are discussed in detail in [Althoff and Magdici 2016; Chou et al. 2020; Gindele et al. 2015; Mozaffari et al. 2020; Rudenko et al. 2020; Wu, Ruenz, et al. 2018; Yoon et al. 2021]. These studies track VRUs based on different learning methods and motion prediction models. Although they are useful for tracking VRUs, they have difficulties with the precision or accuracy of their tracks. The precision of their tracks indicates how well the 2D position and speed of a VRU are estimated. The accuracy of the tracks expresses how many mistakes the tracker made in terms of false positives, the number of tracked VRUs, or the number of VRUs with wrong IDs. The following facts can affect the precision and accuracy of a track:

- Noisy position measurements
- Occlusions

Noisy position measurement includes specification of detectors, environmental situations, such as weather and lighting conditions, and cluttered backgrounds, like trees, traffic signs, and buildings. Image analysis algorithms do not have the same result due to different illumination conditions during the daytime and at night. Occlusions mean objects occlude VRUs entirely or partially, which limits the process of monitoring VRUs. Tracking a VRU in the presence of noisy position measurements and occlusions is a non-trivial task. Therefore, a VRU tracker should satisfy the following requirements:

- Require an algorithm to associate noisy measurements with tracks.
- Contain a state estimation model to predict and describe the movements of each VRU.

The data association is a process of matching the measurements of VRUs with a tracker. The measurements can be about the position and the velocity of each VRU. Data association in an urban environment usually suffers from having multiple false alarms and clutters, such as measurement origin uncertainties, besides VRUs generated observations. Therefore, in this situation, data association is faced with many challenges. When data association confirms a track, a tracker's state estimator will continue to estimate its state vector. After confirming a track, a tracker can predict VRUs trajectories and maintain their identities, regardless of data association errors. It means that incorrect data association leads to potentially catastrophic results. The data association's initialization primary aim is to provide a guess to decide whether a new filter must be created. The data association should be initialized, confirm or refute a track in a short time in order of a millisecond. All of these facts make the data association complicated. Figure 3.1 shows the place of the initialization in a hypothesis-based tracker. Our VRU tracker receives detections as input (sensory data), as shown in Figure 3.1. It also shows a connection between data association and state estimator. It means that, based on a state estimator, the hypotheses are created or updated. The following sections describe that how this chapter estimates and validates the state vector.

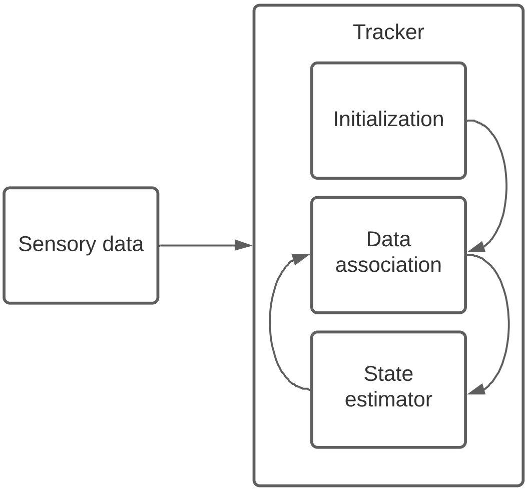


Figure 3.1: A framework to represent the place of the data association's initialization.

The association of the measurements and process models can underlie the hypothesis tree [Cox and Leonard 1994]. This paper incorporates multiple-hypothesis tracking (MHT) with a process model to fulfill the requirements. Solving data association is a pre-requisite for reliable state estimation. Based on how the data association is initialized, most of the existing trackers can be grouped into two categories: model-free-tracking and tracking-by-detection [Sun, Chen, et al. 2020]. To initialize a data association, the model-free tracking algorithms require a fixed number of VRUs and tracking-by-detection algorithms need variations in measurements [Sun, Chen, et al. 2020]. For instance, Blackman [Blackman 2004] takes five scans of data to initialize its data association. Postponing initialization too long may lead to a late or worse response. Simultaneously, initializing a data association algorithm after a single measurement increases the risk of introducing false positives, which leads to an uncomfortable driving experience.

To initialize a data association for a VRU tracker, this work determines a step to initialize a VRU tracker with a probabilistic model. The step helps to choose the correct hypotheses by determining values for parameters within the probabilistic model. In this work, contributions are as follows:

- This chapter proposes a step to find a trade-off between the parameters in a probabilistic model. This step helps data association to reduce the lag between the first detection and track.
- This chapter validates the step's performance by evaluating it on simulation data, custom-build data, and KITTI raw data.

The remainder of this chapter organized, as follows: Section 3.2 discusses literature related to the initialization problem. Section 3.3 describes MHT to fuse and associate data, and Section 3.4 presents the VRU tracker. In Section 3.5, the procedure is validated using different data sets. The conclusion and outline future research directions are available in Section 3.6.

## 3.2 Related work

Among the various studies regarding the initialization of data association of trackers, some of the researchers decide to skip some frames to achieve information regarding the detection [Gunawan et al. 2017; Köpf et al. 2020; Liao and Zhang 2017; Radac and Precup 2019; Zhang and van der Maaten 2013; Zhang and Van Der Maaten 2013]. The skipped frames mean that they remember the detections and start to 'trust' a measurement if it appears multiple times. The numbers of skipped frames are varied based on test situations, such as the numbers of VRUs.

Ding et al. [Ding et al. 2016] use prior knowledge of past detections to initialize the track. They define the probability of the existence of the object based on the past measurements for each object. However, in [Zou et al. 2019], the authors show that [Ding et al. 2016] faces more clutters and false tracks when new detections and tracked detections overlap. Leibe et al. [Leibe et al. 2007] utilize several previous frames to initialize a new track. Azim et al. [Azim and Aycard 2010] compute the Euclidean distance between the predicted position and new measurements for this purpose. When the distance is less than a certain threshold, they assume that the tracker is initialized. Morimitsu et al. [Morimitsu et al. 2017] use a fixed number of objects in the first frame. Subsequently, they localize those objects in the subsequent frames. Singh et al. [Singh et al. 2008] use the global statistics of tracks, linear motion models, and color models to initialize a hypothesis. Schulter et al. [Schulter et al. 2017] localize objects in each frame. Next, they connect those objects to the existing trajectories. Postponing initialization can have a negative impact not only on the quality of a data association, but also on the overall VRU tracker.

As mentioned, a delay in data association makes tracking hard. Different approaches can be used to deal with the data association. For example, nearest neighbor standard filter [Li and Bar-Shalom 1996], global nearest neighbor approach [Blackman and Popoli 1999], joint probabilistic data association [Fortmann et al. 1983], MHT, and finite set statistics [Vo et al. 2005]. Among them, MHT is one of the popular approaches, since it considers data association across multiple input data and multiple hypotheses. MHT grows a tree of hypotheses, based on deterministic branching decisions [Kim, Li, et al. 2015]. To increase the performance of MHT in a VRU tracker, researchers investigate the effects of several parameters, such as tuning VRUs detection, optimizing hypotheses, motion modeling, and initialing tracks. Although initialization is the first step of data associations, it has received less attention than other parameters. In most VRU trackers, the authors utilize similar Poisson-based approaches to achieve prior knowledge and initialize tracks.

In probability theory and statistics, researchers utilize the Poisson distribution of variables to initialize a hypothesis tree. Pan et al.

and Moraffah et al generate the first track by using the average spatial density of new and false-positive object detections in order to address the initialization with Poisson distribution [Moraffah and Papandreou-Suppappola 2019; Pan et al. 2017]. They use the Poisson distribution for modeling the number of objects in a fixed interval of space or time. Pollard et al. [Pollard et al. 2011] use the Napierian natural logarithm to calculate a score of a track. This score is defined as a probability of the corresponding track. They calculate the first score of each track by utilizing $ln$ of the associated false-positive and correct detection. In another work, Pollard et al. propose a global weight to initialize a track [Pollard et al. 2009]. To do this, they use the track score and a statistical distance between the peak and predict track. The quick initialization of a data association can allow any trackers to dive into tracking with a lower error, fewer false positives, and the minimal time delay between the first detection and initialization of the VRU tracker.

To achieve a quick initialization, this chapter proposes steps to reveal the trade-offs between stochastic model parameters. Using the steps, this chapter can minimize the lag between first detection and appearance in a hypothesis-based tracker. This work proposes an initialization procedure for a hypothesis-based approach. This work evaluates the tracking of the VRU tracker with the results of RNN, GNN, and RNN-GNN.

Recurrent neural network (RNN) is one of the popular methods in multi-target tracking with learned data association methods [Fruhwirth-Reisinger et al. 2020]. Gradient-based neural network (GNN) is also an online time-varying learning-based model for object trajectory [Zhang, Chen, et al. 2009]. The authors in [Zhang, Chen, et al. 2009] combine the RNN and GNN (RNN-GNN) to improve traditional online multi-target tracking. In their work, data association depends on the intersection between tracks and detections.

## 3.3 Multiple hypothesis tracker

This part describes the idea underlying MHT, and then presents the reason to select MHT.

### 3.3.1 Basic idea

A successful VRU tracker must handle the data association. Data association assigns the correct measurement with the correct track, initializing new tracks and detecting and rejecting measurements. There are different methods for implementing the data association; MHT is a method for solving the data association problem. The reader refers to [Cox and Leonard 1994] for a more elaborate explanation of the data association problem and ways to handle it.

MHT generates a tree with several branches. Each branch is a hypothesis, and it forms with different possible associations for each measurement. The principle idea of the MHT is to consider all possible hypotheses in parallel, which means that data association decisions can be deferred until uncertainties on data association are resolved [Kim, Li, et al. 2015].

In this work, each hypothesis $H_n^k$ contains a list of VRUs, the estimation of their 2D position, and their velocity. Where $n$ is hypothesis index $n = 1, \ldots, N$ and N is the number of hypotheses at time $k$. Furthermore, by applying MHT, it is possible to update a state of VRUs in a probabilistic way with data from the current time.

If no measurement is compatible with one of the existing hypotheses, a new hypothesis or a clutter (a false detection) should be formed. For each hypothesis, the measurement can be explained differently. The VRU tracker fixes the following probabilities for each hypothesis.

- The probability $P_{new}$ of a new VRU. $P_{new}$ indicates the probability that the measurement originates from VRUs not present in the scenes. It means that the data association finds no match between a detected VRU and one specific hypothesis.

- The probability $P_e$ of an existing VRU. Different hypotheses have different numbers of VRUs at different locations. Therefore, the number of existing VRUs will vary among hypotheses. $P_e$ represents the probability that the measurement originates from the VRUs that are already present in the hypothesis tree.
- The probability $P_c$ of a clutter.

The probabilities of the hypotheses being correct is calculated using Bayes theorem.

$$P(H_n^k|Z^k) = \frac{P(Z^k|H_n^k, Z^{k-1})P(H_n^k)P(H_n^{k-1}|Z^{k-1})}{P(Z(k)|Z^{k-1})} \tag{3.1}$$

where $Z^k$ means all of the measurements received up to time $k$ and $Z(k)$ means all the measurements received at time $k$. $P(H_n^k|Z^k)$ is the posterior probability of the hypothesis with index $n$ given all input measurements up to at time $k$, $P(Z^k|H_n^k, Z^{k-1})$ is the likelihood, which is a conditional probability for a given measurement at time step $k$. $P(H_n^k)$ is a prior probability at time $k$. $P(H_n^{k-1}|Z^{k-1})$ is the posterior probability of the parent hypothesis. Besides, $P(Z(k)|Z^{k-1})$ shows the normalization term. In this work, the prior probability is a function of the $P_{new}, P_e$, and $P_c$.Each hypothesis is a parent of multiple hypotheses with the constant ($P_{new}$,$P_e$,$P_c$) at time $k+1$. When VRUs are tracked, the probabilities of each hypothesis will be computed, and then the VRU tracker continuously updates the state of VRUs with the next measurements at $k+1$.

Because the VRU tracker has no prior knowledge regarding the number of VRUs, it needs to consider all different hypotheses. Enumerating all of the hypotheses would lead to memory overload fast. Therefore, the growth of the hypothesis tree should be managed by merging or pruning the least probable hypotheses and keeping the most probable one [Kim, Li, et al. 2015]. That way, the tree does not grow exponentially with its depth. Figure 3.2 gives a schematic of the tree. Each black dot represents a hypothesis and a red dot represents a pruned hypothesis.

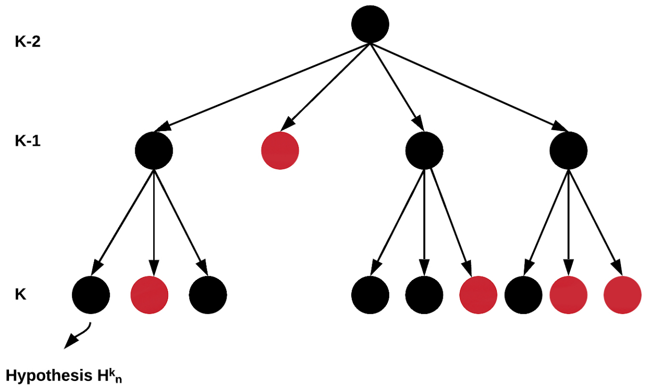


Figure 3.2: Tree representation of the formed multiple hypotheses. Black dot represents a hypothesis, and a red dot represents a hypothesis that pruned.

As mentioned, each VRU in a hypothesis has three probabilities. These fixed probabilities have a significant role in the management of the tree. Therefore, correct initialization helps to select and keep the most probable hypotheses. For instance, in a case $P_{new}$ is high when compared to the other probabilities, detections quickly lead to new VRUs (according to the most probable hypothesis). As a result, the delay between the first detection and appearance in the VRU tracker's most likely hypothesis is low.

On the other hand, this also means that more clutters will enter the VRU tracker. Setting $P_{new}$ very low leads to more considerable delays but, at the same time, minimizes the appearance of false positives in the VRU tracker. This paper is about finding a trade-off between these three probabilities for a hypothesis tree.

### 3.3.2 Reason

Maintaining multiple hypotheses and the ability to correct previous conclusions that are based on new detections are the reasons for choosing the MHT.

## 3.4 VRU tracker

To track VRUs, this work uses an approach introduced [Elfring et al. 2013]. They present a probabilistic environmental description for indoor applications. By changing their state estimator and initialization, this chapter adapts their work to utilize as a VRU tracker for outdoor purposes. This chapter uses Chapter 2 to change their state estimator.

### 3.4.1 State estimator

This chapter assumes that the relative velocity of the VRU with respect to the car is constant over the prediction in order to estimate the 2D position and the velocity of VRUs in a hypothesis. Thus, a Kalman filter with a constant velocity process model is used in the process model. A state estimator provides the position $(m)$ and velocity $(m/s)$ of pedestrians' multiple joints related to a detector.

Figure 3.3 shows the car frame coordinate system. This work denotes components of the relative distance of each VRU to the car frame as $x$ and $y$. $V_x$ and $V_y$ also define the velocity of VRU with respect to the car.

- To track VRUs, the VRU tracker collects data from a camera that was mounted on the top of the car. In this work, measurement contains the 2D position of VRUs. In the next step, the VRU tracker with the MHT framework generates a hypothesis tree in which each branch of the tree is one possible set of data associations. For each hypothesis, the process model estimates the 2D position and the velocity of VRUs. The VRU tracker, based on the output of the data association and the process model, decides to update the tree or add new branches by receiving a new measurement. Subsequently, it chooses the most probable hypothesis. The probability of person $A$ is a new/existing/clutter VRU, and person $B$ is a new/existing/clutter VRU.

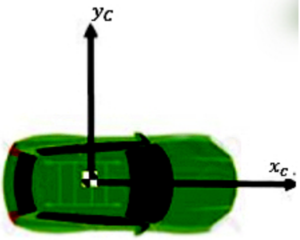


Figure 3.3: The car frame coordinate system. $x_c$ denotes the direction of driving, and $y_c$ indicates the side direction of the car.

### 3.4.2 Performance

In this chapter, multiple Object Tracking Precision (MOTP) and a Multiple Object Tracking Accuracy (MOTA) are selected as performance metrics to evaluate the effects of the initialization on a data association of a VRU tracker [Bernardin et al. 2006]. These two types of metrics quantify different relevant aspects. MOTA and MOTP can compare the effect of the initialization on the VRU tracker with other state-of-the-art algorithms.

The difference between the result of our VRU tracker and the ground truth state vector for the most probable hypothesis is $(SE)$. MOTP is an average of $SE$ and the total number of VRUs in the hypothesis $c(k)$. MOTA consists of the number of false detections $fp(k)$, the number of ID switches $IDs$, and the number of missed VRUs in the most probable hypothesis $m(k)$.

$$MOTA = 1 - \frac{\sum_k (m(k) + fp(k) + IDs)}{\sum_k g(k)} \tag{3.2}$$

where $g(k)$ is the number of objects present at time k.

$$\begin{aligned} MOTP &= \frac{\sum_{j,k} \sqrt{(x_j(k) - xc_j(k))^2 + (y_j(k) - yc_j(k))^2}}{\sum_k c_k} \\ &= \frac{SE}{\sum_k c(k)} \end{aligned} \tag{3.3}$$

where $xc_j(k)$ and $yc_j(k)$ are the ground truth position of each VRU with respect to the car. We briefly discuss an example to obtain a better understanding of MOTA.

- This work assumes that a detector recognizes true positive VRUs. Although a VRU tracker correctly receives measurements, it defines one of the VRU as a $fp(k)$. Besides, it tracks more VRUs $m(k)$ that are not available in the input. Therefore, $m(k)$ indicates a mismatch between the number of VRU of a hypothesis and ground truth. Moreover, a VRU tracker may switch the ID of each VRU during tracking that it calls $IDs$.

This chapter is about finding the balance that optimizes the data association performance in terms of MOTA and MOTP. Therefore, it proposes sets of values for $P_{new}$, $P_e$, and $P_c$, which leads to maximizing the metrics in various circumstances. In order to do so, this chapter analyzes simulated and real-world data with different numbers of VRUs entering the scene. They are recorded with different vehicles, sensor sets, detection algorithms, and different countries. The results show that there is a possibility to identify sets of values $(P_{new}, P_e, P_c)$. These sets can be used to initialize hypothesis-based data association and maximize the two performance metrics. In general, these sets for each VRU are constant, and in the following situation they can be updated :

- Whenever a tracker starts tracking a new VRU.
- When a tracker faces a partial and complete occlusions or the number of VRUs of the current measurement is different from previous ones.

## 3.5 Results

This chapter defines a test case to check the data association's performance in terms of the metrics. In this test case, the car receives the 2D position of the VRUs while the VRUs are crossing a road. In all of the simulations, although one test case is repeated, the following parameters are varied. It means that these parameters are constant during each test, and before starting each test they are fixed manually. By changing each of them, the metrics are analyzed. Moreover,there is an expectation that these values are varied in real-world applications.

- $R$ represents the measurement noise covariance matrix of zero-mean additive Gaussian measurement noise. Based on sensors' specifications and calibration, this work defined a valid range for variation of $R$. Kalman filters use R for estimating the state of VRUs. To estimate $R$ in the simulations, this work uses [Bavdekar et al. 2011]. $R$ is changed in the simulation to investigate whether sensors with different amounts of noise require different values $(P_{new}, P_e, P_c)$.

- The number of moving and standing VRUs. This work defines a state estimator for each VRU. In each time, the hypothesis tree changes based on the measurements. Moreover, the car may encounter many scenarios in which the number of VRUs and the way they move varies. Therefore, this parameter is changed in the simulations to have an optimal data association.

- The probability of new VRUs $P_{new}$, probability of clutter $P_c$ , and probability of existing VRUs $P_e$. As discussed earlier, these probabilities are used to compute the posterior probabilities of all hypotheses given the measurements. The values of three probabilities have to sum up to one, so, by having two of them, the third one can be computed. The probabilities can be set in different ways, depending on the preferences of the user.

Real-world experiments on the university campus are performed to validate the sets of values. This chapter validates the procedure using KITTI data to benchmark performance on a broader range of scenarios and compare with existing work [Geiger, Lenz, Stiller, et al. 2013]. KITTI contains the following information at 10 Hz in a city, residential area, campus, and road:

- 3D Velodyne point clouds that we use as ground-truth measurements.

- 3D GPS/IMU data.

- Calibration of sensors working at different rates to obtain ground truth data.
- 3D object track-list labels. Classify objects as a class of pedestrians, cyclists, cars, and trucks.
- Raw and processed color stereo sequences. Figure 3.4 shows one of the individual benchmarks of KITTI. To collect the data, they equipped a standard station wagon with two high-resolution color and grayscale video cameras [Geiger, Lenz, Stiller, et al. 2013].

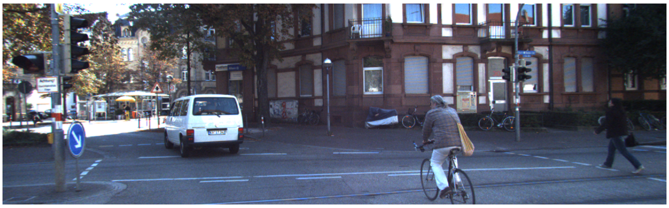


Figure 3.4: Sequence ‘5’ from the KITTI raw dataset.

Subsequently, all of the results are used to find an experimental and analytical relation between MOTP, MOTA, and different sets of probabilities. In the last step of the validation, the data association error due to Poisson’s initialization and the proposed sets of probabilities are compared. Robot Operating System (ROS) is used on Linux OS (Ubuntu 16.04) to achieve all the results. It should note that the calculations were on an Intel Core i7-6700HQ, CPU 2.6 GHz. To log or playback the data, ROSBAG are used. The average size of each ROSBAG file in the simulation was 200 (kB), in the experiment was 50 (MB). For the KITTI benchmark, the average size of each test was 1 (GB).

### 3.5.1 Simulation

False-negative detections can have a significant effect on the data association and the hypothesis tree to keep or prune a hypothesis. Therefore, in order to find a relation between relevant parameters and the hypothesis tree, in the simulation part, this work assumes that there is no false-negative detection. Gaussian noise is utilized as an example of the measurement noise.

To optimize the probabilistic models have used within the hypothesis tree,the following assumptions have been taken:

- The sensors deliver the 2D position of VRUs.
- The average VRU walking speed at crosswalks varies between 0 to 1.4 m per second.
- Based on the sensors' field of view, a maximum of 20 VRUs can be detected.

This chapter contains different simulations to investigate whether or not the assumptions influence the values $P_{new}, P_e$, and $P_c$ and the hypothesis tree. Firstly, the simulation are run in an ideal situation without considering the effect of measurement noise. The settings of the simulations vary, as follows:

- Simulate the effect of the number of standing VRUs per area on the VRU tracker. This simulation has been done in an ideal situation.
- Simulate the effect of the walking or/and standing VRUs on the VRU tracker in an ideal situation.
- Repeat the first simulation in a non-ideal situation.
- Repeat the second simulation in a non-ideal situation.

