Context-aware Human Motion Prediction for Robots in Complex Dynamic Environments

Örebro Studies in Technology 91

Andrey Rudenko

Context-aware Human Motion Prediction for Robots in Complex Dynamic Environments

Title: Context-aware Human Motion Prediction for Robots in Complex Dynamic Environments

Publisher: Örebro University, 2021
www.oru.se/publikationer

Printer: Örebro University/Repro 06-2021

ISSN 1650-8580
ISBN 978-91-7529-392-9

Abstract

Andrey Rudenko (2021): Context-aware Human Motion Prediction for Robots in Complex Dynamic Environments. Örebro Studies in Technology 91.

Understanding human behavior is a key skill for intelligent systems that share physical and emotional spaces with humans. One of the main challenges to this end is the ability of such systems to make accurate predictions of human motion. This is a difficult task as human motion is influenced by a large variety of internal and external stimuli, such as own actions, the presence and actions of surrounding agents, social relations, rules and norms between them, or the environment with its topology, geometry, semantics and affordances.

This thesis systematically addresses human motion prediction for autonomous systems by surveying the field, the different requirements to the prediction task, problem formulations and solution classes, and its application domains. Overviewing three decades of prior research from different communities, this thesis proposes a unifying taxonomy for motion prediction methods based on the modeling approach and level of contextual information used, and provides a review of the existing datasets and performance metrics. Furthermore, it discusses limitations of the state of the art and outlines directions for further research.

Predicting human motion in complex dynamic and cluttered environments is particularly challenging due to the high level of required contextual awareness. To acquire, represent and incorporate a large variety of contextual cues is still an open challenge which is why in this thesis, we also make several methodological contributions. We present a planning-based approach that accounts for maps of obstacles and local interactions with social grouping constraints. This method accommodates many desired properties, such as predicting for an arbitrary number of observed people, estimating multi-modal probability distributions, reasoning over intentions, and supporting semantic map input. Apart from reaching state-of-the-art performance, this single method bridges the gap between short-term motion prediction, where social interaction is the most informative cue, and long-term prediction, where goal-orientation and obstacle geometry typically determine people's motion trajectories.

Along the same line, and in addition to contextual cues of the dynamic environment and the topometric map, semantic information about the environment is a highly informative cue for motion prediction. We address the less explored problem of predicting collision risks by inferring occupancy priors of human motion using only semantic maps as input. The proposed method, based on Convolutional Neural Networks, shows superior performance over the state of the art and demonstrates a novel way to use and apply semantics for the prediction task.

Datasets that contain relevant qualities and quantities of difficulty are critical for benchmarking autonomous systems in general and for motion predic-

tion in particular. Surprisingly, the commonly used datasets are rather limited in that they typically consider simple to almost trivial scenarios, contain little contextual cues and partly suffer from annotation issues. To address these issues, this thesis proposes a weakly-scripted data collection protocol for recording diverse and accurate trajectories of people and robots in interactive scenarios. The protocol includes social roles with simple instructions for the participants, dynamically-allocated goals, group motion and varied obstacle positioning. The data, recorded according to the introduced collection protocol, is used in a motion prediction benchmark, designed for thorough performance evaluation in a variety of experiments: accuracy conditioned on several key factors (e.g. prediction horizon, observation length), evaluation of knowledge transfer to a new environment, testing robustness against perception noise.

The results presented in this thesis are relevant for a broad range of prediction problems with applications in robotics, autonomous driving or video surveillance. With the first systematic taxonomy of prediction approaches, new experiments for benchmarking and novel methods that account for particularly rich contextual cues, we contribute to the field by fostering cross-domain exchange and comparison, and by laying the foundations for various directions of future research.

Keywords: robotics, human motion prediction, activity forecasting

Andrey Rudenko, School of Science and Technology
Örebro University, SE-701 82 Örebro, Sweden, andrey.rudenko@oru.se

Acknowledgements

Completing this doctoral dissertation over the course of four years marks a major milestone in my scientific career. This accomplishment, which I am proud and humbled to present, I could not have achieved without the guidance and support of many wonderful people, to whom I would like to express my sincere gratitude.

First, I would like to thank my long-standing mentor, colleague and friend Luigi Palmieri for his outstanding input into this dissertation, support and guidance. His honest, rational and refined conduct in matters of science and beyond made a profound impact on my professional development.

I am greatly indebted to my advisors Kai Arras and Achim Lilienthal for the years of their invaluable feedback, insightful discussions, encouragement and criticisms. It is with their competence, expertise and example I was so fortunate to grow stronger as a researcher.

I thank my co-authors Michael Herman, Kris Kitani, Dariu Gavrila, Tomasz Kucner, Ravi Chadalavada, Chittaranjan Swaminathan, Johannes Döllinger and my student Wanting Huang for the wonderful manuscripts we created together, insightful discussions and their valuable feedback.

I thank Jim Mainprice, Alexandre Alahi, Lamberto Ballan, Pasquale Coscia and Andrea Bajcsy for their collaboration and efforts in making our motion prediction workshops such special events.

I am thankful to my colleagues Timm Linder and Narunas Vaskevicius for their expertise on scientific and technical topics, and valuable comments to my research.

I thank my Bosch colleagues for an engaging, motivating and welcoming working environment: Robert Schirmer for his enduring optimism, Marco Lampacrescia for his weighted judgment, Mirco Colosi for always keeping the high level of discussion, Sergey Alatartsev for his excellent original presentations, Musa, Max, Sebastian and many others.

A big thanks goes to my PhD buddies in the Örebro university: Dino Hüllmann, Han Fan, Basiliki Kondili, Jiawei Hou and Malcolm Miele for all the jolly chatter, skiing trips, fikas, and timebeers.

Finally, I am deeply grateful to my parents for all the love and encouragement on this journey.

Lastly, I thankfully acknowledge that this thesis was partially supported by the European