The horizontal axes are $P_{new}$ and the vertical axes are $P_e$ in order to read the figures. The summation of the probabilities is one, as mentioned earlier. For example, when the point in the lower left has $P_e = 0.2$, $P_{new} = 0$, and, hence, $P_c = 0.8$. Moreover, MOTP is the average overall VRUs/times. In all figures, dark blue means the minimum MOTP, and yellow indicates the maximum MOTP.

Figures 3.5 and 3.6 show varying the probabilities sets to change MOTP. In these two figures, the MOTP of five standing and walking VRUs into an ideal situation are computed. As a result, by changing these probabilities sets, different hypotheses are selected; then, the hypothesis tree delivers different state estimates. Therefore, the hypotheses tree gets the most probable hypothesis faster than the other sets of probabilities by increasing $P_{new}$. Moreover, having the most probable hypothesis shortly after the detection reduces the difference between the ground truth data and process model estimations, especially for walking VRUs.

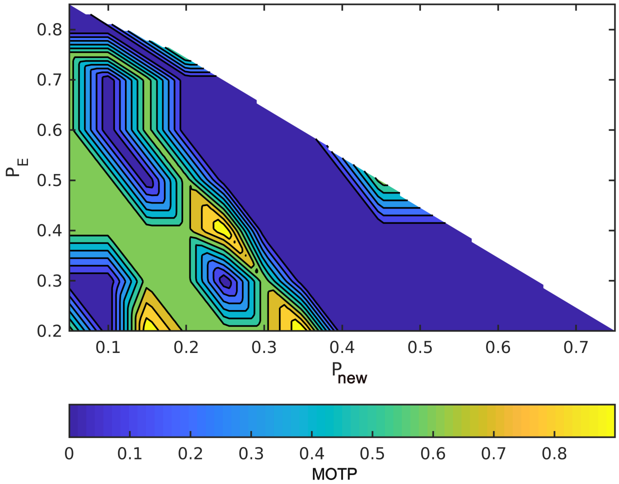


Figure 3.5: Simulate the effect of different sets of probabilities on $MOTP$ with static VRUs and without covariance noises.

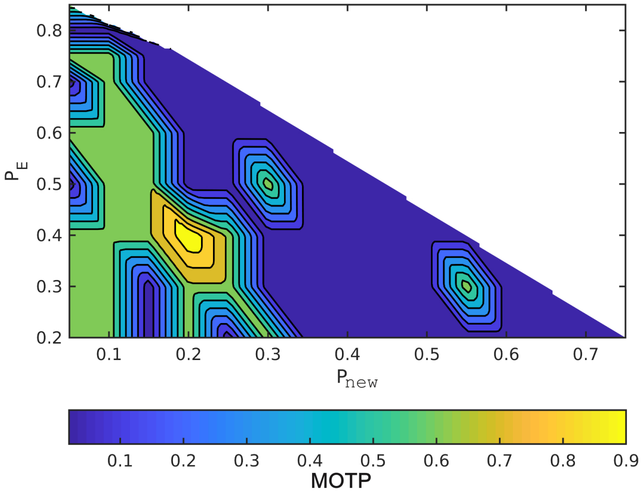


Figure 3.6: Simulate the effect of different sets of probabilities on $MOTP$ with dynamic VRUs and without covariance noises.

Figures 3.7 and 3.8 illustrate the effect of the covariance metric in MOTP. In fact, in these two figures, the previous simulations were repeated using different measurement covariance matrices. Although MOTP increases in Figures 3.7 and 3.8, different sets of $(P_{new}, P_e, P_c)$ can be achieved with lower MOTP than the other sets. These sets are almost similar to the first simulation. For instance, a common region in terms of values for $P_{new}$, $P_e$, and $P_c$ can be achieved by comparing the Figure 3.8 with Figure 3.6. MOTP of this region is lower than other parts. Based on the simulations, a common range of values for $P_{new}$, $P_e$, and $P_c$ is identified. The range of values leads to the best trade-off in terms of MOTP.

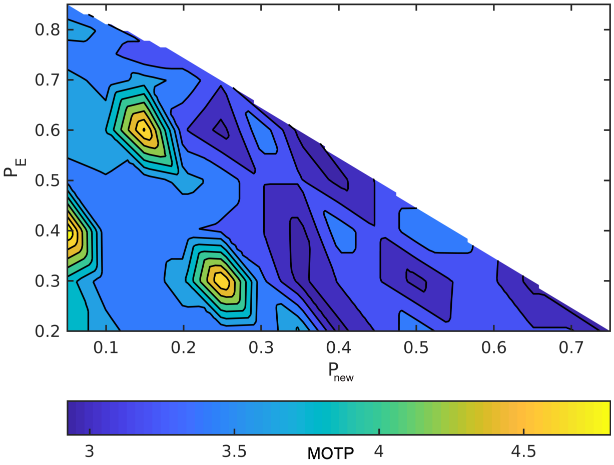


Figure 3.7: Simulate the effect of different sets of probabilities on *MOTP* with static VRUs and with one meter measurement noise.

These simulations are repeated 6400 times in different situations with various speeds, numbers of VRUs, and probabilities for obtaining this region. It means that, in 16 variations of the number of VRUs, walking situation, and measurement noise, this work considered a fix $P_{new}$ and varied the $P_e$ in a range of 0 to 1 and a step of 0.05. Subsequently, this work fixed $P_e$ and repeated the same simulation for variation of $P_{new}$ in a range of 0 to 1 by a step of 0.05. Based on the results of all the simulations,

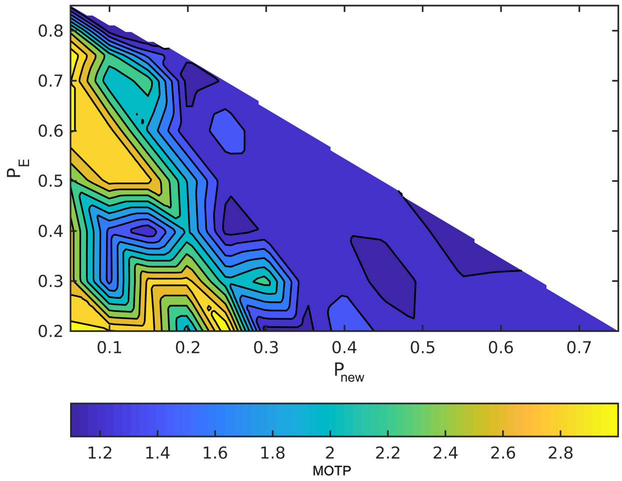


Figure 3.8: Simulate the effect of different sets of probabilities on $MOTP$ with dynamic VRUs and with one meter measurement noise.

Table 3.1 summarizes the minimum MOTP and its probabilities. Where $W$, $S$, and $W-S$ are walking, standing, and both walking and standing VRUs. It is observed that MOTP has the minimum error when the probability of new VRUs bigger than the probability of existing ($P_{new} > P_e$) and the probability of clutter ($P_{new} > P_c$), as shown in the table. Therefore, a region based on the sets of probabilities is obtained.

The minimum value of MOTP means that the data association works with a lower time delay and skips less measurement than other sets of ($P_{new}, P_e, P_c$). It means that the initialization of the data association can help to have a correct guess. Therefore, a correct set of probabilities can decrease the number of false tracks.

### *3.5.2 Experimental setup*

To experimentally evaluate the range of ($P_{new}, P_e, P_c$) procedure, a custom-built autonomous car prototype is used (Toyota Prius, in which sensors and other hardware added). The tests are executed with a vehicle speed of 15 km.h$^{-1}$ on a university

Table 3.1: Probabilities sets that cause the minimum MOTP for each scenario.

| N/VRU | W/S | noise(m) | Min $MOTP$(m) | $P_{new}$ | $P_e$ | $P_c$ |
|---|---|---|---|---|---|---|
| 1 | W | 0 | 0.00 | 0.65 | 0.2 | 0.15 |
| 2 | W | 0 | 0.03 | 0.45 | 0.35 | 0.2 |
| 5 | W-S | 0 | 0.00 | 0.4 | 0.35 | 0.25 |
| 10 | W-S | 0 | 0.00 | 0.45 | 0.4 | 0.15 |
| 20 | W-S | 0 | 0.08 | 0.45 | 0.3 | 0.25 |
| 1 | S | 0 | 0.00 | 0.7 | 0.2 | 0.1 |
| 5 | W-S | 0 | 0.00 | 0.5 | 0.3 | 0.2 |
| 10 | W-S | 0 | 0.80 | 0.5 | 0.3 | 0.2 |
| 1 | W | 1 | 0.04 | 0.5 | 0.3 | 0.2 |
| 2 | W | 1 | 0.06 | 0.4 | 0.35 | 0.25 |
| 5 | W-S | 1 | 0.80 | 0.45 | 0.3 | 0.25 |
| 10 | W-S | 1 | 0.18 | 0.45 | 0.25 | 0.35 |
| 20 | W-S | 1 | 0.25 | 0.5 | 0.3 | 0.2 |
| 5 | S | 1 | 0.40 | 0.5 | 0.3 | 0.2 |
| 5 | W-S | 1 | 0.80 | 0.45 | 0.3 | 0.25 |
| 10 | S | 1 | 0.18 | 0.45 | 0.25 | 0.3 |

campus. Figure 3.9 shows the real test situation; the car detects the pedestrians for the first time. The experiments are repeated one scenario in 11 different periods of one day in Spring. Investigating the effects of measurement noises and external distur-

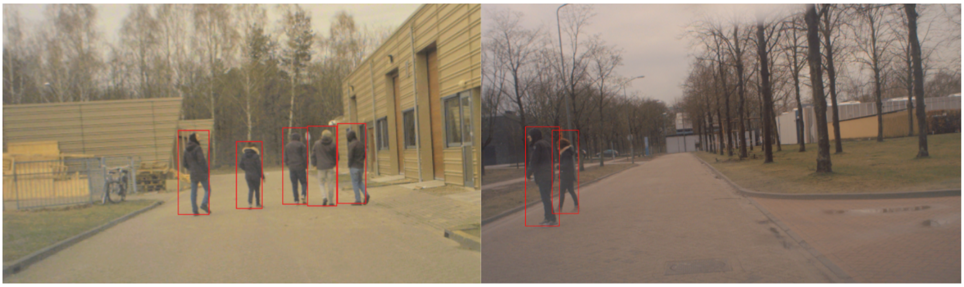


Figure 3.9: Illustration of our experimental test. The car receives the center of each rectangle as the position of pedestrians.

bances, such as weather, light condition, and camera movement on our initialization was our reason for repeating the tests.

The car is equipped with a streaming camera for VRUs detection and a GPS to provide data on its position. In these experiments, Fast Region-based Convolutional Neural Network ($FastR{-}CNN$) are used to detect VRUs [Ren et al. 2015]. During the experiments, the pedestrians cross the road, regardless of the presence of the car. After passing an intersection, the car's camera detected the pedestrians. At the same time, $FastR-CNN$ extracts the pedestrians, and it makes a boundary box around each of them. After that, the VRU tracker receives the center of bounding boxes in the Cartesian coordinate frame as positions of pedestrians at 10 Hz.

Figure 3.10 illustrates MOTP for different probabilities in an experimental test. Table 3.2 shows the minimum and maximum values of MOTP when setting the optimal and non-optimal values for $P_{new}$, $P_e$, $P_f$ in all of the experimental tests. Our reason to compute the maximum value of MOTP is to show the effect of different sets of probabilities on our data association. For instance, in test 1, the VRU tracker estimates the states with 1.8 m error if the set is defined from out of the region.

Referring to Figure 3.10 and Table 3.2, the region that is mentioned in the previous part is in place. To investigate the effect of the region on MOTA, both MOTA and MOTP are computed in all of the experiments. Based on Table 3.3, MOTP and MOTA have more reliable results if the probabilities are selected from the region. The MOTA values indicate that, although there are false positive detections, the data association does not consider false-positive detections. For example, Figure 3.11 indicates a real situation that the tracker receives a false-positive detection as a VRU. Based on the initialisation, the data association selects a hypothesis that assumes that the object is a false-positive detection.

In the next step, the possibility of defining a set of fixed values for $P_{new}$, $P_e$, and $P_c$ are investigated, which leads to good performance in all circumstances. Therefore, the following steps

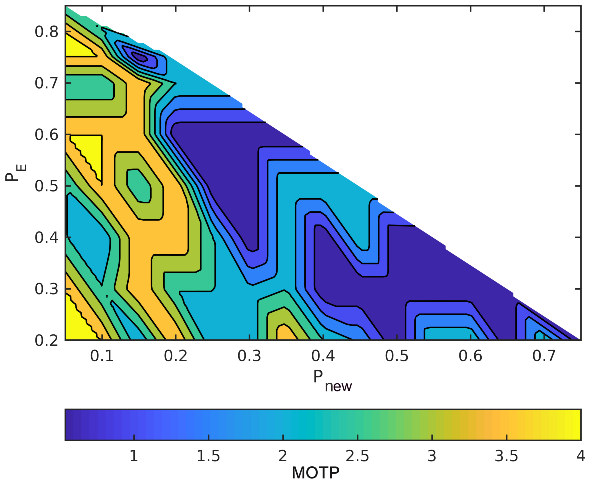


Figure 3.10: Different sets of probabilities affect MOTP of an experiment test.

Table 3.2: Minimum and maximum of MOTP for each experiment

| **No.** | **Min** $SE$**(m)** | $P_n$ | $P_e$ | $P_c$ | **Max** $SE$**(m)** | $P_n$ | $P_e$ | $P_c$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.05 | 0.6 | 0.25 | 0.15 | 1.80 | 0.3 | 0.2 | 0.5 |
| 2 | 0.16 | 0.4 | 0.35 | 0.25 | 3.12 | 0.25 | 0.3 | 0.25 |
| 3 | 0.13 | 0.5 | 0.3 | 0.2 | 2.74 | 0.1 | 0.5 | 0.4 |
| 4 | 0.01 | 0.6 | 0.25 | 0.15 | 1.55 | 0.05 | 0.2 | 0.75 |
| 5 | 0.14 | 0.4 | 0.35 | 0.25 | 2.75 | 0.25 | 0.25 | 0.5 |
| 6 | 0.12 | 0.4 | 0.3 | 0.3 | 2.71 | 0.1 | 0.25 | 0.65 |
| 7 | 0.11 | 0.65 | 0.3 | 0.05 | 1.42 | 0.05 | 0.2 | 0.75 |
| 8 | 0.09 | 0.6 | 0.15 | 0.25 | 1.60 | 0.05 | 0.2 | 0.75 |
| 9 | 0.77 | 0.5 | 0.35 | 0.15 | 3.03 | 0.05 | 0.55 | 0.4 |
| 10 | 0.138 | 0.6 | 0.3 | 0.1 | 2.10 | 0.1 | 0.75 | 0.15 |
| 11 | 0.01 | 0.45 | 0.3 | 0.25 | 1.57 | 0.05 | 0.2 | 0.75 |

have been done:

- Calculate an average of the probabilities in all experiments when we have the minimum errors.

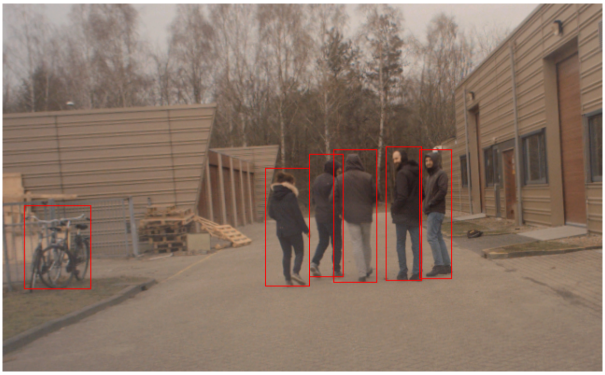


Figure 3.11: This detection contains a false positive since our detection method detected a bicycle as a VRU. Therefore, our tracker should select a hypothesis that assumes those bicycles are false-positives.

Table 3.3: Average of the minimum and maximum MOTP and MOTA for all the experiments

| **MOTP Max** *SE* | **MOTA Max** *SE* | **MOTP Min** *SE* | **MOTA Min** *SE* |
|---|---|---|---|
| 60% | 66% | 91% | 89% |

- Compute both metrics based on the average set of probabilities.

Table 3.4 represents the values of MOTA and MOTP for this set. Although these metrics are lower than Table 3.3 in the minimum *SE*, a constant set of probabilities can be obtained for the initialization. Therefore, there is no need to change the probabilities for different parts of our experiments, since it is inconvenient to dynamically change the probabilities.

The results indicate a significant relationship between the performance of the VRU tracker and the probabilities. Besides, this chapter can find a constant set of probabilities experimentally that can be used in different scenarios. During our experiments, the VRU tracker is associated and tracked up to 0.5 ($s$) faster when constant probabilities are used.

Table 3.4: Average MOTP and MOTA in a set of probabilities

| $P_n$ | $P_e$ | $P_c$ | **MOTP** | **MOTA** |
|---|---|---|---|---|
| 0.5 | 0.3 | 0.2 | 84% | 78% |

### 3.5.3 KITTI raw data

As mentioned before, this work verifies the VRU tracker on raw data recordings of KITTI. Sequences of the raw data containing pedestrian and cyclist categories are used for assessments on the raw KITTI data set. Table 3.5 reveals, for each sequence, the minimum values of $SE$.

Table 3.5: Minimum $SE$ for each sequences

| **No.** | **Minimum error(m)** | $P_{new}$ | $P_e$ | $P_f$ |
|---|---|---|---|---|
| 5 | 0.023 | 0.75 | 0.2 | 0.05 |
| 9 | 0.27 | 0.5 | 0.3 | 0.2 |
| 11 | 0.06 | 0.6 | 0.25 | 0.15 |
| 60 | 0.10 | 0.5 | 0.3 | 0.2 |
| 59 | 0.06 | 0.6 | 0.25 | 0.15 |

Table 3.6 shows the effect of the initialization on [Fruhwirth-Reisinger et al. 2020]. In the same dataset, this work uses their state estimator and replace their data association algorithm. Subsequently, their tracker is initialized with the proposed initialization. The first line of the table is the MOTA of [Fruhwirth-Reisinger et al. 2020] that was published in the KITTI website. In the second line, a hypothesis-based data association is used and initialized based on the best MOTA. The third line represents the hypothesis-based data association using a set of probabilities that achieved in the experimental part ($P_{new} = 0.5$ ,$P_e = 0.3$ , $P_c = 0.2$). It calls the set "define set” ($DS$).

Table 3.6: Compare our work with a state-of-the-art algorithm on KITTI raw dataset

| **Name** | $MOTA$ | $Ours_{min}$ | $Ours_{DS}$ |
|---|---|---|---|
| [Fruhwirth-Reisinger et al. 2020] | 72.1% | 73.2% | 72.6% |
| [Choi 2015] | 57.61% | 57.93% | 57.67% |
| [Zhou et al. 2020] | 53.84% | 54% | 53.86% |

Based on Table 3.6, the initialization produces the highest number of $MOTA$. It means that, during the initialization, our data association outperforms the benchmark method in terms of MOTA. Additionally, the table shows that. Even by setting constant probabilities for all KITTI sequences, this work can perform better than a state-of-the-art algorithm. It means that $P_{new} = 0.5$, $P_e = 0.3$, $P_c = 0.2$ have an acceptable performance to track VRUs. Moreover, a comparison between our result and state-of-the-art in Table 3.6 reveals that our initialization procedure can minimize the number of false-positive and miss VRUs. The reasons for having less false-positive and miss VRUs are as follows:

- The VRU tracker can keep multiple hypotheses to rematch measurement with VRUs. It means that, if our VRU tracker receives a false positive detection, it can correct itself after receiving the measurements in the next timestamps.

- Besides, selecting the probabilities based on the region helps the data association to match data without skipping measurements. Therefore, the probability of missing VRUs is low.

The initialization affects the results of the full track. Therefore, there is a possibility to improve the performance metrics in the initialization phase. Based on the outputs of simulations, KITTI, and experiments, for the range of the measurement noise, our data association can be initialized by selecting the probabilities from the region.

### 3.5.4 Comparison

This chapter compares two different types of data association's initialization, the Poisson distribution and the proposed initialization. Poisson is one of the most popular distributions in the literature. To compare the two methods, the Poisson distribution is utilized to initialize the VRU tracker. The same procedure is taken, as in [Pan et al. 2017]. This reference gives detail regarding setting the Poisson in a hypothesis tree. Besides, this work uses data from the experimental test on the campus. Although two methods have the same $MOTP$ after initialization, the $MOTP$ of the proposed procedure is less than Poisson. The reason for this difference is that Poisson should skip some frames to achieve information regarding the average of false and new detection. Meanwhile, our initialization generates the tree and estimates the state of VRUs by choosing the probabilities from the region.

$SE$ is a difference between the ground-truth data and the results of a state estimator, as shown in Equation (4.1). To have a fair comparison between the initialization procedures' effects, the same tracker and test data are used. The test data were the 11 experiments, and Chapter 2 was used for tracking pedestrians. It should be noted that the same tracker means the same data association and the same state estimator. To initialize data association, the Poisson method should wait for few frames to collect data and build up confidence. Based on Figure 3.12, Poisson waits for three frames to collect data, which, here, the waiting time is 0.3 s. As a result of the 0.3 $(s)$ delay in initializing the hypotheses-based data association, the hypothesis tree assumes a part of the measurements in the first 0.3 s is clutter. Therefore, the state estimator could not estimate the states close to the ground-truth data.

In the meantime, based on Figure 3.12, this trade-off skips only one frame, which means 0.1 s. Therefore, this initialization helps the data association to gain more measurement. Hence, the state estimator can estimate the state vector close to the ground-truth data. As a result of saving 0.2 s, this initialization performs faster than Poisson. Therefore, in Figure 3.12, initialization with the Poisson approach can lead to 2% less MOTP compared to this initialization. In Table 3.7, we applied Poisson and the procedure to initialize the data association. Subsequently, we computed the *SE* of all 11 experiments. Similar to Figure 3.12, Table 3.7 shows that the initialization has less error when compared to Poisson.

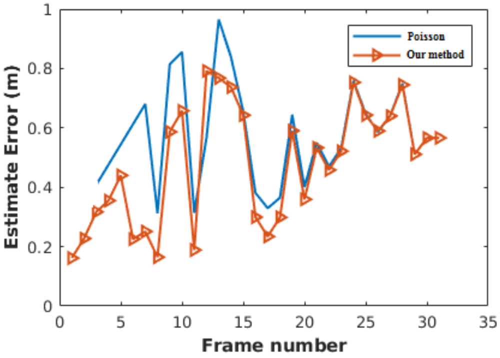


Figure 3.12: *SE* of the data association based on the procedure and the Poisson.

Table 3.7: *SE* of two different initialization procedures.

| No. | Minimum *SE* our method(m) | Minimum *SE* Poisson(m) |
|---|---|---|
| 1 | 0.05 | 0.14 |
| 2 | 0.16 | 0.25 |
| 3 | 0.13 | 0.14 |
| 4 | 0.01 | 0.08 |
| 5 | 0.14 | 0.16 |
| 6 | 0.12 | 0.13 |
| 7 | 0.11 | 0.19 |
| 8 | 0.09 | 0.20 |
| 9 | 0.77 | 0.81 |
| 10 | 0.14 | 0.20 |
| 11 | 0.01 | 0.03 |

## 3.6 Conclusions

This paper presents a collection of probabilistic elements to initialize the data association of VRUs trackers in an urban environment. The initialization of data association can affect the results of the entire track. It means that late initialization could lead to undesired, or even dangerous, situations. Therefore, there is a possibility to improve the performance of trackers by improving their data association. The primary purposes of our initialization are as follows:

- Minimize the delay between first detection and selecting the correct hypothesis.
- Discard false positives from a VRU detection.

This chapter finds a trade-off between the parameters in a probabilistic model. Various simulations, experimental tests, and the KITTI benchmark in different lighting and weather conditions are used to find the trade-off. This work demonstrated that the collection of probabilistic elements are valid for different numbers of VRUs and measurement noise. Moreover, it work shows that the probabilistic sets help to initialize hypothesis-based data association and maximize the performance metrics.

Using the collection of probabilistic elements, a hypothesis-based tracker can match data without skipping measurements, and a tracker can reduce the probability of missing VRUs. The MHT can compute better probabilities for hypotheses and it is more likely to select the correct one. Therefore, the collection of probabilistic elements have a significant role in the management of a hypothesis tree. Besides, our evaluations show that our approach has a superior performance in the simulation, real-time, and KITTI datasets. The results showed that the tracking precision and accuracy increase up to 3.6% with the proposed initialization as compared to the state-of-the-art algorithms in tracking VRUs. Multiple group tracking is also challenging in the field of autonomous cars. In future work, this work plans to investigate the effect of the initialization procedure on a data association of multiple group tracking.

# Chapter 4

# Multiple pedestrian tracking using vision-based sensors and IoT technology

**Abstract:** Pedestrian tracking has an essential role in awareness systems, having applications in autonomous driving and smart cities. One of the significant challenges during tracking is unseen or partially occluded pedestrians in urban scenes. In recent years, multiple sensors have been used to deal with occlusion. However, the tracking of unseen or occluded pedestrians is still an unsolved challenge. Accordingly, this paper proposes a collaborative pedestrian tracker based on the Internet-of-Things (IoT) and on-board vision-based sensors. In this work, IoT and vision-based sensors have different responsibilities. IoT technologies are used as an external source of information to update variables in a hypothesis-based data association algorithm and be aware of existing pedestrians, during times that cameras are unable to detect pedestrians. Besides, vision-based on-board sensors are used to estimate and predict the positions of pedestrians regarding a vehicle. As a proof of concept, the proposed tracker is validated using experiments and simulation data. The experiments showed that the tracking accuracy increased to 11% with the proposed tracker compared to a vision-based pedestrian tracker.

This chapter is based on:
Dolatabadi, M., Elfring, J., van de Molengraft, R. Multiple pedestrian tracking using vision-based sensors and IoT technology. Internet of Things. Submitted

## 4.1 Introduction

The transportation system is facing an increasing logistical demand due to high numbers of commuters [Sun and Boukerche 2020b]. One of the essential concerns in improving the transportation system is estimating the position of pedestrians [Combs et al. 2019]. In an urban environment, different objects can limit a camera-based tracker field of view such as pedestrians, vehicles, cyclists, or infrastructures. According to [Ning et al. 2021], camera-based trackers' accuracy varies due to limited viewing angles of the cameras, illuminating conditions, co- existing obstacles, and the relative position between pedestrians and vehicles. Therefore, camera-based tracking of multiple pedestrians' positions and velocities related to a moving vehicle is challenging [Dollar et al. 2011].

According to a survey by the U.S. Department of Transportation, pedestrians' low visibility is a major cause for pedestrian injuries in more than 75% of cases [Ning et al. 2021]. Hence, detecting and tracking multiple pedestrians who are being occluded is a critical challenge. During occlusion, objects such as vehicles or other pedestrians occlude pedestrians. To deal with occlusion, [Ning et al. 2021] provides a complete survey of current approaches. Based on their results during long-term occlusion problem, the current vision-based tracking algorithms are unable to track all the pedestrians. Based on [Merdrignac et al. 2016], external information can be used to improve a tracker's accuracy and decrease the side effects of low-visibility. Therefore, this work proposes an integrated approach to support vision-based pedestrian trackers and track pedestrians during occlusion. To do this, Internet-of-Thing (IoT) and vision-based data technologies are used.

IoT has led to an increase in the number of networked devices and has influenced vehicle-to-vehicle (V2V), vehicle-to-pedestrian (V2P), and vehicle-to-everything (V2X) connections to computing, storing, or sharing measurements [Gaurav et al. 2021]. IoT devices can provide a wide variety and diverse set of measurements, such as positions, velocities and, IDs. These measurements can be blended with localization and detection systems to enhance tracking accuracy [Shit and Sharma 2018]. This work proposes a tracker that combines vision-based sensors and IoT devices.

The proposed tracker uses various measurements with different accuracy. Therefore, the tracker should deal with data associations with different update rates and accuracy. To make an association between data, the idea of multiple hypothesis tracking (MHT) [Blackman 2004] is used. Therefore, different hypotheses are created to make an association between pedestrians and available IoT devices during tracking pedestrians. This work defines different responsibilities for IoT devices and camera detections. IoT technologies are used as prior knowledge to update probabilistic models in MHT and detect existing pedestrians, at times that as a result of occlusion cameras are unable to detect pedestrians. Contributions of this work are as follows:

- Proposed a tracker based on vision-based on-board sensors and IoT technology to track pedestrians during low-visible and fully occlusion situations.
- Validate the accuracy of the proposed tracker using simulation and experimental data.

The remainder of this paper is organized as follows: In Section II, we give an overview of the state-of-the-art. Then, in Section III, the proposed system architecture is presented. Section IV demonstrates the corresponding simulation and experimental test results for evaluating the proposed work. Finally, Section V concludes the paper.

## 4.2 Related work

Despite several years of research and development, different factors such as environmental conditions, occlusion, and illumination can affect the quality and robustness of vision-based trackers. Based on research carried out by [Su et al. 2015], during occlusion, the accuracy of trackers decrease if measurements are received from a single sensor. Single vision-based trackers are therefore unable to formulate collision avoidance in all possible circumstances [Sun and Boukerche 2020a].

One approach to improve performance of a tracker and deal with occlusion is the use of multiple on-board sensors such as LIDAR and cameras [Kim and Kim 2016; Redmon et al. 2016; Simonyan and Zisserman 2014]. For example, [Anuj and Krishna 2017; Boukerche and Hou 2021; Guo et al. 2015; Liang et al. 2020] have used multiple sensors to determine various features and states. The features have been defined as boundaries, edges, colors, and textures in image sequences, while the states have been defined as positions, velocities, and orientations of pedestrians. They use the features and states to track pedestrian before, during, and after occlusion. In spite of many works on pedestrian detection and tracking, multiple on-board sensors have limitations. Some situations are impossible to handle with on-board sensing [Rawashdeh and Wang 2018b]. Suppose a vehicle using multiple on-board sensors wants to turn into an intersection, and a group of pedestrians occupies the road. In these situations, the on-board sensors cannot detect pedestrian that are outside the field of view. Therefore, turning into the intersection would cause a dangerous situation.

Studies have suggested that using data from external sources can improve performance of existing trackers [Fiore et al. 2019; Gelbal et al. 2020; Van et al. 2021; Wang, Wang, et al. 2019]. Internet-of-Thing (IoT) technologies can be one external source. The IoT can recognize pedestrians that are in the blind spots of on-board sensors; IoT can communicate with cars to transfer data on the positions and velocity vectors of pedestrians. Vehicles can use this information to adjust their route or speed and may track pedestrians. Although a tracker's performance can be improved by using IoT and on-board sensors, a tracker that uses communicated data faces more challenges. The challenges can be defined as follows:

- Robust communication.

- Stable and robust data association.

In wireless communications, data loss is inevitable due to network congestion, sampling frequencies, environmental noise, and time synchronization. Therefore, there is a probability that some data may be corrupted, missed, or contain inaccurate measurements. In theses situations, it is hard to associate all the data from different sources.

Data association is the process of relating sensor to pedestrians. A data association algorithm should confirm or refuse a track in a short time frame. Otherwise, a tracker fails to track multiple pedestrians. The quick data association can allow any trackers to dive into tracking with a lower error, fewer false positives, and a minimal time delay between the first detection and the first track. Association data from multiple sources is needed to ensure the accuracy of the detection [Chen 2021]. As a result of non-robust communication, solving data associations is even more challenging when a tracker uses IoT data.

To associate camera detections and IoT devices, [Van et al. 2021] uses humans' acceleration, orientation, and rotation features. Based on the features, they pair IoT devices and camera detections. In complex environments with multiple pedestrians, there is a high probability that pedestrians walk in a group. Therefore, this association may pair devices and pedestrians wrongly. [Solmaz et al. 2019] shows that combining IoT and machine learning can improve the performance of autonomous driving to identify the crowded area and enhance safety in urban environments. Data association is out of scope of [Solmaz et al. 2019]. Therefore, they assume during crossing the road, they know which IoT device belongs to which pedestrians. Therefore, they only match IDs to make an association. In [Sun and Boukerche 2020a], they use Internet-of-Vehicle (IoV) to develop a practical pedestrian detection technology. In a simulation environment, they receive the position, speed, and acceleration of a pedestrian from other vehicles. Then, based on these measurements, their vehicle has information regarding the occluded pedestrian. The vehicle would need continuous measurements of pedestrians to calculate the corresponding motion status of pedestrians and make an association between pedestrians. Therefore, a pedestrian tracker is needed. [Sun and Boukerche 2020a] uses root mean square distance to make an association. In the real world with multiple pedestrians, this association may cause errors.

This work uses a hypothesis-based approach to make an association between detected pedestrian and available IoT devices. The hypotheses contain different associations, which define that a pedestrian is a new, existing, or even a false detection. Besides, the hypotheses describe which IoT device belongs to which pedestrian. The subsequent measurements will modify the probability of all the hypotheses. Based on [Renaudin et al. 2013], suppose a pedestrian carries a smartphone; then displacement of the devices can represent the displacement of the pedestrian. Therefore, this work uses this idea and estimates the displacement of IoT devices. The displacements, ID of IoT devices, and camera measurements are used to make associations between IoT devices and pedestrians.

Without managing data, the MHT algorithm suffers from the exponential growth of tracking hypotheses. Based on research in the Chapter 3, updating parameters in probabilistic models can improve hypothesis-based tracker performance. Thus, this study proposes the use of a vision-based tracker with IoT data to obtain prior knowledge and update a probabilistic model. In this work, accuracy is improved with respect to decreasing numbers of missed pedestrians, false positives, and identity switches between multiple pedestrians. This research does not examine whether a tracker can use only IoT data to track pedestrians or improve IoT data precision. Rather, it aims to associate low accuracy and low frequency IoT data and camera images to track pedestrians.

## 4.3 Input data

This work consists of a camera measurement, IoT measurement, coordinate transformation, synchronization, state estimation, and MHT for data association. Based on Input data and their update rate, sensors can have different responsibilities. Therefore, this work extracts different variables from the input data. As an example, IoT devices suffer from inaccuracy in position data and low update rates. Therefore, this work uses position data of IoT devices indirectly. In this work to use the IoT data, first, the following errors are removed from data.

- IoT data are received through different servers. Therefore, there is a probability that servers are unable to get new measurements and send communicated data that were saved in a cloud. Based on update rate of IoT devices, the tracker should remove incorrect timestamps. The data which do not correspond to the moment that a vehicle may receive data.
- Considering the Netherlands topology, all points with an altitude value of less than -7 and more than 322 meters above sea level should be removed.

This work assumes that each pedestrian has a maximum of one IoT device. Therefore, by counting the number of valid incoming IoT measurements, the tracker is aware of the minimum number of pedestrians. This number can be used to associate between pedestrians and hypotheses. The camera and IoT parts can offer the following measurement:

- Camera : Image sequences with timestamp and position of multiple joints of pedestrians.
- IoT : In this work, IoT data are received from smart devices of pedestrians. Each device provides the total acceleration from the 3-axis accelerometer after removing the component of gravity, geographical coordinates, and a unique ID. In this work, IoT devices have lower update rates than cameras.

## 4.4 Pedestrian tracker

In this work, the approach described in Chapter 2 is adopted to track pedestrians and estimate the position of multiple joints related to a vehicle. Chapter 2 provides the position $(m)$ and the velocity $(m/s)$ of pedestrians' multiple body key points with respect to a vehicle's camera. In this work, the tracker in Chapter 2 changes as follows:

- In this work, IoT and vision data with different update rates and accuracy are used. Therefore, the measurement model in Chapter 2 is modified to formulate an appropriate measurement model and identify measurement characteristics. Although adding more measurements can improve the performance of the proposed tracker, the tracker faces more challenges in data association compared to Chapter 2.
- The proposed tracker uses more features and measurements characteristics to make associations between pedestrians and measurements. Therefore, the state space is increased compared to Chapter 2. In the following section the new state space will be explained.
- In Chapter 2, constant variables are used in all situations. Although, this work uses the same probabilistic model as Chapter 2, based on situations, IoT data are used to update the probabilistic model.

### *4.4.1 IoT state estimation*

The acceleration and geographical coordinates of each device can give information about the speed and displacement. In this work, these information are used to analyze the displacement (change of position) of pedestrians. This work uses a Kalman filter to estimate the displacement.

This Kalman filter receives the accelerations in combination with geographical coordinates as observed states and estimates displacement of each devices. In this Kalman filter, the measurement noise is considered to be white and Gaussian. The state vector $(x)$ at time $(t)$ for each device in this Kalman filter is as follows:

$$x(t) = [p_x(t), p_y(t), d(t), v(t), q(t)] \tag{4.1}$$

where $p_x$,$p_y$ $d$, $v$, and $q$ represent position of a device in $x$ direction, position of a device in $y$ direction, displacement, velocity and orientation of each device. This work assumes that the position of each device can be modeled with a constant acceleration model. Therefore, the speed of the device at time $(t)$ can be estimated by integrating the acceleration over time. In this work, the distance is computed as a summation of velocities at discrete time instants.

### 4.4.2 Data association

In this work, data association is defined as matching newly detected pedestrians with pedestrians being tracked. Data association can also determine which IoT device belongs to which pedestrian. One of the existing challenges is that the number of detected pedestrians by the camera and IoT are not same over time. Therefore, the data association is more complicated than using only camera measurements. To deal with this challenge, MHT generates a hypothesis tree with branches. Camera detections and IoT devices can be associated with an existing pedestrian, clutter, or pedestrian that was not tracked before.

Each branch is a collection of hypotheses and can be formed with different possible associations. Hypotheses are considered in parallel. It means that data association decisions can be deferred until uncertainties on data association are resolved. The tree expands by receiving a new measurement at a time of $(t+1)$. Each hypothesis's probability is computed to pick the most probable hypothesis. [Elfring et al. 2013] explains how these probabilities are estimated. The probability for each pedestrian can be defined as follows:

- The probability of a new pedestrian. $P_{new}$ indicates the probability that the measurement originates from pedestrians not present in the tree. It means that the data association finds no match between a detected device, pedestrian, and one specific hypothesis.
- The probability of an existing pedestrian $P_{exist}$. Different hypotheses have different numbers of pedestrians at

different locations with different IoT devices and displacements. Therefore, the number of existing pedestrians will vary among hypotheses. $P_{exist}$ represents the probability of an association. This probability shows the measurement originates from a pedestrian and a device that is already present in the hypothesis tree.

- The probability $P_{clutter}$ of a clutter.

In this work, IoT data, camera measurements, the Manhattan distance, and the Poisson distribution are used to update the probabilities. Based on the ID of IoT devices in previous seconds and the number of pedestrians in each hypothesis, the MHT can realize whether the IoT devices in each hypothesis are existing or new. In the next step, Manhattan distance is estimated between available IoT devices and pedestrians. If no updates from a device are received for one second, the MHT should prune that hypothesis. The defined seconds based on update rate of sensors and their uncertainty would be variable. In this work, the following situations are considered to update the probabilities during driving while both IoT and vision sensors detect pedestrians.

- The number of pedestrians detected by a camera ($n_{camera}$) is equal to the number of devices detected by IoT ($n_{IoT}$). In this case, if the Manhattan distance is less than a threshold, then probabilities remain constant. The threshold is function of various parameters such IoT device noise, status of the car, and kinematic of pedestrian.

- $n_{camera} = n_{IoT}$ and the Manhattan distance is more than a threshold. In this situation, the Poisson distribution is used. The value of the Poisson is added to $P_{new}$.

- $n_{camera} < n_{IoT}$ and MHT recognizes there are new devices. In this situation, the Poisson distribution is used. The value of the Poisson is added to $P_{new}$.

- $n_{camera} < n_{IoT}$ and MHT recognizes there are existing devices. In this situation, the Poisson distribution is used. The value of the Poisson is added to $P_{exist}$.

In the Poisson distribution, if the mean number of detected pedestrians with vision-based sensors per interval is $(\lambda)$, the probability of observing undetected pedestrians $(k)$ in a given interval is given by

$$P = e^{-\lambda}\frac{\lambda^k}{k!} \tag{4.2}$$

## 4.5 Performance evaluation

IoT devices can improve the performance of the vision-based tracker. It means that the tracker would be aware of all pedestrians surrounding a vehicle without knowing their accurate position. Therefore, in this work, multiple object tracking accuracy $(MOTA)$ is selected as a performance metric to evaluate the effects of the proposed tracker. In timestamp $t$, MOTA consists of the number of false detections $f_p(t)$, the number times an ID of a pedestrian switches $IDs$, and the number of missed pedestrians in the most probable hypothesis $m(t)$. $g(t)$ is the total number of detections in each time.

$$MOTA = 1 - \frac{\Sigma_k(m(t) + f_p(t) + IDs)}{\Sigma_k(g(t))} \tag{4.3}$$

To evaluate the performance, this work adopts C++ to conduct the simulation and experiment tests. Moreover, a custom-built autonomous vehicle prototype is used (Toyota Prius, in which sensors and other hardware were added) to collect a series of experimental data. The tests were executed with a vehicle speed of 15 $(km/h)$ on a university campus. All the experimental tests are open source.

### 4.5.1 Simulation

Occlusion can significantly affect the performance of tracking and the data association to keep or prune a track. Aim of this part is finding answer for the following questions:

- Is it possible to generate MHT and track pedestrians before camera detection?

- Can an external source of information change the performance of the hypothesis tree?

- How can fusion of IoT and on-board sensors decrease the side effects of occlusion and improve the performance of a pedestrian tracker?

In order to find answers, the following scenarios are considered:

- Scenario 1: To answer the first question, this scenario is simulated. Before a vehicle enters a road that could not be observed before, its camera could not detect anything. It means the camera has no information about the new area. Fig.4.1 illustrates this situation. In this situation, the tracker can receive data through IoT connections.

- Scenario 2: To answer the two other questions, this scenario is simulated. During a turn or moving on a straight line with a constant velocity, the vehicle's camera detects pedestrians partially. It means that pedestrians are not entirely in the camera's field of view, or other objects in the scene such as other vehicles, cyclists, pedestrians occlude pedestrians partially. Fig.4.1 illustrates an example of this situation. In this situation, the tracker can receive data through IoT connections and measurements from a camera.

Figure 4.1: In the situation number 1,the vehicle has no vision data only IoT data. In the situation number 2, the vehicle receives IoT data and a partial measurement from its camera. The arrow shows vehicle's moving direction.

The assumptions for these simulations are as follows:

- Vehicle speed is constant and it is 15 $(km/h)$.
- The right amount of displacement and rotation of the vehicle are available.
- Pedestrian speed is varied in the range of 0-1.2 $(m/s)$.
- The number of pedestrians was varied between 1-30.
- The threshold for the Manhattan distance is 2 $(m)$.
- The camera measurement's noise is 0.5 $(m)$ in all the scenarios.
- To find a relation between IoT, on-board sensors and the proposed tracker's performance, in the simulations, it is assumed that there is no false detections.

### 4.5.2 Scenario 1

This scenario shows that a hypothesis tree can be generated or initialized using IoT devices where it is impossible for on-board sensors to detect pedestrians. Therefore, IoT measurements are used to initialize objects. As mentioned before, the measurements are IDs, accelerations, and geographical coordinates.

In this scenario, a vehicle receives data from nearby IoT devices; then, an MHT with several branches is created. Each branch could contain multiple hypotheses, and it forms with different probabilities of possible associations for IoT devices. As a result of assuming no false detections, these probabilities show that a device is new or exists. Based on the probabilities, each hypothesis can explain the displacements and IDs of IoT devices differently. In this scenario, the probability of new pedestrians is equal to one $(P_{new} = 1)$, if the pedestrian was not and could not have been observed before. Basically, if an ID was detected in a previous time, the $P_(exist)$ of that device is equal to one.

In the simulation environment, the number and displacement of pedestrians are given as ground truth data. Therefore, the tracker's accuracy is estimated using the ground truth and the

Table 4.1: A fixed range of noise for each test case

| Case | a standard deviation(m) | Case | a standard deviation(m) |
|---|---|---|---|
| 1 | [0,0.5] | 3 | [1,1.5] |
| 2 | [0.5,1] | 4 | [1.5,2] |

best hypothesis. The best hypothesis mean selecting the hypothesis with the most probable explanation about IoT devices. In this scenario, the following steps are considered:

- After an intersection, a rectangle (length 25 (m) and width 7 (m)) is defined. The size of rectangle is selected based on the safety reasons.
- Through wireless connection, the vehicle receives data of IoT devices that are inside of the rectangle.
- IoT measurements are received 1 $Hz$.
- Duration of each simulation is 7 sec. Based on the pedestrians speed range, it assumes that during 7 sec they have enough time to leave the rectangle.
- This scenario is simulated in five cases and each case is repeated for 25 times. In each case, IoT devices have a fixed range of measurement noise, Table 4.1 defines the range of the noise.

Based on the average MOTA in each case, Fig.4.2 indicates that there is a possibility that a tracker becomes aware of pedestrians before on-board sensors detection. In the first three cases, the accuracy is close to 100%, which means false hypotheses have less chance to be selected as correct hypotheses. Thus, the tree can grow and manage similar to the ground truth data. Therefore, there is a possibility to update the tree after receiving the first on-broad sensors detection without initializing the tree. In the last case, the accuracy is around 60%, which means the hypotheses can be wrongly selected. Therefore, the MHT should be initialized as soon as the on-broad sensors detect pedestrians. Although the accuracy of 60% may be sufficient in the real world, in simulations with ideal situations and no false detections is insufficient.

As mentioned in this work, a linear model is used to estimate the displacement of IoT devices. Suppose the tracker receives data from multiple IoT devices with the maximum amount of noise. In that case, there is a probability that the MHT makes a wrong association between IDs and displacements. Moreover, in each hypothesis, only pedestrians inside of the rectangle should be considered. Therefore, as a result of the noise and linear model, there is a probability that a pedestrian left a rectangle. Although, the position of the IoT device is still inside the rectangle. To conclude this scenario, the results indicate that generating MHT and tracking pedestrians without on-board sensors are a function of the accuracy of IoT devices. The following scenario shows the effect of generating the MHT before on-board sensors detection on the performance of entire tracks.

### *4.5.3 Scenario 2*

This scenario aims to answer the two other questions and show the added value of IoT in vision-based trackers. Therefore, tracker in Chapter 2 and the proposed tracker are used. This scenario

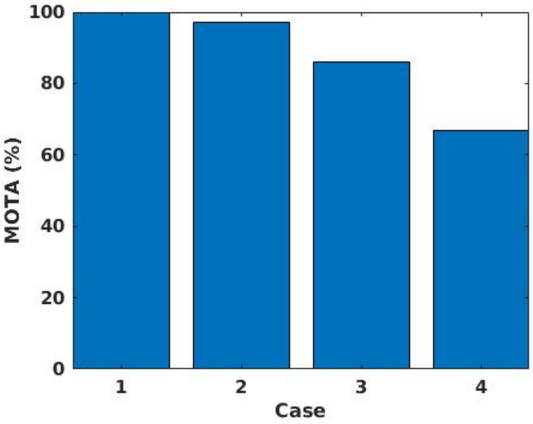


Figure 4.2: Average MOTA for each test case

occurs when a vehicle turns at an intersection or moves in a straight route. Then, both trackers face complete and partial detections. Partial detections mean that on-board sensors detect pedestrians but as a result of occlusion, sensors are unable to detect the whole body. In this situation, although Chapter 2 needs leg measurement to estimate and predict the position of pedestrians related to the vehicle, the proposed tracker can use IoT and image sequences to increase the accuracy of detection and data association. Similar to scenario one, a rectangle in a vehicle frame is defined. It means the center of the rectangle is in the center of the bumper in front of the vehicle.

- Each case repeats 100 times. Each case contains a different number of pedestrians and variable IoT noises. For 50 times in each case, the vehicle moves with a constant speed, and 50 times, the vehicle turns an intersection.

- Pedestrians walk differently, such as walk laterally or longitudinally related to the vehicle or in diagonal lines.

- The duration of each test is 6 seconds.

- The IoT updates its measurements 1 $(Hz)$ while the camera update rate is 10 $(Hz)$.

As mentioned before, MHT uses different probabilities to compute the posterior probabilities of all hypotheses given the measurements. In Chapter 2, these probabilities are constant. However, different parameters such as weather conditions and the number of pedestrians can change the probabilities. Besides, to initialize MHT, Chapter 2 should skip frames; this work uses IoT data to update the probabilities. In this scenario, similar to the previous one, the MHT is created based on IoT devices; then, after the first camera detections, the tree is updated. It means that if no camera measurement is compatible with one of the existing hypotheses, a new hypothesis or a clutter (a false detection) should be formed. Table.4.2 shows the results of this scenario.

In Table.4.2, one reason that the accuracy of the proposed tracker is better than Chapter 2 is skipping fewer measurements to ini-

Table 4.2: Average MOTA(%) in the proposed tracker with both camera images and IoT devices in the simulation environment.

| A standard deviation (m) | MOTA(IoT+Camera) | MOTA(Camera) |
|---|---|---|
| 0-1 | 80 | 65 |
| 1-1.5 | 73 | 65 |
| 1.5-2 | 57(IoT 1 $(Hz)$) | 65 |
| 1.5-2 | 70.81(IoT 10 $(Hz)$) | 65 |

tialize the tree. In Chapter 2, they should skip frames to collect sufficient prior knowledge about the surrounding environment. In the same situation, the proposed tracker does not require collecting prior knowledge after on-broad sensors detections. It means that initializing the data association with camera measurements and IoT data can help to select the correct hypotheses. Moreover, the results show that IoT can help robustness. Therefore, a tracker can decrease the number of false tracks.

The update rate of IoT is slower than the camera. This different update rate can cause an error. Therefore, in one case, the accuracy of the proposed tracker is smaller than Chapter 2. As an example, a pedestrian is walking in a non-straight line. Therefore the movement of a pedestrian may not be linear. In this situation, if the tracker estimates the displacement with a linear model, more measurements in a shorter time are required. Different update rates and nonlinear movement of pedestrians are the reasons to have the lowest MOTA for pedestrians with IoT noise in a range of 1.5-2 meters. To be confident that having a different and lower update rate of IoT compared to the camera is one of the reasons; the update rate of IoT devices is changed. The same update rate of 10 $Hz$ for both the camera and IoT is utilized in the simulation environment. The last row of the table shows the result. This result shows that the performance of the tracker can be improved by increasing the update rate of IoT devices. In fact, data association will be improved if IoT data are received with a higher update rate.

### 4.5.4 Experiment

This section evaluates the accuracy of the proposed tracker in an experimental situations. The experimental situations are same as the last simulation scenario. In these experiments, pedestrians were asked to carry a smart device. The smart devices send their acceleration and position to the vehicle through 4G connections. Fig.4.3 illustrates the test scenario.

In this part, on-board sensor is an RGB camera. During the test day, on-board sensors were able to detect multiple joints of pedestrians at 10 $(Hz)$ with a maximum distance of 25 meters from the on-board sensors. Fig.4.4 shows the output of sensors. The testing process is as follow:

- A pedestrian (blue icon) sends a taxi request. At that time, an autonomous vehicle is in the “Start” location.
- When the vehicle starts to drive, two pedestrians that one of them occludes another one,cross the road in front of the vehicle (green icon). Each of these two pedestrians has a IoT device.
- After they cross the road, the vehicle continues to drive.

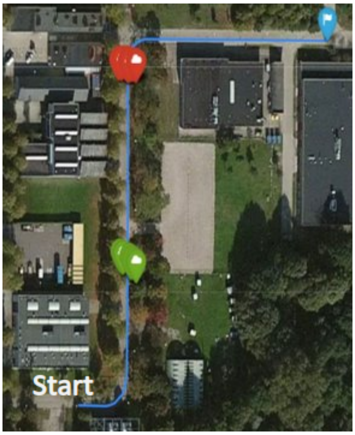


Figure 4.3: A schematic of experimental tests, a vehicle should start driving autonomously from start point to the blue flag.

The vehicle's vision sensors detects two other pedestrians perfectly (red icon) while they are crossing the road. Two pedestrians have a different IoT device.

- After they cross the road, the vehicle continues to drive to meet a pedestrian in the blue icon.

During the test, the tracker faces various environmental noises that were not considered in the simulation. These environmental noises include weather conditions, the amount of charge in the IoT devices, and road structure. These noises can affect the connection between the IoT device, its server, and the vehicle. Fig.4.5 illustrates the quality of an image during weather conditions.

Based on the IoT, the hypothesis tree generates a tree with two new pedestrians. After sensors detection, if the sensors detect fewer pedestrians than IoT, the tree keeps the hypothesis of existing a pedestrian. It means that there is a high probability that on-board sensors miss a pedestrian. Therefore, the tracker knows there is a pedestrian that the on-board sensors could not detect. As soon as the sensors detect another pedestrian, the tracker can provide the position and displacement. These ex-

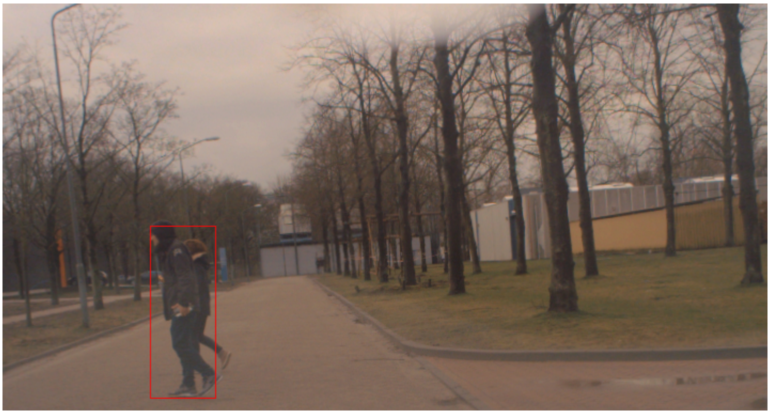


Figure 4.4: One pedestrian occludes another pedestrian. Therefore camera is not able to detect one of them, while both of them have IoT devices.

Table 4.3: Difference between average MOTA in the two different cases.

| Testing scenario | Average MOTA (%) |
|---|---|
| Track pedestrians using image sequences without IoT (Chapter 2) | 64 |
| Track pedestrians using image sequences and IoT | 75 |

periments were recorded 25 times. The accuracy of the tracker are evaluated in the following situations:

- Only the on-board sensors are used to estimate the accuracy.
- The data of IoT and on-board sensors are used. The probabilities of MHT updates based on IoT data.

Table.4.3 shows that among two cases, the accuracy of the tracker can be improved if a tracker uses both sources of data and updates the probabilities. The reason for this improvement is that the tracker can generate multiple hypotheses before detection. It means that if the tracker receives a false detection, it can generate a hypothesis and correct itself after receiving the measurements in the next timestamps. Moreover, updating the probabilities helps the data association match the vision sensors' data

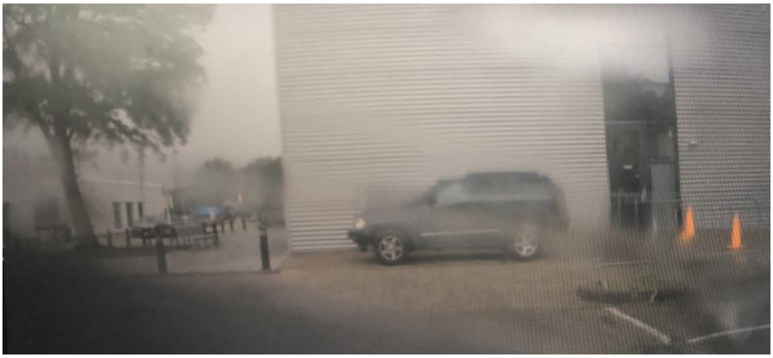


Figure 4.5: One of the environmental noises that our vehicle faced during tests.

without skipping measurements. Therefore, the probability of missing pedestrians is low.

## 4.6 Conclusion

In recent years, several pedestrian trackers have been developed to help awareness systems and improve pedestrian safety. The current existing trackers usually rely on multiple on-board sensors. In reality, the large number of objects such as cyclists, pedestrians, other vehicles can block sensors' field of view. As a result, on-board sensors are unable to track pedestrians in their blind detection area. To overcome this problem, this work proposes a pedestrian tracker. This tracker uses on-board sensors and IoT techniques and generates a hypothesis-based data association algorithm before detecting pedestrians with on-board sensors. Therefore, this tracker is aware of unseen or semi-occluded pedestrians. Moreover, in this work, constant parameters in a hypothesis-based data association algorithm are updated. This work shows that updating parameters can minimize the delay between the first detection and selecting the correct hypothesis.

# Chapter 5

# Intention Prediction and Action Recognition of Pedestrians

**Abstract:** Intention prediction and action recognition are two critical tasks to drive safely and smoothly. Pedestrian safety could be significantly improved if cars could predict and recognize each pedestrian's intentions and actions. This work proposes a framework that recognizes the current action and predicts intention. To recognize action, this work proposes a set of body features that are distinctive among pedestrian actions. The action recognizer of the proposed framework has been trained with the Human Gait Database (HuGaDB) and the MAREA gait database. With respect to intention prediction, this work focuses on the intention of crossing/not crossing in front of the car. We tackle the intention prediction by observing pedestrians' distance to the car, action, and using available spatio-temporal context information such as traffic signs, environmental factors, zebra-crossings, pedestrians' occlusions with elements in the scene, their gaze information,their hand gesture, and weather conditions. As a proof of concept, this work compares this framework with state-of-the-art methods. The Joint Attention for Autonomous Driving (JAAD) dataset is used to validate this framework. The results indicate that this framework can recognize the action with an accuracy of 75%. The action could be walking, standing, staring, and stopping. We predict the final intention of pedestrians in the JAAD dataset and report an average 91%

accuracy.

Dolatabadi, M., Elfring, J., Aboutalebian, B. , Van de Molengraft, R. (2021). Intention Prediction and Action Recognition of Pedestrians Using Body Features and Contextual Information In Automotive Applications. Robotics and Autonomous Systems, Submitted.

## 5.1 Introduction

According to an article published by European mobility and transport, pedestrians are much more vulnerable to accidents than other road users [Commission 2017]. Based on the world health organization (WHO), more than one-fifth of road traffic deaths worldwide are pedestrians [Organization et al. 2018]. Due to the high number of fatalities, several studies have been done on creating a system that can offer more safety for pedestrians [Kwak et al. 2016; Völz et al. 2015].

One of the main challenges in transportation systems is an understanding of pedestrians' behaviors [Rasouli and Tsotsos 2019]. As shown in [Quintero et al. 2017; Schulz and Stiefelhagen 2015], non-verbal communication with pedestrians can improve braking systems' performance. Mainly because of the high variability of movement, pedestrians can change their actions quickly. For instance, they can suddenly start or stop walking. Action recognition and intention prediction can be used to predict human motion and may help to have safe and comfortable driving [Fang and López 2018; Ferguson et al. 2015; Gu, Hashimoto, et al. 2016; Rudenko et al. 2020].

Intention predictors can predict the intentions based on the most recent estimated action [Li, Zhang, et al. 2018]. For example, suppose a car receives the position of a pedestrian near a curb of a sidewalk. In that case, using the most recent action of a pedestrian such as walking, running, or standing, the car can predict intention. For instance, the car will predict that the pedestrian is stepping onto the road and will cross. Therefore, the vehicle can start to slow down to reduce the risk of a collision [Fang and López 2018]. As a result of uncertainties regarding pedestrians' impending motion, the pedestrians' intention prediction and action recognition are not trivial tasks [Ferguson et al. 2015]. In this work, the actions are considered as walking, starting to walk, standing, and stopping. Using the motion of pedestrians, [Rasouli, Kotseruba, et al. 2017a] predicts intention at the first moment pedestrians are assessing the environment and expressing their crossing or not crossing intention. In contrast with [Rasouli, Kotseruba, et al. 2017a], this research believes that intention is not a moment in time. Moreover, this work would not expect that pedestrians actively inform an approaching car about their intention. As an example, suppose pedestrians intend to cross a street where there is no priority sign. Only in an ideal case may pedestrians use body posture or establish eye contact with an approaching car to ensure that the car is aware of them. Therefore, in this work, intention is defined as crossing or not-crossing of pedestrians on a piece of road that is used by the car.

Most of the works on this topic [Chaabane et al. 2020; Fang and López 2019; Saleh et al. 2019a; Schneemann and Heinemann 2016; Wang and Papanikolopoulos 2020] have tackled the action or the intention problem by observing features such as position and velocity in a single point on the body of the pedestrian. Based on [Sztyler et al. 2017], a different part of the body provides information regarding the action or intention of pedestrians. Therefore, only considering the variation of the features, in a single point, can significantly degrade the performance of an action recognizer or an intention predictor or lead to their failure. Thus, this chapter decides to use multiple points in the body and extract distinctive features among pedestrian actions and intentions.

The focus in this work is proposing a framework that will recognize the action and will predict intention. The framework performs based on a series of features using available measurements. The measurements are information on spatio-temporal contexts such as traffic signs, traffic flow, and other environmental features, pedestrian gaze status, hand gesture, occlusion information, and weather conditions. As a result of this contexts information and environmental conditions, pedestrians may change their behavior at the curb before crossing. Therefore, using the current action of pedestrians, their intentions are not always predictable. This framework shows that a combination of actions and contextual information improves intention prediction compared to prediction based on the pedestrians' current actions or only contextual information. Contributions of this work are as follows:

- Predict the current action of pedestrians, variation in legs' frequency and variation in distance of pedestrians from an approaching car are used.
- Combine the recognized actions, hand gestures, weather conditions, and context information to predict crossing intention.
- Show effects of a sudden change of behavior in intention prediction.
- Similar to [Afolabi et al. 2018; Bouhsain et al. 2020; Gujjar and Vaughan 2019; Liu, Adeli, et al. 2020; Marginean et al. 2019; Pop et al. 2019; Rasouli, Kotseruba, et al. 2018; Styles et al. 2019; Yang, Zhang, et al. 2021], this chapter reports results for the Autonomous Driving dataset (JAAD) dataset [Rasouli, Kotseruba, et al. 2017b]. This dataset allows us to address the intention and action classifications in realistic driving conditions. This work compares the accuracy of the intention prediction and action recognition with methods that used the same dataset.

The rest of this paper is arranged as follows: Section II gives an overview of the state-of-the-art in the area of pedestrian action recognition and intention prediction. Section III introduces the framework's architecture. Sections IV and V describe how this work recognize actions and predict intentions. Section VI explains the classifier. Section VII contains the evaluation procedure and validation. Section VIII presents conclusions and outlines future directions in this research.

## 5.2 Related work

This section discusses the state-of-the-art to recognize the action and predict pedestrian intention. It should be noted that numerous works address this problem; in [Chen, Ma, et al. 2020; Hou et al. 2021; Rasouli 2020], the authors provide a complete survey regarding human intention prediction, mainly for indoor robots. In automotive applications, due to factors such as illumination conditions, clutter backgrounds, and occlusions, the recognition of actions and predicting intention are challenging [Marginean et al. 2019].

Based on a theory of mind research, there is a hypothesis that others people's intentions are known by observing their actions [Blakemore and Decety 2001]. According to this hypothesis, intention prediction can be improved by using the current action. For example, a pedestrian near a street with the action of walking is more likely to cross than one with the action standing still. In this work, this hypothesis is validated.

A group of researchers in this field use pedestrians' silhouettes and a feature-based approach in monocular videos [Angelini et al. 2018; Jalal et al. 2019; Klinger and Arens 2009; Marzoli et al. 2019]. They usually extract their measurements and use them in a ragdoll model to recognize the action or intention. The major drawback of ragdoll models is that they cannot predict stop and start actions based on images of a camera on a moving car.

Several recent works [Afolabi et al. 2018; Bouhsain et al. 2020; Chaabane et al. 2020; Fang and López 2019; Gujjar and Vaughan 2019; Styles et al. 2019] tackled the action or intention problem by observing a variation in the position, velocity, head orientation,or gesture. All these works use different state estimators to estimate their feature. Then, to predict intention or action, they use a feature. As cited in [Schulz and Stiefelhagen 2015], a feature alone is not particularly useful for pedestrians who intend to stop or cross the road. Therefore, in order to be conclusive in more scenarios, adding different types of features is recommended. Besides, as [Bassett et al. 2017] has demonstrated, the velocity variation of pedestrians are accurate at speeds of 1.34 $(m/s)$ and above, but at 0.44 $(m/s)$ is hard to detect the variation of velocity for a camera-based setup on a moving car. In [Danion et al. 2003], they confirm that only focusing on the effect of walking speed may not enough to predict action of a pedestrian.Thus, another group of researchers add features from images or uses prior knowledge.

To learn intention of pedestrians from images, recent works [Goldhammer et al. 2019; Kotseruba et al. 2020; Kwak et al. 2016; Pop et al. 2019; Wakim et al. 2004] extract features from sequence of images, then they utilize different neural networks and video-based motion classifiers to predict the future movement of vulnerable road users (VRUs). They have a bounding box around each VRU. Usually, they predict intention or action based on the center of bounding boxes. Although these algorithms perform well, there is a possibility that they could not predict the intention of a pedestrian as a result of occlusion. The main problem is that estimating a position using a single point may lead to inaccurate position and velocity estimates due to the fact the center of a bounding box does not always represent the same body part. It can even be that the center is not part of the pedestrian but part of the background. As shown in [Yan et al. 2004], during occlusions, the center of the bounding box is not always enough.

To increase the accuracy of action recognition and intention prediction, a group of works uses measurement such as body-pose key-points in multiple image sequences [Cadena et al. 2019; Piccoli et al. 2020; Wang and Papanikolopoulos 2020]. They extract features from body key-points to predict the crossing/not-crossing intention. To extract the features, they use different motion estimation models or neural networks. Then, features from body key-points are used as input to a classifier, such as support vector machine (SVM), Hidden Markov model (HMM), or Random Forest. Next, the classifier provides a probability of crossing/not-crossing intention. Although they extract multiple key-points, they only use the most stable ones. Based on [Fang and López 2018], the key-points belong to shoulders, and legs are the most stable compared to other key-points. However, [Fukuda et al. 2021] shows during walking, arm movements are more flexible than leg movements. Therefore, arm movements cannot be repeatedly acquired action or intention similar to leg movements. [Zhang, Abdel-Aty, et al. 2021] shows that an important variable for predicting intention is the movement of the angle between knee and ankle. Hence, compared to other body joints, the leg movements can provide more information regarding pedestrian action or intention.

The legs execute continue/start walking or stopping/standing actions. Moreover, [Yan et al. 2004] shows that when pedestrians are preparing themself to walk, their legs have flexion and extension. However, the center of their body has no movement. It means that the velocity of the center of the body is tending to zero. However, the frequencies of the legs are not equal to each other. The legs are the first moving parts that start/stop moving when a pedestrian walks/stops. For that reason, it makes sense to look at the positions of legs. Besides, ankle/knee are well-defined parts of the body. As an example, suppose a pedestrian wants to cross the road similar to Fig.5.1. In Fig.5.1, based on the frequencies of the legs, there is a probability that she prepares herself to walk.

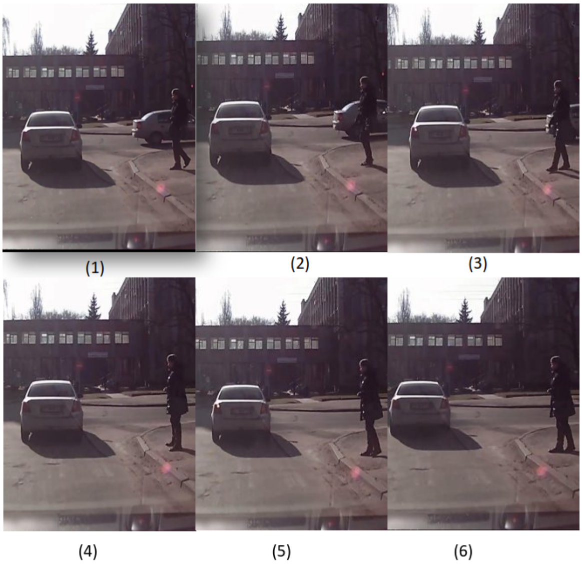


Figure 5.1: Pedestrian is walking in place and is preparing herself to walk. The number below each image indicates its sequence order [Rasouli, Kotseruba, et al. 2017b].

Although using body keypoints we can predict the intention, adding even more information may help improve performance even further. To use more information regarding pedestrians, in [Liu, Adeli, et al. 2020], they investigate the effects of crosswalks and sign information on the intention of pedestrians. Similar to us, they believe that the intention prediction problem from image sequences can be improved by using environmental information. Therefore, they consider crosswalk and sign information and the variation of the 2D position of the center of the pedestrians' bounding boxes in pixel. Although the current action of pedestrians can be vital to predicting their intention, [Liu, Adeli, et al. 2020] does not consider the effect of actions explicitly. [Marginean et al. 2019] uses a set of features based on pose estimations and context information from sequences of images to predict intention. They define different weights for all relevant and irrelevant features and train their model based on the defined weights. Based on their conclusion, using the weights can cause an imbalance effect. In [Yang, Zhan, et al. 2021], they use a 3D convolutional neural network (CNN) and context information such as zebra and stop signs to recognize intention of pedestrians. Based on their conclusion, their framework cannot be used for real-time applications since the introduction of their high computational load neural network.

From this discussion, it is clear that it is possible to use more information to improve the performance of intention prediction. Therefore, this work utilizes features in sequences of images. Then, the features are used as input for a classifier. To extract features, legs are tracked, since ankle, knee, hip are well defined parts of the human body. Moreover, context information, and weather conditions are used.

To recognize action, [Lei et al. 2016] combines the CNN network's feature learning ability and the HMM model's sequence dynamic modeling ability. Based on [Lei et al. 2016], for weakly labeled action, HMM model can be used as label information to train the CNN network. [Wu, Song, et al. 2020] also shows that HMM has good performance in action recognition using arm movements. [Gu, Liu, et al. 2021] shows that in the decision-level fusion method, the computational cost and training times of HMMs are less than deep learning methods. Besides, [Wissel et al. 2013] shows that the HMM does not need to await further samples for a decision. It means that HMM has an implicit adaptation in real-time applications. Therefore, this research uses HMM to classify action and intention.

One of the contributions of this work is showing that the combination of the action, weather condition, hand gesture, the context information, and gaze status can improve the performance of the intention predictor. Therefore, we recognize actions as an explicit step in the intention prediction framework. To recognize action, we use the body-key points. Then, we extract frequencies of two legs as features to recognize the action. Then, based on the features, we train our classifier to recognize the actions of pedestrians. To classify intention as crossing or not-crossing in front of our car, we used five types of measurements: the actions, context information, gaze status, hand gesture, and weather conditions. It should be noted that extracting the context information, gaze status from images are not trivial tasks. Extracting such information is out of the scope of this work. Getting this information from camera images can be done using existing algorithms. Therefore, we use the context information, gaze status, hand gesture, and weather conditions that the JAAD dataset provides.

## 5.3 Framework architecture

This work is focused on pedestrians in urban traffic environments. Fig.5.2 illustrates a block diagram of this framework. Although in the following sections blocks are described, here the most important components in the framework and their relations are defined:

- After detecting a pedestrian in a bounding box, Openpose library [Cao et al. 2017b] is used to extract body key points. The body keypoints belong to left hip, left knee, left ankle, right hip, right knee, and right ankle.
- Having access to tracked pedestrian positions is a prereq-

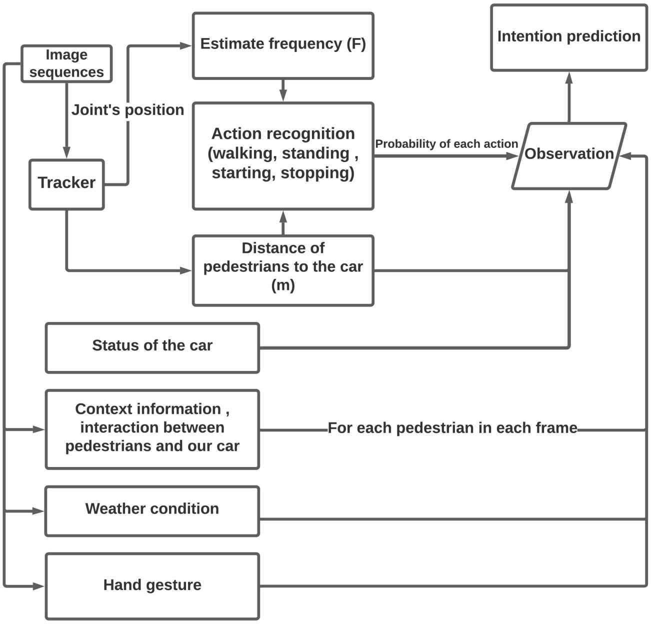


Image sequences
Joint's position
Tracker
Estimate frequency (F)
Action recognition (walking, standing , starting, stopping)
Probability of each action
Observation
Intention prediction
Distance of pedestrians to the car (m)
Status of the car
Context information , interaction between pedestrians and our car
For each pedestrian in each frame
Weather condition
Hand gesture

Figure 5.2: Block diagram of this framework.

uisite for this work, however, it is outside the scope of this paper. For that reason, the approach described in Chapter 2 is adopted.

- Based on body keypoints, Chapter 2 estimates the position, the velocity, and legs' frequency of pedestrians.

- If the library fails to extract all key-points, the center of bounding box around each pedestrian is used to estimate the distance between pedestrians and an approaching car.

- In this work, the position of pedestrian is estimated based on the right and left hip.

- Based on a variation of the leg frequencies and position of pedestrian, a probability for each action is estimated.

- Knee is the first joint that moves at the beginning of a gait cycle. Therefore, to compute the variation of stride frequency, the knee (left and right) frequencies are used.

- The frequency needs frames to converge as soon as a detector detects a pedestrian. Moreover, different parameters such clothes or leg occlusion can affect the frequency estimation. Therefore, position of pedestrian is also used to recognize action. More details about the effects of the convergence speed and its impact on estimation results will be given later.
- Action recognition computes the probabilities for each action.
- To predict the intention, a combination of action, the distance of pedestrians to the car, gaze status, hand gesture, weather condition, contextual information, and status of the car in each frame are used.
- Intention are predicted as crossing or not crossing in front of a car.
- The reasons that this work uses actions instead of the variation of the velocity are as follows:
  - Estimating the accurate velocity variation is very challenging, especially if pedestrian velocities are low.
  - This work assumes that for estimating intentions, a precise velocity is not needed. Moreover, it assumes that knowing how a pedestrian moves on a coarse scale is sufficient to predict the intention.

## 5.4 Action

[Goldhammer et al. 2019] shows that the following four actions are helpful to prevent collision with car and estimate the trajectory of VRUs. Therefore, this work uses the same action as [Goldhammer et al. 2019].

1. Standing. It indicates that a pedestrian stands at a fixed position while he/she might move the upper body.
2. Starting. The first step after standing when the knee starts to bend until a pedestrian reaches his/her constant walk-

ing velocity. It means starting can be occur in multiple frames.

3. Stopping. It shows that based on previous sequences, the pedestrian is tending to a standstill. The last frames before standing that the frequency of the leg is not constant but decreasing is stopping.

4. Walking. It is defined as pedestrian movements between the last starting action and the first stopping action.

To recognize action, this work focuses on body features that are distinctive among pedestrian actions. The measurement is as follows:

- Derivative of stride frequency: This work defines the stride frequency as the number of steps that a pedestrian takes in a second. This work uses derivative of the stride frequency in each frame compared to the previous frame.

- Variation of distance : The tracker estimates the position of multiple joints related to an approaching car. Variation of the position in each frame is used to recognize action. In this work, the distance of a pedestrian is defined based on the middle of hip joints. As shown in Fig.5.3, this pedestrian carries a shopping trolley, and the trolley occludes her legs. In this kind of situation, the frequency estimation has inaccurate results. Therefore, without knowing the pedestrian's distance variation, the algorithm cannot recognize the actions correctly.

After estimating the frequencies and distance, they are used to recognize the action.

### *5.4.1 Action recognition*

Fig.5.4 depicts the state transitions. As Fig.5.4 shows, the transitions between standing and walking are always separated by starting or stopping. These states are not directly observable by a detector.

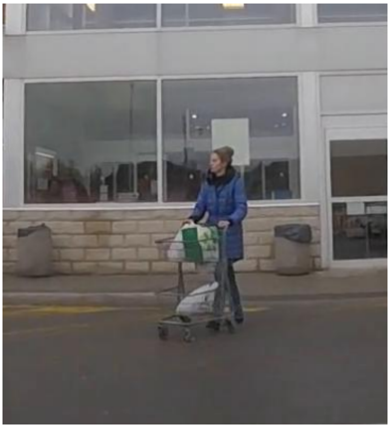


Figure 5.3: Due to occlusion, there is wrong information regarding the frequency [Rasouli, Kotseruba, et al. 2017b].

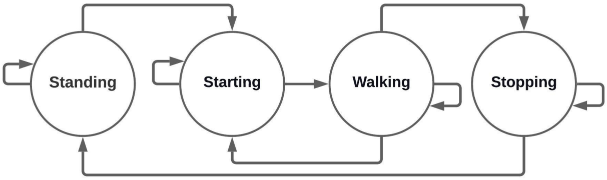
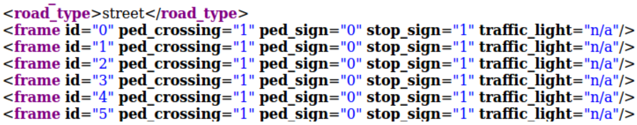


Figure 5.4: State diagram for modeling pedestrian actions.

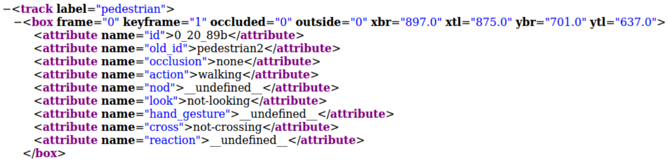
## 5.5 Intention

Traffic information and the environmental information of pedestrians are helpful for an improved intention prediction performance. As shown in Fig.5.2, to predict intention and actions, available information from the traffic scenes where pedestrians locate is used. This work uses JAAD dataset [Rasouli, Kotseruba, et al. 2017b] to get access to the traffic scene information. [Rasouli, Kotseruba, et al. 2017b] contains a large number of pedestrian samples with temporal correspondence, a subset of which are annotated with behavior information by algorithms. This part describes the last two blocks and starts with the context information, gaze situation, and occlusion. The following input features are used mainly because they contain global context



```
<road_type>street</road_type>
<frame id="0" ped_crossing="1" ped_sign="0" stop_sign="1" traffic_light="n/a"/>
<frame id="1" ped_crossing="1" ped_sign="0" stop_sign="1" traffic_light="n/a"/>
<frame id="2" ped_crossing="1" ped_sign="0" stop_sign="1" traffic_light="n/a"/>
<frame id="3" ped_crossing="1" ped_sign="0" stop_sign="1" traffic_light="n/a"/>
<frame id="4" ped_crossing="1" ped_sign="0" stop_sign="1" traffic_light="n/a"/>
<frame id="5" ped_crossing="1" ped_sign="0" stop_sign="1" traffic_light="n/a"/>
```

Figure 5.5: Information regarding traffic scene [Rasouli, Kotseruba, et al. 2017b].



```
–<track label="pedestrian">
  –<box frame="0" keyframe="1" occluded="0" outside="0" xbr="897.0" xtl="875.0" ybr="701.0" ytl="637.0">
      <attribute name="id">0_20_89b</attribute>
      <attribute name="old_id">pedestrian2</attribute>
      <attribute name="occlusion">none</attribute>
      <attribute name="action">walking</attribute>
      <attribute name="nod">__undefined__</attribute>
      <attribute name="look">not-looking</attribute>
      <attribute name="hand_gesture">__undefined__</attribute>
      <attribute name="cross">not-crossing</attribute>
      <attribute name="reaction">__undefined__</attribute>
    </box>
```

Figure 5.6: Information regarding pedestrians annotation in the JAAD dataset [Rasouli, Kotseruba, et al. 2017b].

information. The global context indicates the semantic segmentation of road and pedestrians and vehicles [Yang, Zhang, et al. 2021]. This block contains the following measurements that are input to the proposed model:

- A tag for a car actions, such as moving fast, speeding up, moving slow, decelerating, stop in each frame.

- Information regarding status of traffic signs includes availability of car stop sign, pedestrian stop sign, pedestrian crossing area for each frame. These annotations are one per image frame. Fig.5.5 shows a sample of this input.

- The measurement also includes looking and occlusion information. Looking shows that is a pedestrian looking towards the host vehicle or not. Occlusion indicates that a pedestrian is occluded or not. The occlusion values are either 0 (no occlusion), 1 (partial occlusion >25%) or 2 (full occlusion >75%). This information is one per frame per pedestrian. Fig.5.6 shows a sample of pedestrians annotation in the dataset.

- To associate looking, occlusion information, and hand gestures with pedestrians, ground truth bounding boxes are used as an input to OpenPose.

- Moreover, the measurement includes hand gestures and weather conditions.

## 5.6 HMM

To estimate the action and predict intention, an HMM is used. Based on Markov assumptions, the hidden states are only dependent upon previous states, and all observations are independent given the state. Therefore, HMM is one of the suitable classifiers to classify hidden states based on a time sequence [Razin et al. 2017]. Recognizing action is one of the contributions. This work shows that using the current action, intention can be predicted more accurately than only using context information. To make sure that the action part is accurate, first, this work recognizes action separately and evaluate the performance of this intermediate step. Then, actions and the intention estimation are validated. HMM includes hidden states, features, a state transition probability matrix, and an emission probability matrix.

### 5.6.1 Measurement and state

Before using HMM, the hidden states and measurements are defined. During recognizing action separately the hidden states are walking, starting, stopping, and standing. For intention prediction the state vector contains

- Intention: Crossing, not-crossing in front of our car.

To estimate the hidden states, the following measurements are utilized:

- Action: The rate of change in frequency.
- Intention: The current action, pedestrians' distance to the car, context information, looking and occlusion information.

There are transition probabilities between the hidden states. These probabilities are shown as a transition matrix with $(i, j)$th element. Based on the definitions in [Rabiner 1989], each transition probability from state $s_i$ to state $s_j$ is defined by:

$$a_{s_i j} = \frac{number\ of\ times\ state\ s_j\ follows\ state\ s_i}{number\ of\ times\ state\ s_j\ occurs} \tag{5.1}$$

Where $s$ means hidden state and $k$ indicates a time instant. The probability of correctly classifying $s$ is maximized by choosing a sequence of measurement ($x$ ) that has a maximum posterior probability. A sequence of measurement likelihoods or emission probabilities expresses the probability of a measurement ($x$ ) generated from a hidden state $s$. The emission probability in state $s_j$ is formulated as [Rabiner 1989]:

$$b_{s_j}(k) = \frac{number\ of\ times\ in\ state\ s_j\ and\ observing\ x}{expected\ number\ of\ times\ in\ state\ s_j} \tag{5.2}$$

The initial probability is assumed to be uniformly distributed because this work does not know the hidden states in $k = 1$. The initial state distributions is defined as:

$$\pi_j = P(q_1 = S_j), 1 \leq j \leq N. \tag{5.3}$$

$N$ is the number of states. A HMM can be characterized by a triplet $\lambda = (a_{s_{ij}}, b_{s_j}(k), \pi_j)$.

### 5.6.2 Prediction

To predict, the maximum likelihood probability of the hidden state is computed. To have the maximum likelihood, the Viterbi algorithm is used. The Viterbi algorithm finds the most likely state sequence in the maximum a posterior probability [Johnson et al. 2010]. The goal of the Viterbi algorithm is to make an inference based on a trained model and measurements. To compute the maximum likelihood and all HMM training, the standard approach available in [Johnson et al. 2010; Rabiner 1989] is used.

### 5.6.3 Training HMM

To confidently use the HMM for action recognition or intention prediction, the HMM is trained to represent measurements accurately. Training determines optimal estimates for $\lambda$. This work uses the Baum-Welch algorithm to estimate the transition and emission probabilities [Sammut and Webb 2010]. Baum-Welch algorithm starts with an initial estimate of $\lambda$ and a set of training sequences $L$. It then updates the estimates of transition and emission by calculating forward and backward probabilities at each iteration.

To train HMM in the action part, three datasets are used [Chereshnev and Kertész-Farkas 2017; Khandelwal and Wickström 2017; Luo et al. 2020]. These datasets provide detailed gait data of the legs in indoor and outdoor environments for healthy people. Using the gait data, the features are estimated. In these datasets, they use wearable sensors on the right and the left thigh, shin, and foot. It means that the matrices can be trained without any occlusion or incorrect data association. The datasets provide the height, gender, and age of all participates. As ground truth (GT) data, these datasets provide action labels. Therefore, based on the sensory data and GT that are provided by the datasets, the action recognizer is trained. The state vector in the action part contains walking, standing, starting, and stopping. These datasets provide gait information of people; therefore, the frequency is computed. Then, each time by receiving new measurements from the tracker, the state transition matrix and the emission matrix are updated.

To train the HMM in the intention part, this work uses the first 70% of the JAAD dataset. The JAAD dataset contains various complex interactions and situations that may impact the traffic participants' behavior. The JAAD dataset recorded over 240 hours of driving footage. The dataset was recorded in six months in North America and Eastern Europe [Rasouli, Kotseruba, et al. 2017b]. They use Convolutional Neural Networks (CNNs) to detect and analyze the context of the scenes and pedestrians' behavioral cues. The JAAD dataset has temporal correspondences between the frames, and each pedestrian has a unique id throughout the sequence. These IDs are used during the training part and validation part. As mentioned before, crossing means moving on a piece of road used by the car in the near future before the car has passed that piece of road. The JAAD contains information regarding crossing and not-crossing, which are the states in the intention part. Moreover, this dataset contains our required measurements such as annotation for the location, glancing, looking duration, a gesture of a pedestrian, and environmental context[Rasouli, Kotseruba, et al. 2017b]. JAAD dataset was recorded at 30 fps. If the average distance between pedestrians and their recording car is between 6 to 20 meters, they provide behavioral data. Otherwise, they report the bounding box coordinates, Id, and occlusion information [Rasouli, Kotseruba, et al. 2017b]. The tracker detects pedestrians. Therefore, only the bounding box coordinates in the JAAD dataset are used to validate this work.

## 5.7 Evaluation

To have a clear and understandable evaluation, this work utilizes accuracy, precision, recall, and F1score [Goldhammer et al. 2019]. True positive ($TP$), false positive ($FP$), true negative ($TN$), and false negative ($FN$) are used to compute these metrics. These metrics are illustrated as follows:

- Accuracy: the proportion of correctly classified pedestrians among all the pedestrians.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{5.4}$$

- Precision: the proportion of correctly classified pedestrians among classified positive pedestrians.

$$Precision = \frac{TP}{TP + FP} \tag{5.5}$$

- Recall: The proportion of correctly classified pedestrians among actual positive pedestrians.

$$Recall = \frac{TP}{TP + FN} \tag{5.6}$$

- F1score : a weighted average of precision and recall.

$$F1score = \frac{2 * Recall * Precision}{Recall + Precision} \tag{5.7}$$

It should be noted that in the action recognition part, $TP$,$FP$,$TN$,and $FN$ are computed based on their average for each action.

### 5.7.1 Action Experimental Evaluation

The JAAD dataset does not directly provide ground truth actions for pedestrians. Therefore, behavioral labels and vehicle labels are used. Based on the previous definition, the last frame before standing is stopping. Using the definition for action recognition, the ground truth (GT) on the JAAD dataset are created as follows:

1. If the behavior label of a pedestrian (available in the dataset) is standing, the GT action is standing.
2. If the behavior label of a pedestrian (available in the dataset) is walking, and the pedestrian is standing in the next frame. Then, GT action for the frames before standing are stopping.
3. If the behavior label of a pedestrian is standing, and the pedestrian is walking in the next frame. Then, GT action for the frames before walking are starting.
4. If the above items are not applicable, the GT action is walking.

Fig.5.7 illustrates probabilities of actions that this work has estimated for one pedestrian in JAAD dataset. This figure shows the sequences between states. In this specific experiment in the JAAD dataset, a pedestrian is crossing a two-lane road. Therefore, after passing the first lane, he stops. Then, he starts to walk and passes the second lane. Fig.5.7 also indicates the GT actions. As Fig.5.7 shows for the part around 3 to 5.5 seconds $p(walking) = 0.9$ and $p(starting) = 0.1$, which the ground truth indicates the pedestrian is walking. Although we also have a high probability for walking, the pedestrian does not completely follow a constant velocity motion. Therefore, there is a low probability for starting to walk. This means that walking has been defined by moving at constant but non-zero speed. Around 6 seconds, the pedestrian stops then stands for a while. Based on the ground truth at this point, the pedestrian was near the first line of the road and he stops, checks the road and starts to walk again.

Fig.5.8 shows another pedestrian in one of the test cases in the JAAD dataset. This pedestrian was walking with a group. As Fig.5.8 shows, sometimes in the figure, the probability of starting is more than walking, and the probability of stopping is more than standing, such as between 0.1 and 0.3 seconds. The reason for this behavior is that this person was walking in a group. Thus it would be possible that the tracker tracks body key-points wrongly. Moreover, the behavior of a pedestrian in a group is different, which might affect the action sequences.

As can be seen in these two figures, the features that this work chooses can recognize actions correctly. The word "correctly" means that the action with the highest probability is the same as the ground truth. As mentioned before, to have GT for starting and stopping, only one frame ahead is used. Based on these two figures and also other results, we believe that one frame is not enough for two phases of starting and stopping. Moreover, when a pedestrian continuously changes his/her walking frequency, the proposed framework face different challenges. For instance, in each walking step, a pedestrian can decide to stop. Training with more a dataset that also covers starting and stopping can be a solution and it is recommended for future work.

Table 5.1 compares our action recognition results to an algorithm that used the same dataset. The algorithm uses the same dataset to train and test their action.

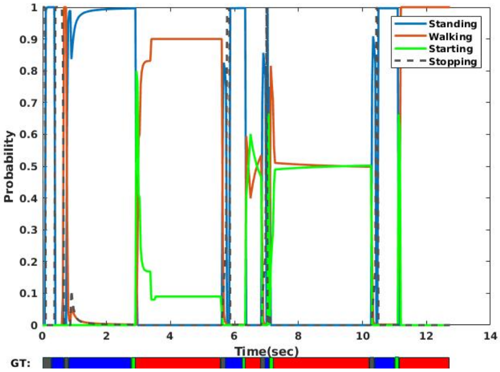


Figure 5.7: Results of our action recognizer for one of the videos in the JAAD dataset. GT means ground truth actions.

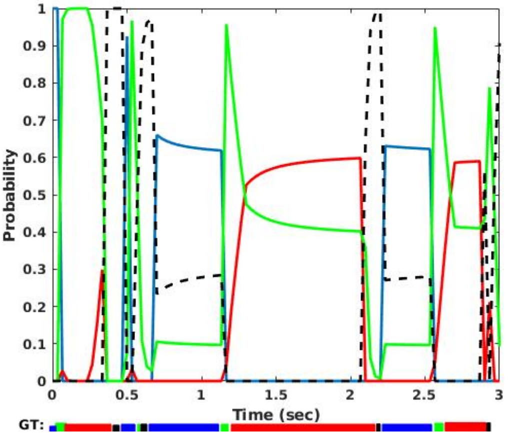


Figure 5.8: Results of our action recognizer for one of the videos in the JAAD dataset. GT means ground truth actions.

Table 5.1: Action recognizing accuracy for the JAAD dataset

| Method | Accuracy |
|---|---|
| Ours (all the dataset) | 78% |
| Ours (the last 30% of the dataset) | 83% |
| [Gujjar and Vaughan 2019] | 75 % |

Table 5.1 shows the result in two situations. In the first line, all the videos of the dataset are used to test the proposed model. As mentioned before, other datasets are used to train the action recognizer. In [Gujjar and Vaughan 2019] their actions are walking and standing. In [Gujjar and Vaughan 2019], they use the bounding boxes position that they detect to categorize action as walking or standing. They feed the position data into a binary classification network. Therefore to have a fair comparison in the second line of the table, the testing and training plan are changed as follows:

- Similar to [Gujjar and Vaughan 2019], this work used the last 30% of the JAAD dataset for testing the recognizer. The first 70% of the JAAD dataset are used to train. It means that in this part did not use other datasets to train the model.
- If actions were walking, starting, and stopping, they are considered as walking. Otherwise, the action is classified as standing. Therefore, this work has the same classification as [Gujjar and Vaughan 2019].
- In the this comparison similar to [Gujjar and Vaughan 2019], the action of pedestrians with crossing intention is recognized. It means that, in contrast with the first row, the second row shows action of a group of detected pedestrians.

Although this work has higher accuracy compared to [Gujjar and Vaughan 2019], the algorithm also has limitations. Having no measurement regarding the gait data of pedestrians can be a reason to have inaccurate action recognition such as Fig.5.3. A shopping trolley occludes legs of a pedestrian. In this situation the action recognizer has no information about leg frequency and is expected to fail.

### *5.7.2 Intention Experimental Evaluation*

Next step after recognizing the action is predicting intention. This part uses action and other measurements to predict intention. As mentioned before, the intention is predicted with three different measurements. This work used the first 70% of the JAAD dataset to train and the last 30% to test the model. The measurements are as follows:

- I : The recognized actions are used.
- II : The context and gaze information, weather condition, hand gesture, occlusion status, and the pedestrian's distance to the car are used.
- III : This measurement provided the recognized actions, the context and gaze information, weather condition, hand gesture, occlusion status, and the pedestrian's distance to an approaching car.

Table 5.2 compares this work with state-of-the-art intention prediction models on the same dataset. Similar to this work, [Fang and López 2018] and [Wang and Papanikolopoulos 2020] take advantage of body key points in sequences of images. Using libraries to extract key points in some situations, such as crowded areas, may contain errors. The errors in key points can cause inaccuracy for intention prediction. Because of this in [Fang and López 2018], they predict intention only if there is no occlusion. In Table 5.2, the average precision of testing result of [Yang, Zhan, et al. 2021] is higher compared to this work. [Yang, Zhan, et al. 2021], they use the first 70% of the JAAD dataset for training, the middle 20% for validation, and the remaining 10% for testing. As mentioned, this work uses the 30% of the JAAD dataset for testing. It should be noted that this dataset contains many situations, such as Fig.5.9. Situations that a pedestrian crosses in front of our car without eye contact or following traffic rules.

Fig.5.9 illustrates that using the context information is not enough. We have to utilize other measurements parallel to the context information. Comparing the results measurement III and [Yang, Zhang, et al. 2021] indicates the benefits of using action instead of a variation of speed and position. [Yang, Zhang, et al. 2021] uses the same context information but without considering ac-

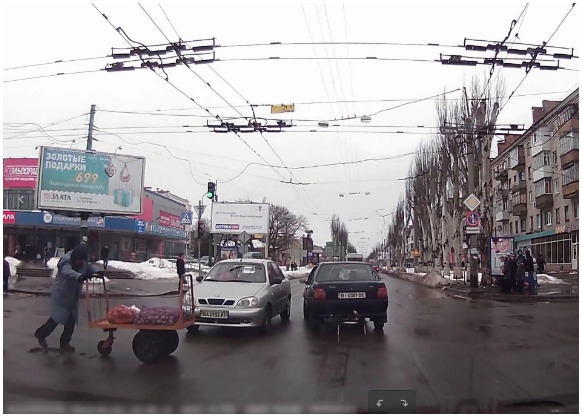


Figure 5.9: A sequence of the JAAD dataset that a pedestrian crosses the main road at a random location and breaks the traffic rules [Rasouli, Kotseruba, et al. 2017b].

tion, weather condition, and hand gesture. [Liu, Adeli, et al. 2020] concludes that the movement of vehicles in a sense can affect the pedestrian's crossing behavior. Therefore, they predict intention based on the movement of vehicles and traffic light conditions. In contrast with [Liu, Adeli, et al. 2020], Table 5.2 indicates that the combination of action and the context information can affect the intention the most, rather than the movement of other vehicles. To predict intention, [Kotseruba et al. 2020] uses the current action and trajectory of pedestrians. Therefore, it would be fair to compare it with measurement I. This comparison also shows that the action recognition based on the leg frequency can outperform the intention prediction.

As demonstrated in Table 5.2, using action and context information can improve intention prediction. The aim of Table 5.2 is to show that using all the features, we can predict intention better than using one measurement. Therefore, Table 5.2 contains different columns. The second column is $T$ frames, which means that after tracking a pedestrian, we predict intention $T$ frames before a decision point. The decision point is when a pedestrian decides to cross or not cross in front of our car. If in the dataset $T = 0$, we have to predict intention as not crossing in front of our car as far as a pedestrian is in the field of view of our camera.

The JAAD dataset contains a decision point for each pedestrian, which means this pedestrian will cross in front of our car at frame $T$. We predict final intention in T=18, T=15, and T=1 frames before decision points in JAAD. It means that we predict a chain of states using our HMM. In this dataset, $T = 18$ is 0.6 seconds, $T = 15$ that it 0.5 seconds, and $T = 1$ which is 0.033 seconds before the ground truth decision point.

Based on the table for short prediction (1 frame before decision point), measurement (I) has the lowest accuracy. Reasons for this low accuracy are as follows:

- The frequency needs some frames to converge as soon as a person starts to walk.
- One frame is not enough to decrease the effects of wrong frequencies if (a) a pedestrian stands still but moves his/her legs a bit or (b) the car moves.
- Different parameters such as weather conditions, status of an approaching car, traffic signs, and laws can cause a sudden change of direction and can change pedestrians' intention.

To show the effect of weather conditions on the action of pedestrians, Fig.5.10 shows the average frequency of walking during rainy days and clear days in the JAAD dataset. Based on this figure, pedestrians can have a different stride frequency during various weather conditions and they may walk faster on rainy days. Therefore, they may change their decision to cross the road due to weather conditions. This work added weather condition to measurement I. Then, it compared the result with measurement I. The results show the precision can improved up to 2% by adding weather conditions. It means that various parameters can influence pedestrians' behaviour. Similar to weather conditions, other parameters can affect the final intention of pedestrians. Hence, adding more information can help to increase the performance of the intention predictors.

Table 5.2: Intention recognizing accuracy for the JAAD dataset. (N.A means not available)

| Method | T | Accuracy | Precision | Recall | F1score |
|---|---|---|---|---|---|
| I | 1<br>15<br>18 | 0.88<br>0.81<br>0.83 | 0.78<br>0.74<br>0.67 | 0.87<br>0.82<br>0.83 | 0.82<br>0.77<br>0.74 |
| II | 1<br>15<br>18 | 0.91<br>0.67<br>0.65 | 0.82<br>0.79<br>0.79 | 0.92<br>0.62<br>0.56 | 0.86<br>0.70<br>0.65 |
| III | 1<br>15<br>18 | 0.92<br>0.90<br>0.89 | 0.86<br>0.78<br>0.7 | 0.93<br>0.92<br>0.87 | 0.89<br>0.84<br>0.77 |
| [Fang and López 2018] | 1<br>14 | 0.81<br>0.88 | N.A<br>N.A | N.A<br>N.A | N.A<br>N.A |
| [Wang and Papanikolopoulos 2020] | 1 | 0.81 | N.A | N.A | N.A |
| [Yang, Zhan, et al. 2021] | 1 | N.A | 0.89 | N.A | N.A |
| [Yang, Zhang, et al. 2021] | 1 | 0.83 | 0.51 | 0.81 | 0.63 |
| [Liu, Adeli, et al. 2020] | 1 | 0.79 | N.A | N.A | N.A |
| [Kotseruba et al. 2020] | 1 | 0.83 | 0.79 | 0.85 | 0.81 |

Table 5.2 shows that the context information is helpful for one frame prediction. Moreover, the results show that:

- As a result of measurement bias, weather conditions, measurement noise, and having no accurate velocity of the car the accuracy of the state estimator was not accurate enough to compute the distance. Therefore, predicting 15 frames and 18 frames before the decision point causes the

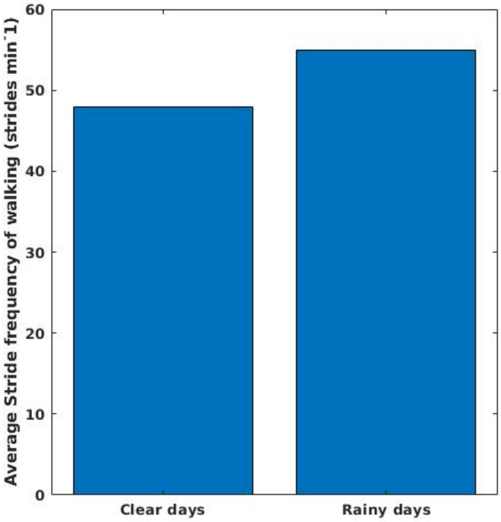


Figure 5.10: Average stride frequency of walking pedestrians in JAAD dataset during rainy days and clear days.

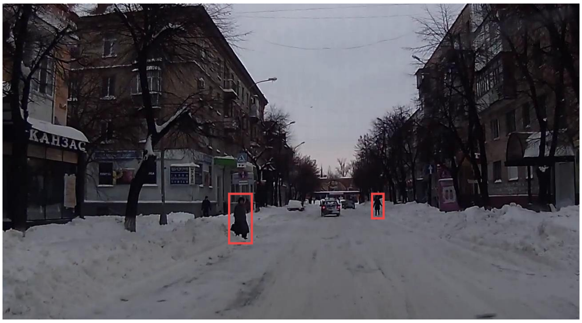


Figure 5.11: There is no measurements regarding the context information [Rasouli, Kotseruba, et al. 2017b].

worse intention prediction than the measurement I.

- In a situation such as Fig.5.11 that there is no traffic sign or light, and the pedestrian does not look at the car, the proposed model does not have enough measurements. Therefore, the prediction contains errors.

Measurements I and II indicate that the intention prediction can not perform well when using each measurement separately. By fusing all the available data, the intention predictor can become much more capable of performing at its best. Therefore, in measurement III, we have different sources of information. Results of measurement III illustrate that if we use the context information, the distance, the weather condition, and the action, we can improve our intention predictor's accuracy. The results of the third measurement validate the hypothesis in the theory of mind that adding information on the current action improves the results compared to only using the data in measurement set II. Moreover, Table 5.2 indicates that inaccurate state estimations can cause incorrect intention prediction. Besides, the environmental factor can affect the intentions. Therefore, using the temporal information and features independent of a state estimator's accuracy, we can increase an intention prediction accuracy.

## 5.8 Conclusion

The paper proposes a framework to recognize the action and predict intention. This framework is vision-based, and it includes a detector, tracker, action recognizer, and intention predictor. In this work, we only focus on the action recognition and intention prediction parts. To recognize action, we use stride frequency and position variation for each pedestrian. We predict intention with three different sets of measurements. We show that we can improve intention prediction and action recognition results using measurements that indirectly depend on state estimators' or detectors' accuracy.

Our work shows that combining action, distance, weather, context, and interaction with elements in the scene can improve prediction results. We evaluate our framework in natural driving conditions (JAAD dataset). Our framework has achieved remarkable results compared to the literature's approaches that used the same testing data.

In future work, we will extend the proposed framework to support a group intention prediction. It means predicting the intention of each pedestrian in a group. Moreover, we have to improve the interaction between the car's speed and pedestrians' crossing intentions. Besides, HMM may not be suitable if the prediction is employed in complex traffic scenes. Therefore, in future work, a different classifier will be used.

# Chapter 6

# Conclusions and Future Work

## 6.1 Conclusions and Recommendation

One of the essential steps in increasing the safety of driving is developing an awareness system. With an awareness system, it is possible to identify road users, analyze their behavior, communicate with them, predict their future actions, and choose an appropriate vehicle response. Based on the types of road users, an awareness system acts differently. In this thesis, there is a focus on pedestrian awareness systems. In the research shown in Chapters 2 to 5, it is suggested that the proposed methods and trackers can improve the accuracy and precision of a pedestrian awareness system in urban areas. In this research, the position and velocity of pedestrians related to an approaching vehicle can be estimated using pedestrian trackers.

The research was validated using various experimental and simulation tests. Moreover, to make a fair comparison with the state-of-the-art, the available datasets were utilized. This chapter summarises and discusses the lessons learned. Furthermore, this chapter presents recommendations for future work.

### *Conclusions 1:*

In Chapter 2, a multiple joints pedestrian tracker based on human kinematic constraints and a physical model was proposed.

This tracker estimates and predicts the position and velocity of legs' joints of each pedestrian.

An important conclusion is the importance of a proper state estimation model and data association algorithm during semi-occlusion situation or poor detection conditions. In these situations, non-linear models based on human kinematics can improve the prediction and estimation of states compared with linear models.

*Conclusions 2:*

A pedestrian tracker faces ambiguous situations in which different associations about the appearance of pedestrians are reasonable. Therefore, multiple hypotheses are considered, and each hypothesis is associated with a probability of being correct. In Chapter 3, it is described that the results of the entire track may be affected by the initialization of data association. Based on Chapter 3, a tracker can re-initialize its data association when it faces partial and complete occlusions, or when the number of pedestrians of the current measurement differs from previous ones.

A lesson that is learned from this chapter is that the success rate of a data association can be improved by initializing multiple hypotheses. It means that data can be matched by a hypothesis-based tracker without skipping measurements, and the probability of a tracker misses the pedestrians can be reduced using such initialization of multiple hypotheses. One of the limitations of this contribution is that sets of values were achieved based on a series of simulation tests. Therefore, based on environmental conditions it is unclear how the set can be updated. This limitation is addressed in the third contribution of this thesis. In Chapter 4, an external source of information is used to estimate and update the probabilities.

*Conclusions 3:*

In Chapter 4, a pedestrian tracker is proposed based on the measurements of vision systems and Internet-of-things (IoT) devices. The contributions in Chapters 2 and 3 are used in this tracker. An external source of information and the displacement and ID of IoT devices were added to the tracker proposed in Chapter 2. In this tracker, the advantage of IoT was considered. The importance of using prior knowledge during the data-association process is shown in Chapter 4. Furthermore, this chapter shows that detection robustness of an awareness system can be improved by fusing IoT data and on-board vision sensors. A valuable lesson learned in Chapter 4 is that pedestrians in blind spots of on-board sensors can be tracked by using IoT technology. Based on this chapter, IoT data could maintain, update, and generate a data association algorithm.

*Conclusions 4:*

In Chapter 5, it was shown that the action of pedestrians and their intentions could be recognized and predicted, respectively, by an awareness system. In this chapter, one of the important lessons learned is that the current action has an impact on predicting the intention of crossing the road. Besides, in this chapter, it is illustrated that the actions of pedestrians can be recognized without using their accurate positions related to an approaching vehicle. It means that actions can be recognized by the pedestrian awareness system using discrete features. Moreover, this chapter shows that spatio-temporal context information such as traffic signs, environmental factors, zebra-crossings can help to improve pedestrian intention prediction.

Another important lesson learned is related to the way intentions and actions are represented. In Chapter 5, action and intention were represented in a probabilistic approach. It means that the probabilities of other actions and intentions are considered by the pedestrian awareness system. Then each time, the ones with the highest probabilities are selected by it. Considering probabilities for actions and intentions can increase the accuracy of pedestrian awareness systems.

## 6.2 Future work

There are various ways to build further on the results of this research project to move toward a pedestrian awareness system. In this section, directions for future research are suggested.

- **Improve pedestrian motion model**

  To track pedestrian, this work used pedestrians' legs. Suppose other objects, such as other pedestrians, cars, or cyclists, occlude a pedestrians' legs. Although a camera detected the upper body of the pedestrians, the tracker is unable to track them based on the leg joints. Therefore, adding key-points of the upper body is recommended to improve the accuracy and precision of the tracker.

- **Analyse group behavior**

  In practice, pedestrians often walk in groups. Therefore, it is necessary to understand local interactions to develop a reliable pedestrian awareness system for urban infrastructures, traffic management, and pedestrian safety during mass events. However, the characteristics of the motion of pedestrian groups have not yet been empirically studied. It is unknown how moving group members interact with other pedestrians and with other groups. How such groups organize in space and how these spatial patterns affect pedestrian flow dynamics also need further study.

- **Improve noise covariance matrices**

  The accuracy of the pedestrian awareness system is highly dependent on the accuracy of its process and measurement noise covariance matrices. Pedestrians may be lost, and their identities are mixed and overall accuracy and precision of awareness systems may be reduced in the presence of improper noise covariance matrices. In this thesis, an ad-hoc procedure is used to tune these two matrices, in which the matrices are assumed to be constant and fix during the estimation and are manually adjusted by trial and error approaches. Based on applications and various conditions these two can be varied. In future work, real-time adaptations to Kalman filter is suggested.

- **Multiple pedestrians and multiple IoT devices**

  This work assumes that each pedestrian carries one IoT device, which, in reality, is not always a valid assumption. Suppose that a pedestrian takes more than one device. In that case, data association between IoT devices and onboard sensors measurements is more challenging. Therefore, it is suggested to improve the data association algorithm to remove this assumption.

- **Validate the proposed approach in nonideal situations**

  In this thesis, different datasets were used to validate the proposed models. Besides, various experimental and simulation tests are used to validate the proposed models. The experimental trials were held during weekends at the Tu/e campus. During experiments, pedestrians were asked to walk a given route, where the route was carefully defined. It is highly recommended to collect more experimental results on a more realistic scenario by testing on a real-world situation with more environmental noises and various speeds of a vehicle.

# Acknowledgments

Earning a Ph.D. degree is not solely about growing knowledge; it is a masterclass in personal development. My Ph.D. was an inflection point in my life. My journey would not have been successful without several wonderful people in my life.

I would like to start by thanking my first promoter, supervisor, mentor, and teacher, René van Molengraft. René, words fail to express how thankful I am for you. You have always believed in and supported me, no matter the circumstances. You have inspired me to increase my confidence, knowledge, and strength. Through our weekly meetings, you have noticed all my concerns and challenges as an international Ph.D. student. Your insights have been valuable. Thank you for your never-ending support. You taught me how to create a balance in my life. I will forever remember our countless spirited discussions about research and culture, as well as your positive outlook on both my professional and personal life.

Next, I would like to thank my second promoter, Maarten Steinbuch. I was thrilled when you provided me with this opportunity and became my promoter. You have succeeded in making the group seem like a family, and I thank you for that.

Completing this thesis would not have been possible without the support of my co-promoter, Jos Elfring. Jos, I owe my deepest gratitude to you. You made my day when you sent me that email offering me a collaboration. Thank you for your continuous guidance and interest in my work. I highly value your sharp mind and honesty, and I am forever grateful for the chance to work with you.

I wish to thank other committee members separately. I truly appreciate the time and effort you dedicated to reviewing my thesis.

My journey wouldn't have been such a memorable experience without the contributions of my CST colleagues and friends. Many thanks to Nancy and Roos for their help and support during the graduation phase. I would like to thank Jos den Ouden for his managing of the experiments and deadlines of the project.

I have had the pleasure of working among the most wonderful colleagues. Puck, Wouter Kuijpers, Wouter Houtman, Bob, Roy, Yanick, Hao, Cezar, Jesse, and Jordy: Thank you for your constant encouragement, kindness, and support. Because of you guys, our office was like home. I knew nothing about the topic when I first came to the office, but you supported me. Thank you for giving me a chance to learn from you. By the way, I will miss all of our coffee breaks.

My time in the Netherlands would have been formidable because of my friends. Zohreh, we have known each other for more than half of our lives. Your friendship, individuality, and uplifting conversations have been a source of comfort to me. Thank you for our friendship. Aida, thank you for being around me whenever I needed you. From the first time we met each other, we knew we would develop a great friendship filled with tears and laughter. Sina, we met each other in a course based on a mistake I made. When I look back at our friendship, I am so happy that I bought the wrong ticket. Thanks for all the coffee breaks, hiking trips, and gym sessions. Ariyan, you came to the Netherlands and made my life easier. Thank you for all your help and support. You know me very well, including the fact that I often forget to listen to you, so you always explain scientific topics in a way that intrigues me. Bahareh, first of all, thanks for accepting the PDEng project. That project changed the course of my Ph.D. Whenever I want to shop, hike, or have brunch, I always know that I can count on you. Thank you for always being there and for your excellent photography skills. Sahar and Sina, thanks to COVID-19, we found each other. I wish you all the love. Thank you for all the fun times we had. The nights that, whenever I think about them, I cannot control my laughter. My dear Soheil, thank you for all the conversations we had, your kindness, and for being part of my story. Arghavan, Mirhossein, and Ahmadreza, thank you for every hour we spent together. Mohammad, you have been a lovely friend who always cheered me up. Thank you for your motivational words. Antje, the joy you bring with you always makes my day. Thank you for your positive energy. Behnam, thank you for all the laughs and stories you have shared. Chyannie, thanks for your understanding, kindness, and friendship. I am proud of you and impressed by your achievements.

To my second family in the Netherlands, Shima, Ali, and Arash. Shima, my lovely sister, after every storm I experienced in my life, you came like a hope-giving rainbow. Thank you for your unconditional support, help, and love. Ali, thanks for being such a great and strong brother. I will never forget the many trips we took together. My dearest Arash, you are the most influential person in my life. You have always believed in me and encouraged me to keep going. You have supported me ever since I decided to leave Iran, and I thank you for that.

Last but not least, my warmest gratitude and tokens of love go to my dear family. Maman, Baba, Malihe, Alireza, Omid, Fatemeh and my love Arya, no words can adequately express how thankful I am for your unfailing support and love. I have always surprised you with my decisions, but you have always trusted me. Your love and trust made me a happy and independent girl. Thank you for always being by my side. I love you all so much.

Marzieh Dolatabadi
Eindhoven, November 2021

# Curriculum Vitae

Marzieh Dolatabadi was born on October 09, 1988, in Tehran, Iran. She received her Bachelor of Science degree in aerospace engineering from Iran University of Science and Research (IAU), Tehran, Iran, in 2012. She received her Master of Science degree in aerospace engineering with a specialization in aerodynamic engineering from University of Tehran (UT), Tehran, Iran, in 2014.

In April 2017, she started to pursue a Ph.D. degree within the Control Systems Technology group at the Department of Mechanical Engineering at Eindhoven University of Technology (TU/e), under the supervision of Ren é van de Molengraft. Her research project was part of a research program "Autopilot" funded from the European Union's Horizon 2020 Framework Program. The main results of her research are printed in this dissertation.

# Bibliography

Abbasi-Kesbi, R. and A. Nikfarjam (2018). “A miniature sensor system for precise hand position monitoring”. *IEEE Sensors Journal* vol. 18, no. 6, pp. 2577–2584.

Abuelsamid, S., D. Alexander, and L. Jerram (2017). “Navigant research leaderboard report: automated driving”. *Chicago: Navigant*.

Afolabi, O., K. Driggs–Campbell, R. Dong, M. J. Kochenderfer, and S. S. Sastry (2018). “People as sensors: Imputing maps from human actions”. *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 2342–2348.

Althoff, M. and S. Magdici (2016). “Set-based prediction of traffic participants on arbitrary road networks”. *IEEE Transactions on Intelligent Vehicles* vol. 1, no. 2, pp. 187–202.

Angelini, F., Z. Fu, S. A. Velastin, J. A. Chambers, and S. M. Naqvi (2018). “3d-hog embedding frameworks for single and multi-viewpoints action recognition based on human silhouettes”. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 4219–4223.

Anuj, L. and M. G. Krishna (2017). “Multiple camera based multiple object tracking under occlusion: A survey”. *2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*. IEEE, pp. 432–437.

Areta, J., Y. Bar-Shalom, M. Levedahl, and K. R. Pattipati (2006). “Hierarchical Track Association and Fusion for a Networked Surveillance System.” *J. Adv. Inf. Fusion* vol. 1, no. 2, pp. 140–157.

Asvadi, A., L. Garrote, C. Premebida, P. Peixoto, and U. J. Nunes (2018). "Multimodal vehicle detection: fusing 3D-LIDAR and color camera data". *Pattern Recognition Letters* vol. 115, pp. 20–29.

Azim, A. and O. Aycard (2010). "Multiple pedestrian tracking using Viterbi data association". *2010 IEEE Intelligent Vehicles Symposium*. IEEE, pp. 706–711.

Baghdadi, A., L. A. Cavuoto, and J. L. Crassidis (2018). "Hip and trunk kinematics estimation in gait through Kalman filter using IMU data at the ankle". *IEEE Sensors Journal* vol. 18, no. 10, pp. 4253–4260.

Bai, H., S. Cai, N. Ye, D. Hsu, and W. S. Lee (2015). "Intention-aware online POMDP planning for autonomous driving in a crowd". *2015 ieee international conference on robotics and automation (icra)*. IEEE, pp. 454–460.

Bajracharya, M., B. Moghaddam, A. Howard, S. Brennan, and L. H. Matthies (2009). "A fast stereo-based system for detecting and tracking pedestrians from a moving vehicle". *The International Journal of Robotics Research* vol. 28, no. 11-12, pp. 1466–1485.

Banerjee, K., D. Notz, J. Windelen, S. Gavarraju, and M. He (2018). "Online camera lidar fusion and object detection on hybrid data for autonomous driving". *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, pp. 1632–1638.

Bao, S.-D., X.-L. Meng, W. Xiao, and Z.-Q. Zhang (2017). "Fusion of inertial/magnetic sensor measurements and map information for pedestrian tracking". *Sensors* vol. 17, no. 2, p. 340.

Bar-Shalom, Y., F. Daum, and J. Huang (2009). "The probabilistic data association filter". *IEEE Control Systems Magazine* vol. 29, no. 6, pp. 82–100.

Bar-Shalom, Y., P. K. Willett, and X. Tian (2011). *Tracking and data fusion*. Vol. 11. YBS publishing Storrs, CT, USA:

Bassett, D. R., L. P. Toth, S. R. LaMunion, and S. E. Crouter (2017). "Step counting: a review of measurement considerations and health-related applications". *Sports Medicine* vol. 47, no. 7, pp. 1303–1315.

Bavdekar, V. A., A. P. Deshpande, and S. C. Patwardhan (2011). "Identification of process and measurement noise covariance for state and parameter estimation using extended Kalman filter". *Journal of Process control* vol. 21, no. 4, pp. 585–601.

Baxter, R. H., M. J. Leach, S. S. Mukherjee, and N. M. Robertson (2014). "An adaptive motion model for person tracking with instantaneous head-pose features". *IEEE Signal Processing Letters* vol. 22, no. 5, pp. 578–582.

Bennett, T., R. Jafari, and N. Gans (2013). "An extended kalman filter to estimate human gait parameters and walking distance". *2013 American Control Conference*. IEEE, pp. 752–757.

Bernardin, K., A. Elbs, and R. Stiefelhagen (2006). "Multiple object tracking performance metrics and evaluation in a smart room environment". *Sixth IEEE International Workshop on Visual Surveillance, in conjunction with ECCV*. Vol. 90. 91. Citeseer.

Bertram, J. E. and A. Ruina (2001). "Multiple walking speed–frequency relations are predicted by constrained optimization". *Journal of theoretical Biology* vol. 209, no. 4, pp. 445–453.

Bhuvaneswari, S. and T. Subashini (2014). "TRACKING MANUALLY SELECTED OBJECT IN VIDEOS USING COLOR HISTOGRAM MATCHING." *Journal of Theoretical & Applied Information Technology* vol. 67, no. 3.

Bilal, M. (2017). "Algorithmic optimisation of histogram intersection kernel support vector machine-based pedestrian detection using low complexity features". *IET Computer Vision* vol. 11, no. 5, pp. 350–357.

Blackman, S. S. (2004). "Multiple hypothesis tracking for multiple target tracking". *IEEE Aerospace and Electronic Systems Magazine* vol. 19, no. 1, pp. 5–18.

Blackman, S. and R. Popoli (1999). *Design and Analysis of Modern Tracking Systems. Norwood, MA, USA: Artech House*.

Blakemore, S.-J. and J. Decety (2001). "From the perception of action to the understanding of intention". *Nature reviews neuroscience* vol. 2, no. 8, pp. 561–567.

Bouhsain, S. A., S. Saadatnejad, and A. Alahi (2020). "Pedestrian intention prediction: A multi-task perspective". *arXiv preprint arXiv:2010.10270*.

Boukerche, A. and Z. Hou (2021). “Object detection using deep learning methods in traffic scenarios”. *ACM Computing Surveys (CSUR)* vol. 54, no. 2, pp. 1–35.

Bruneliere, H., E. Burger, J. Cabot, and M. Wimmer (2019). “A feature-based survey of model view approaches”. *Software & Systems Modeling* vol. 18, no. 3, pp. 1931–1952.

Cadena, P. R. G., M. Yang, Y. Qian, and C. Wang (2019). “Pedestrian graph: Pedestrian crossing prediction based on 2d pose estimation and graph convolutional networks”. *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, pp. 2000–2005.

Cao, Z., T. Simon, S.-E. Wei, and Y. Sheikh (2017a). “Realtime multi-person 2d pose estimation using part affinity fields”. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7291–7299.

Cao, Z., T. Simon, S.-E. Wei, and Y. Sheikh (2017b). “Realtime multi-person 2d pose estimation using part affinity fields”. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7291–7299.

Chaabane, M., A. Trabelsi, N. Blanchard, and R. Beveridge (2020). “Looking ahead: Anticipating pedestrians crossing with future frames prediction”. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2297–2306.

Chau, D. P., F. Bremond, and M. Thonnat (2013). “Object tracking in videos: Approaches and issues”. *arXiv preprint arXiv:1304.5212*.

Chavez-Garcia, R. O. and O. Aycard (2015). “Multiple sensor fusion and classification for moving object detection and tracking”. *IEEE Transactions on Intelligent Transportation Systems* vol. 17, no. 2, pp. 525–534.

Chen, C. W. (2021). “Drones as internet of video things front-end sensors: challenges and opportunities”. *Discover Internet of Things* vol. 1, no. 1, pp. 1–12.

Chen, C., P. Zhao, C. X. Lu, W. Wang, A. Markham, and N. Trigoni (2020). “Deep-learning-based pedestrian inertial navigation: Methods, data set, and on-device inference”. *IEEE Internet of Things Journal* vol. 7, no. 5, pp. 4431–4441.

Chen, L., N. Ma, P. Wang, J. Li, P. Wang, G. Pang, and X. Shi (2020). “Survey of pedestrian action recognition techniques for autonomous driving". *Tsinghua Science and Technology* vol. 25, no. 4, pp. 458–470.

Chereshnev, R. and A. Kertész-Farkas (2017). "Hugadb: Human gait database for activity recognition from wearable inertial sensor networks". *International Conference on Analysis of Images, Social Networks and Texts*. Springer, pp. 131–141.

Choi, W. (2015). "Near-online multi-target tracking with aggregated local flow descriptor". *Proceedings of the IEEE international conference on computer vision*, pp. 3029–3037.

Chou, F.-C., T.-H. Lin, H. Cui, V. Radosavljevic, T. Nguyen, T.-K. Huang, M. Niedoba, J. Schneider, and N. Djuric (2020). "Predicting motion of vulnerable road users using high-definition maps and efficient convnets". *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, pp. 1655–1662.

Combs, T. S., L. S. Sandt, M. P. Clamann, and N. C. McDonald (2019). "Automated vehicles and pedestrian safety: exploring the promise and limits of pedestrian detection". *American journal of preventive medicine* vol. 56, no. 1, pp. 1–7.

Commission, E. (2017). *A Road Transport Strategy for Europe.* Tech. rep.

Corbetta, A., J. A. Meeusen, C.-m. Lee, R. Benzi, and F. Toschi (2018). "Physics-based modeling and data representation of pairwise interactions among pedestrians". *Physical Review E* vol. 98, no. 6, p. 062310.

Cox, I. J. and J. J. Leonard (1994). "Modeling a dynamic environment using a Bayesian multiple hypothesis approach". *Artificial Intelligence* vol. 66, no. 2, pp. 311–344.

Danion, F., E. Varraine, M. Bonnard, and J. Pailhous (2003). "Stride variability in human gait: the effect of stride frequency and stride length". *Gait & posture* vol. 18, no. 1, pp. 69–77.

Deb, S., M. Yeddanapudi, K. Pattipati, and Y. Bar-Shalom (1997). "A generalized SD assignment algorithm for multisensor-multitarget state estimation". *IEEE Transactions on Aerospace and Electronic systems* vol. 33, no. 2, pp. 523–538.

Dimitrievski, M., P. Veelaert, and W. Philips (2019). "Behavioral pedestrian tracking using a camera and lidar sensors on a moving vehicle". *Sensors* vol. 19, no. 2, p. 391.

Ding, R., M. Yu, H. Oh, and W.-H. Chen (2016). “New multiple-target tracking strategy using domain knowledge and optimization”. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* vol. 47, no. 4, pp. 605–616.

Dollar, P., C. Wojek, B. Schiele, and P. Perona (2011). “Pedestrian detection: An evaluation of the state of the art”. *IEEE transactions on pattern analysis and machine intelligence* vol. 34, no. 4, pp. 743–761.

Elfring, J., S. van den Dries, M. Van De Molengraft, and M. Steinbuch (2013). “Semantic world modeling using probabilistic multiple hypothesis anchoring”. *Robotics and Autonomous Systems* vol. 61, no. 2, pp. 95–105.

Ess, A., K. Schindler, B. Leibe, and L. Van Gool (2010). “Object detection and tracking for autonomous navigation in dynamic environments”. *The International Journal of Robotics Research* vol. 29, no. 14, pp. 1707–1725.

Fang, H.-S., S. Xie, Y.-W. Tai, and C. Lu (2017). “Rmpe: Regional multi-person pose estimation”. *Proceedings of the IEEE international conference on computer vision*, pp. 2334–2343.

Fang, Z. and A. M. López (2018). “Is the pedestrian going to cross? answering by 2d pose estimation”. *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, pp. 1271–1276.

Fang, Z. and A. M. López (2019). “Intention recognition of pedestrians and cyclists by 2d pose estimation”. *IEEE Transactions on Intelligent Transportation Systems* vol. 21, no. 11, pp. 4773–4783.

Fang, Z., D. Vázquez, and A. M. López (2017). “On-board detection of pedestrian intentions”. *Sensors* vol. 17, no. 10, p. 2193.

Feichtenhofer, C., A. Pinz, and A. Zisserman (2017). “Detect to track and track to detect”. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3038–3046.

Feng, W., Z. Hu, W. Wu, J. Yan, and W. Ouyang (2019). “Multi-object tracking with multiple cues and switcher-aware classification”. *arXiv preprint arXiv:1901.06129*.

Ferguson, S., B. Luders, R. C. Grande, and J. P. How (2015). “Real-time predictive modeling and robust avoidance of pedestrians with uncertain, changing intentions”. *Algorithmic Foundations of Robotics XI*. Springer, pp. 161–177.

Fiore, U., A. Florea, and G. Pérez Lechuga (2019). "An Interdisciplinary Review of Smart Vehicular Traffic and Its Applications and Challenges". *Journal of Sensor and Actuator Networks* vol. 8, no. 1, p. 13.

Fod, A., A. Howard, and M. Mataric (2002). "A laser-based people tracker". *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292)*. Vol. 3. IEEE, pp. 3024–3029.

Fortmann, T., Y. Bar-Shalom, and M. Scheffe (1983). "Sonar tracking of multiple targets using joint probabilistic data association". *IEEE journal of Oceanic Engineering* vol. 8, no. 3, pp. 173–184.

Fruhwirth-Reisinger, C., G. Krispel, H. Possegger, and H. Bischof (2020). "Towards data-driven multi-target tracking for autonomous driving". *Proceedings of the 25th Computer Vision Winter Workshop (CVWW), Rogaska Slatina, Slovenia*, pp. 3–5.

Führ, G. and C. R. Jung (2014). "Combining patch matching and detection for robust pedestrian tracking in monocular calibrated cameras". *Pattern Recognition Letters* vol. 39, pp. 11–20.

Fukuda, S., M. Nishiyama, and Y. Iwai (2021). "Reduction in Communication via Image Selection for Homomorphic Encryption-based Privacy-protected Person Re-identification." *VISIGRAPP (5: VISAPP)*, pp. 36–47.

Gaurav, A. K., N. Sahu, A. P. Dash, G. Chalapathi, V. Chamola, et al. (2021). "A survey on computation resource allocation in IoT enabled vehicular edge computing". *Complex & Intelligent Systems*, pp. 1–23.

Geiger, A., P. Lenz, C. Stiller, and R. Urtasun (2013). "Vision meets robotics: The kitti dataset". *The International Journal of Robotics Research* vol. 32, no. 11, pp. 1231–1237.

Geiger, A., P. Lenz, and R. Urtasun (2012). "Are we ready for autonomous driving? the kitti vision benchmark suite". *2012 IEEE conference on computer vision and pattern recognition*. IEEE, pp. 3354–3361.

Gelbal, S. Y., B. Aksun-Guvenc, and L. Guvenc (2020). "Collision avoidance of low speed autonomous shuttles with pedestrians". *International journal of automotive technology* vol. 21, no. 4, pp. 903–917.

Ghori, O., R. Mackowiak, M. Bautista, N. Beuter, L. Drumond, F. Diego, and B. Ommer (2018). "Learning to forecast pedestrian intention from pose dynamics". *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, pp. 1277–1284.

Gindele, T., S. Brechtel, and R. Dillmann (2015). "Learning driver behavior models from traffic observations for decision making and planning". *IEEE Intelligent Transportation Systems Magazine* vol. 7, no. 1, pp. 69–79.

Goldhammer, M., S. Köhler, S. Zernetsch, K. Doll, B. Sick, and K. Dietmayer (2019). "Intentions of Vulnerable Road Users—Detection and Forecasting by Means of Machine Learning". *IEEE transactions on intelligent transportation systems* vol. 21, no. 7, pp. 3035–3045.

Goodfellow, I. (2016). "Nips 2016 tutorial: Generative adversarial networks". *arXiv preprint arXiv:1701.00160*.

Gu, Y., Y. Hashimoto, L.-T. Hsu, and S. Kamijo (2016). "Motion planning based on learning models of pedestrian and driver behaviors". *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, pp. 808–813.

Gu, Y., M. Liu, W. Sheng, Y. Ou, and Y. Li (2021). "Sensor fusion based manipulative action recognition". *Autonomous Robots* vol. 45, no. 1, pp. 1–13.

Gujjar, P. and R. Vaughan (2019). "Classifying pedestrian actions in advance using predicted video of urban driving scenes". *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 2097–2103.

Gunawan, A. A., W. Jatmiko, and A. M. Arymurthy (2017). "Fast and Optimal Visual Tracking based on Spectral Method". *Procedia computer science* vol. 116, pp. 571–578.

Guo, Y., F. Sohel, M. Bennamoun, J. Wan, and M. Lu (2015). "A novel local surface feature for 3D object recognition under clutter and occlusion". *Information Sciences* vol. 293, pp. 196–213.

He, K., X. Zhang, S. Ren, and J. Sun (2016). "Deep residual learning for image recognition". *Proceedings of the IEEE con-*

*ference on computer vision and pattern recognition*, pp. 770–778.

He, M., H. Luo, B. Hui, and Z. Chang (2019). “Fast online multi-pedestrian tracking via integrating motion model and deep appearance model”. *IEEE Access* vol. 7, pp. 89475–89486.

Ho, N.-H., P. H. Truong, and G.-M. Jeong (2016). “Step-detection and adaptive step-length estimation for pedestrian dead-reckoning at various walking speeds using a smartphone”. *Sensors* vol. 16, no. 9, p. 1423.

Hou, M., J. Cheng, F. Xiao, and C. Wang (2021). “Distracted behavior of pedestrians while crossing street: a case study in China”. *International journal of environmental research and public health* vol. 18, no. 1, p. 353.

Hsu, Y.-L., J.-S. Wang, and C.-W. Chang (2017). “A wearable inertial pedestrian navigation system with quaternion-based extended Kalman filter for pedestrian localization”. *IEEE Sensors Journal* vol. 17, no. 10, pp. 3193–3206.

Iqbal, U., A. Milan, and J. Gall (2017). “Posetrack: Joint multi-person pose estimation and tracking”. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2011–2020.

Jalal, A., S. Kamal, and C. A. Azurdia-Meza (2019). “Depth maps-based human segmentation and action recognition using full-body plus body color cues via recognizer engine”. *Journal of Electrical Engineering & Technology* vol. 14, no. 1, pp. 455–461.

Jeung, H., H. T. Shen, and X. Zhou (2007). “Mining trajectory patterns using hidden Markov models”. *International Conference on Data Warehousing and Knowledge Discovery*. Springer, pp. 470–480.

Johnson, L. S., S. R. Eddy, and E. Portugaly (2010). “Hidden Markov model speed heuristic and iterative HMM search procedure”. *BMC bioinformatics* vol. 11, no. 1, pp. 1–8.

Khandelwal, S. and N. Wickström (2017). “Evaluation of the performance of accelerometer-based gait event detection algorithms in different real-world scenarios using the MAREA gait database”. *Gait & posture* vol. 51, pp. 84–90.

Kim, C., F. Li, A. Ciptadi, and J. M. Rehg (2015). “Multiple hypothesis tracking revisited”. *Proceedings of the IEEE international conference on computer vision*, pp. 4696–4704.

Kim, S. and M. Kim (2016). “Occluded pedestrian classification using gradient patch and convolutional neural networks”. *Advances in Computer Science and Ubiquitous Computing*. Springer, pp. 198–204.

Klinger, V. and M. Arens (2009). “Ragdolls in action–action recognition by 3d pose recovery from monocular video”. *Proceedings of the IADIS International Conference on Computer Graphics, Visualization, Computer Vision and Image Processing*, pp. 219–223.

Kong, W., A. Hussain, and M. H. M. Saad (2013). “Essential human body points tracking using kalman filter”. *Proceedings of the World Congress on Engineering and Computer Science*. Vol. 1, pp. 503–507.

Kooij, J. F., F. Flohr, E. A. Pool, and D. M. Gavrila (2019). “Context-based path prediction for targets with switching dynamics”. *International Journal of Computer Vision* vol. 127, no. 3, pp. 239–262.

Köpf, F., S. Ramsteiner, L. Puccetti, M. Flad, and S. Hohmann (2020). “Adaptive dynamic programming for model-free tracking of trajectories with time-varying parameters”. *International Journal of Adaptive Control and Signal Processing* vol. 34, no. 7, pp. 839–856.

Kotseruba, I., A. Rasouli, and J. K. Tsotsos (2020). “Do they want to cross? understanding pedestrian intention for behavior prediction”. *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, pp. 1688–1693.

Kurz, M. J. and N. Stergiou (2007). “Hip actuations can be used to control bifurcations and chaos in a passive dynamic walking model”.

Kwak, J.-Y., E.-J. Lee, B. Ko, and M. Jeong (2016). “Pedestrian’s intention prediction based on fuzzy finite automata and spatial-temporal features”. *Electronic Imaging* vol. 2016, no. 3, pp. 1–6.

Kwon, D., S. Park, S. Baek, R. K. Malaiya, G. Yoon, and J.-T. Ryu (2018). “A study on development of the blind spot detection system for the IoT-based smart connected car". *2018 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, pp. 1–4.

Kwon, S. K., E. Hyun, J.-H. Lee, J. Lee, and S. H. Son (2017). "Detection scheme for a partially occluded pedestrian based on occluded depth in lidar–radar sensor fusion". *Optical Engineering* vol. 56, no. 11, p. 113112.

Laursen, T., N. B. Pedersen, J. J. Nielsen, and T. K. Madsen (2012). "Hidden Markov Model based mobility learning fo improving indoor tracking of mobile users". *2012 9th Workshop on Positioning, Navigation and Communication*. IEEE, pp. 100–104.

Lei, J., G. Li, S. Li, D. Tu, and Q. Guo (2016). "Continuous action recognition based on hybrid CNN-LDCRF model". *2016 International Conference on Image, Vision and Computing (ICIVC)*. IEEE, pp. 63–69.

Leibe, B., K. Schindler, and L. Van Gool (2007). "Coupled detection and trajectory estimation for multi-object tracking". *2007 IEEE 11th International Conference on Computer Vision*. IEEE, pp. 1–8.

Li, S., L. Zhang, and X. Diao (2018). "Improving human intention prediction using data augmentation". *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, pp. 559–564.

Li, X. R. and Y. Bar-Shalom (1996). "Tracking in clutter with nearest neighbor filters: analysis and performance". *IEEE transactions on aerospace and electronic systems* vol. 32, no. 3, pp. 995–1010.

Liang, J., U. Patel, A. J. Sathyamoorthy, and D. Manocha (2020). "Realtime collision avoidance for mobile robots in dense crowds using implicit multi-sensor fusion and deep reinforcement learning". *arXiv preprint arXiv:2004.03089*.

Liao, X. L. and C. Zhang (2017). "Toward situation awareness: a survey on adaptive learning for model-free tracking". *Multimedia Tools and Applications* vol. 76, no. 20, pp. 21073–21115.

Linder, T. and K. O. Arras (2014). “Multi-model hypothesis tracking of groups of people in RGB-D data”. *17th International conference on information fusion (FUSION)*. IEEE, pp. 1–7.

Liu, B., E. Adeli, Z. Cao, K.-H. Lee, A. Shenoi, A. Gaidon, and J. C. Niebles (2020). “Spatiotemporal relationship reasoning for pedestrian intent prediction”. *IEEE Robotics and Automation Letters* vol. 5, no. 2, pp. 3485–3492.

Liu, H. and W. Wu (2017). “Interacting multiple model (IMM) fifth-degree spherical simplex-radial cubature Kalman filter for maneuvering target tracking”. *Sensors* vol. 17, no. 6, p. 1374.

Luo, Y., S. M. Coppola, P. C. Dixon, S. Li, J. T. Dennerlein, and B. Hu (2020). “A database of human gait performance on irregular and uneven surfaces collected by wearable sensors”. *Scientific data* vol. 7, no. 1, pp. 1–9.

Mahler, R. P. (2007). *Statistical multisource-multitarget information fusion*. Artech House, Inc.

Marginean, A., R. Brehar, and M. Negru (2019). “Understanding pedestrian behaviour with pose estimation and recurrent networks”. *2019 6th International Symposium on Electrical and Electronics Engineering (ISEEE)*. IEEE, pp. 1–6.

Marzoli, D., A. Pagliara, G. Prete, G. Malatesta, C. Lucafò, C. Padulo, A. Brancucci, and L. Tommasi (2019). “Lateralized embodiment of ambiguous human silhouettes: data on sex differences”. *Data in brief* vol. 25, p. 104009.

Masoud, O. and N. P. Papanikolopoulos (2001). “A novel method for tracking and counting pedestrians in real-time using a single camera”. *IEEE transactions on vehicular technology* vol. 50, no. 5, pp. 1267–1278.

Mathew, W., R. Raposo, and B. Martins (2012). “Predicting future locations with hidden Markov models”. *Proceedings of the 2012 ACM conference on ubiquitous computing*, pp. 911–918.

Merdrignac, P., O. Shagdar, and F. Nashashibi (2016). “Fusion of perception and v2p communication systems for the safety of vulnerable road users”. *IEEE Transactions on Intelligent Transportation Systems* vol. 18, no. 7, pp. 1740–1751.

Moon, S., Y. Park, D. W. Ko, and I. H. Suh (2016). “Multiple kinect sensor fusion for human skeleton tracking using Kalman filtering". *International Journal of Advanced Robotic Systems* vol. 13, no. 2, p. 65.

Moraffah, B. and A. Papandreou-Suppappola (2019). "Random infinite tree and dependent Poisson diffusion process for nonparametric Bayesian modeling in multiple object tracking". *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5217–5221.

Morimitsu, H., I. Bloch, and R. M. Cesar-Jr (2017). "Exploring structure for long-term tracking of multiple objects in sports videos". *Computer Vision and Image Understanding* vol. 159, pp. 89–104.

Mozaffari, S., O. Y. Al-Jarrah, M. Dianati, P. Jennings, and A. Mouzakitis (2020). "Deep learning-based vehicle behavior prediction for autonomous driving applications: A review". *IEEE Transactions on Intelligent Transportation Systems*.

Murthy, K. G. (1968). "An algorithm for ranking all the assignments in order of increasing costs". *Operations research* vol. 16, no. 3, pp. 682–687.

Muscholl, N., A. Poibrenski, M. Klusch, and P. Gebhard (2020). "Simp3: Social interaction-based multi-pedestrian path prediction by self-driving cars". *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, pp. 2731–2738.

Nguyen, U., F. Rottensteiner, and C. Heipke (2019). "Confidence-aware pedestrian tracking using a stereo camera". *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 4 (2019), Nr. 2/W5* vol. 4, no. 2/W5, pp. 53–60.

Nielsen, J. J., N. Amiot, and T. K. Madsen (2013). "Directional hidden markov model for indoor tracking of mobile users and realistic case study". *European Wireless 2013; 19th European Wireless Conference*. VDE, pp. 1–6.

Ning, C., L. Menglu, Y. Hao, S. Xueping, and L. Yunhong (2021). "Survey of pedestrian detection with occlusion". *Complex & Intelligent Systems* vol. 7, no. 1, pp. 577–587.

Nwaizu, H., R. Saatchi, and D. Burke (2016). "Accelerometer based human joints' range of movement measurement". *2016 10th International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP)*. IEEE, pp. 1–6.

Ojala, T., M. Pietikainen, and T. Maenpaa (2002). “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns”. *IEEE Transactions on pattern analysis and machine intelligence* vol. 24, no. 7, pp. 971–987.

Organization, W. H. et al. (2018). *Global status report on road safety 2018: summary*. Tech. rep. World Health Organization.

Pan, S., Q. Bao, and Z. Chen (2017). “An efficient TO-MHT algorithm for multi-target tracking in cluttered environment”. *2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*. IEEE, pp. 705–708.

Pfeiffer, M., G. Paolo, H. Sommer, J. Nieto, R. Siegwart, and C. Cadena (2018). “A data-driven model for interaction-aware pedestrian motion prediction in object cluttered environments”. *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 5921–5928.

Piccoli, F., R. Balakrishnan, M. J. Perez, M. Sachdeo, C. Nunez, M. Tang, K. Andreasson, K. Bjurek, R. D. Raj, E. Davidsson, et al. (2020). “Fussi-net: Fusion of spatio-temporal skeletons for intention prediction network”. *2020 54th Asilomar Conference on Signals, Systems, and Computers*. IEEE, pp. 68–72.

Pollard, E., B. Pannetier, and M. Rombaut (2009). “Convoy detection processing by using the hybrid algorithm (gmcphd/vs-immc-mht) and dynamic bayesian networks”. *2009 12th International Conference on Information Fusion*. IEEE, pp. 907–914.

Pollard, E., B. Pannetier, and M. Rombaut (2011). “Hybrid algorithms for multitarget tracking using MHT and GM-CPHD”. *IEEE Transactions on Aerospace and Electronic Systems* vol. 47, no. 2, pp. 832–847.

Pop, D. O., A. Rogozan, C. Chatelain, F. Nashashibi, and A. Bensrhair (2019). “Multi-task deep learning for pedestrian detection, action recognition and time to cross prediction”. *IEEE Access* vol. 7, pp. 149318–149327.

Privat, G. (2012). “Extending the Internet of things”. *Communications & Strategies*, no. 87, pp. 101–119.

Quan, R., L. Zhu, Y. Wu, and Y. Yang (2021). “Holistic LSTM for pedestrian trajectory prediction”. *IEEE transactions on image processing* vol. 30, pp. 3229–3239.

Quintero, R., I. Parra, J. Lorenzo, D. Fernández-Llorca, and M. Sotelo (2017). “Pedestrian intention recognition by means of a hidden markov model and body language”. *2017 IEEE 20th international conference on intelligent transportation systems (ITSC)*. IEEE, pp. 1–7.

Rabiner, L. R. (1989). “A tutorial on hidden Markov models and selected applications in speech recognition”. *Proceedings of the IEEE* vol. 77, no. 2, pp. 257–286.

Radac, M.-B. and R.-E. Precup (2019). “Data-Driven model-free tracking reinforcement learning control with VRFT-based adaptive actor-critic”. *Applied Sciences* vol. 9, no. 9, p. 1807.

Rasmussen, C. and G. D. Hager (1998). “Joint probabilistic techniques for tracking multi-part objects”. *Proceedings. 1998 ieee computer society conference on computer vision and pattern recognition (cat. no. 98cb36231)*. IEEE, pp. 16–21.

Rasouli, A. (2020). “Deep learning for vision-based prediction: A survey”. *arXiv preprint arXiv:2007.00095*.

Rasouli, A., I. Kotseruba, and J. K. Tsotsos (2017a). “Agreeing to cross: How drivers and pedestrians communicate”. *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, pp. 264–269.

Rasouli, A., I. Kotseruba, and J. K. Tsotsos (2017b). “Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior”. *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 206–213.

Rasouli, A., I. Kotseruba, and J. K. Tsotsos (2018). “It’s Not All About Size: On the Role of Data Properties in Pedestrian Detection”. *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0.

Rasouli, A. and J. K. Tsotsos (2019). “Autonomous vehicles that interact with pedestrians: A survey of theory and practice”. *IEEE transactions on intelligent transportation systems* vol. 21, no. 3, pp. 900–918.

Rawashdeh, Z. Y. and Z. Wang (2018a). “Collaborative automated driving: A machine learning-based method to enhance the accuracy of shared information”. *2018 21st International*

*Conference on Intelligent Transportation Systems (ITSC)*. IEEE, pp. 3961–3966.

Rawashdeh, Z. Y. and Z. Wang (2018b). “Collaborative automated driving: A machine learning-based method to enhance the accuracy of shared information”. *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, pp. 3961–3966.

Razin, Y. S., K. Pluckter, J. Ueda, and K. Feigh (2017). “Predicting task intent from surface electromyography using layered hidden Markov models”. *IEEE Robotics and Automation Letters* vol. 2, no. 2, pp. 1180–1185.

Redmon, J., S. Divvala, R. Girshick, and A. Farhadi (2016). “You only look once: Unified, real-time object detection”. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788.

Reed, M., S. Ebert, M. Jones, and B. Park (2019). “US National Highway Traffic Safety Administration”. *Washington, DC: US National Highway Traffic Safety Administration*.

Ren, S., K. He, R. Girshick, and J. Sun (2015). “Faster r-cnn: Towards real-time object detection with region proposal networks”. *Advances in neural information processing systems* vol. 28, pp. 91–99.

Renaudin, V., V. Demeule, and M. Ortiz (2013). “Adaptative pedestrian displacement estimation with a smartphone”. *International conference on indoor positioning and indoor navigation*. IEEE, pp. 1–9.

Riley, K. (2006). “Book Review: Mathematical Methods for Physics and Engineering: a Comprehensive Guide, /Cambridge University Press, 2006”. *The Observatory* vol. 126, p. 431.

Rudenko, A., L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras (2020). “Human motion trajectory prediction: A survey”. *The International Journal of Robotics Research* vol. 39, no. 8, pp. 895–935.

Saleh, K., M. Hossny, and S. Nahavandi (2019a). “Contextual recurrent predictive model for long-term intent prediction of vulnerable road users”. *IEEE Transactions on Intelligent Transportation Systems* vol. 21, no. 8, pp. 3398–3408.

Saleh, K., M. Hossny, and S. Nahavandi (2019b). “Real-time intent prediction of pedestrians for autonomous ground vehicles via spatio-temporal densenet”. *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 9704–9710.

Sammut, C. and G. Webb (2010). *Baum-welch algorithm*.

Schlosser, J., C. K. Chow, and Z. Kira (2016). “Fusing lidar and images for pedestrian detection using convolutional neural networks”. *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 2198–2205.

Schneemann, F. and P. Heinemann (2016). “Context-based detection of pedestrian crossing intention for autonomous driving in urban environments”. *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, pp. 2243–2248.

Schulter, S., P. Vernaza, W. Choi, and M. Chandraker (2017). “Deep network flow for multi-object tracking”. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6951–6960.

Schulz, A. T. and R. Stiefelhagen (2015). “A controlled interactive multiple model filter for combined pedestrian intention recognition and path prediction”. *2015 IEEE 18th International Conference on Intelligent Transportation Systems*. IEEE, pp. 173–178.

Shit, R. C. and S. Sharma (2018). “Localization for Autonomous Vehicle: Analysis of Importance of IoT Network Localization for Autonomous Vehicle Applications”. *2018 International Conference on Applied Electromagnetics, Signal Processing and Communication (AESPC)*. Vol. 1. IEEE, pp. 1–6.

Sidenbladh, H., M. J. Black, and D. J. Fleet (2000). “Stochastic tracking of 3D human figures using 2D image motion”. *European conference on computer vision*. Springer, pp. 702–718.

Simonyan, K. and A. Zisserman (2014). “Very deep convolutional networks for large-scale image recognition”. *arXiv preprint arXiv:1409.1556*.

Singh, V. K., B. Wu, and R. Nevatia (2008). “Pedestrian tracking by associating tracklets using detection residuals”. *2008 IEEE Workshop on Motion and video Computing*. IEEE, pp. 1–8.

Soldatos, J., N. Kefalakis, M. Hauswirth, M. Serrano, J.-P. Calbimonte, M. Riahi, K. Aberer, P. P. Jayaraman, A. Zaslavsky, I. P. Žarko, et al. (2015). “Openiot: Open source internet-of-things in the cloud”. *Interoperability and open-source solutions for the internet of things*. Springer, pp. 13–25.

Solmaz, G., E. L. Berz, M. F. Dolatabadi, S. Aytaç, J. Fürst, B. Cheng, and J. d. Ouden (2019). “Learn from IoT: pedestrian detection and intention prediction for autonomous driving”. *Proceedings of the 1st ACM Workshop on Emerging Smart Technologies and Infrastructures for Smart Mobility and Sustainability*, pp. 27–32.

Song, W., G. Xiong, and H. Chen (2016). “Intention-aware autonomous driving decision-making in an uncontrolled intersection”. *Mathematical Problems in Engineering* vol. 2016.

Spincemaille, P., T. D. Nguyen, M. R. Prince, and Y. Wang (2008). “Kalman filtering for real-time navigator processing”. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* vol. 60, no. 1, pp. 158–168.

Steinbring, J., C. Mandery, F. Pfaff, F. Faion, T. Asfour, and U. D. Hanebeck (2016). “Real-time whole-body human motion tracking based on unlabeled markers”. *2016 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE, pp. 583–590.

Streit, R. L. and T. E. Luginbuhl (1995). *Probabilistic multi-hypothesis tracking*. Tech. rep. NAVAL UNDERWATER SYSTEMS CENTER NEWPORT RI.

Streubel, R. and B. Yang (2016). “Fusion of stereo camera and MIMO-FMCW radar for pedestrian tracking in indoor environments”. *2016 19th International Conference on Information Fusion (FUSION)*. IEEE, pp. 565–572.

Styles, O., A. Ross, and V. Sanchez (2019). “Forecasting pedestrian trajectory with machine-annotated training data”. *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, pp. 716–721.

Su, H., S. Maji, E. Kalogerakis, and E. Learned-Miller (2015). “Multi-view convolutional neural networks for 3d shape recognition”. *Proceedings of the IEEE international conference on computer vision*, pp. 945–953.

Sun, P. and A. Boukerche (2020a). "A novel internet-of-vehicles assisted collaborative low-visible pedestrian detection approach". *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, pp. 1–6.

Sun, P. and A. Boukerche (2020b). "Challenges and potential solutions for designing a practical pedestrian detection framework for supporting autonomous driving". *Proceedings of the 18th ACM Symposium on Mobility Management and Wireless Access*, pp. 75–82.

Sun, Z., J. Chen, L. Chao, W. Ruan, and M. Mukherjee (2020). "A survey of multiple pedestrian tracking based on tracking-by-detection framework". *IEEE Transactions on Circuits and Systems for Video Technology* vol. 31, no. 5, pp. 1819–1833.

Sundaresan, A., R. Chellappa, and R. RoyChowdhury (2004). "Multiple view tracking of humans modelled by kinematic chains". *2004 International Conference on Image Processing, 2004. ICIP'04.* Vol. 2. IEEE, pp. 1009–1012.

Swalaganata, G., Y. Affriyenni, et al. (2018). "Moving object tracking using hybrid method". *2018 International Conference on Information and Communications Technology (ICOIACT)*. IEEE, pp. 607–611.

Sztyler, T., H. Stuckenschmidt, and W. Petrich (2017). "Position-aware activity recognition with wearable devices". *Pervasive and mobile computing* vol. 38, pp. 281–295.

Troje, N. F. (2002). "Decomposing biological motion: A framework for analysis and synthesis of human gait patterns". *Journal of vision* vol. 2, no. 5, pp. 2–2.

Tsang, D. J., M. Lukac, and A. E. Martin (2019). "Characterization of statistical persistence in joint angle variation during walking". *Human movement science* vol. 68, p. 102528.

Van, L.-D., L.-Y. Zhang, C.-H. Chang, K.-L. Tong, K.-R. Wu, and Y.-C. Tseng (2021). "Things in the air: tagging wearable IoT information on drone videos". *Discover Internet of Things* vol. 1, no. 1, pp. 1–13.

Vasquez, D., T. Fraichard, and C. Laugier (2009). "Incremental learning of statistical motion patterns with growing hidden markov models". *IEEE Transactions on Intelligent Transportation Systems* vol. 10, no. 3, pp. 403–416.

Vermesan, O., P. Friess, et al. (2014). *Internet of things-from research and innovation to market deployment*. Vol. 29. River publishers Aalborg.

Vo, B.-N., S. Singh, and A. Doucet (2005). "Sequential Monte Carlo methods for multitarget filtering with random finite sets". *IEEE Transactions on Aerospace and electronic systems* vol. 41, no. 4, pp. 1224–1245.

Völz, B., H. Mielenz, G. Agamennoni, and R. Siegwart (2015). "Feature relevance estimation for learning pedestrian behavior at crosswalks". *2015 IEEE 18th International Conference on Intelligent Transportation Systems*. IEEE, pp. 854–860.

Waddell, M. L. and E. L. Amazeen (2017). "Evaluating the contributions of muscle activity and joint kinematics to weight perception across multiple joints". *Experimental brain research* vol. 235, no. 8, pp. 2437–2448.

Wakim, C. F., S. Capperon, and J. Oksman (2004). "A Markovian model of pedestrian behavior". *2004 ieee international conference on systems, man and cybernetics (ieee cat. no. 04ch37583)*. Vol. 4. IEEE, pp. 4028–4033.

Wang, H., Z. Deng, B. Feng, H. Ma, and Y. Xia (2017). "An adaptive Kalman filter estimating process noise covariance". *Neurocomputing* vol. 223, pp. 12–17.

Wang, Y., X. Wang, D. Tian, X. Duan, H. Liu, Y. Gong, Z. Sheng, and V. C. Leung (2019). "A Multi-object Detection Method Based on Connected Vehicles". *Proceedings of the 9th ACM Symposium on Design and Analysis of Intelligent Vehicular Networks and Applications*, pp. 89–96.

Wang, Z. and N. Papanikolopoulos (2020). "Estimating pedestrian crossing states based on single 2D body pose". *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 2205–2210.

Wasik, A., P. U. Lima, and A. Martinoli (2020). "A robust localization system for multi-robot formations based on an extension of a Gaussian mixture probability hypothesis density filter". *Autonomous Robots* vol. 44, no. 3, pp. 395–414.

Wieser, A., M. G. Petovello, and G. Lachapelle (2004). "Failure scenarios to be considered with kinematic high precision relative GNSS positioning". *Proceedings of the 17th International*

*Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS 2004)*, pp. 1448–1459.

Wissel, T., T. Pfeiffer, R. Frysch, R. T. Knight, E. F. Chang, H. Hinrichs, J. W. Rieger, and G. Rose (2013). "Hidden Markov model and support vector machine based decoding of finger movements using electrocorticography". *Journal of neural engineering* vol. 10, no. 5, p. 056020.

Wongthongtham, P., J. Kaur, V. Potdar, and A. Das (2017). "Big data challenges for the Internet of Things (IoT) paradigm". *Connected Environments for the Internet of Things*. Springer, pp. 41–62.

Wu, J., W. Song, X. Lai, and X. Li (2020). "Upper Arm Action Recognition for Self Training with a Smartphone". *Journal of Physics: Conference Series*. Vol. 1616. 1. IOP Publishing, p. 012102.

Wu, J., J. Ruenz, and M. Althoff (2018). "Probabilistic map-based pedestrian motion prediction taking traffic participants into consideration". *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, pp. 1285–1292.

Xie, C., J. Tan, L. Zhou, L. He, J. Zhang, and Y. Bu (2012). "A Joint Object Tracking Framework with Incremental and Multiple Instance Learning". *2012 Fourth International Conference on Digital Home*. IEEE, pp. 7–12.

Yan, L., R. S. Allison, and S. K. Rushton (2004). "New simple virtual walking method-walking on the spot". *Proceedings of the IPT Symposium*. Citeseer, pp. 1–7.

Yang, B., W. Zhan, P. Wang, C. Chan, Y. Cai, and N. Wang (2021). "Crossing or Not? Context-Based Recognition of Pedestrian Crossing Intention in the Urban Environment". *IEEE Transactions on Intelligent Transportation Systems*.

Yang, D., H. Zhang, E. Yurtsever, K. Redmill, and Ü. Özgüner (2021). "Predicting Pedestrian Crossing Intention with Feature Fusion and Spatio-Temporal Attention". *arXiv preprint arXiv:2104.05485*.

Yang, F., H. Lu, and M.-H. Yang (2013). "Robust visual tracking via multiple kernel boosting with affinity constraints". *IEEE Transactions on Circuits and Systems for Video Technology* vol. 24, no. 2, pp. 242–254.

Yi, S., H. Li, and X. Wang (2015). “Understanding pedestrian behaviors from stationary crowd groups”. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3488–3496.

Yoon, Y.-C., D. Y. Kim, Y.-m. Song, K. Yoon, and M. Jeon (2021). “Online multiple pedestrians tracking using deep temporal appearance matching association”. *Information Sciences* vol. 561, pp. 326–351.

Zhang, L. and L. van der Maaten (2013). “Structure preserving object tracking”. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1838–1845.

Zhang, L. and L. Van Der Maaten (2013). “Preserving structure in model-free tracking”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* vol. 36, no. 4, pp. 756–769.

Zhang, S., M. Abdel-Aty, Y. Wu, and O. Zheng (2021). “Pedestrian Crossing Intention Prediction at Red-Light Using Pose Estimation”. *IEEE Transactions on Intelligent Transportation Systems*.

Zhang, Y., K. Chen, and H.-Z. Tan (2009). “Performance analysis of gradient neural network exploited for online time-varying matrix inversion”. *IEEE Transactions on Automatic Control* vol. 54, no. 8, pp. 1940–1945.

Zhao, H. and R. Shibasaki (2005). “A novel system for tracking pedestrians using multiple single-row laser-range scanners”. *IEEE Transactions on systems, man, and cybernetics-Part A: systems and humans* vol. 35, no. 2, pp. 283–291.

Zhao, X., P. Sun, Z. Xu, H. Min, and H. Yu (2020). “Fusion of 3D LIDAR and camera data for object detection in autonomous vehicle applications”. *IEEE Sensors Journal* vol. 20, no. 9, pp. 4901–4913.

Zhou, X., V. Koltun, and P. Krähenbühl (2020). “Tracking objects as points”. *European Conference on Computer Vision*. Springer, pp. 474–490.

Zhuang, B., H. Lu, Z. Xiao, and D. Wang (2014). “Visual tracking via discriminative sparse similarity map”. *IEEE Transactions on Image Processing* vol. 23, no. 4, pp. 1872–1881.

Zou, Y., W. Zhang, W. Weng, and Z. Meng (2019). "Multi-vehicle tracking via real-time detection probes and a markov decision process policy". *Sensors* vol. 19, no. 6, p. 1309.

*There is no end to stories!*

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY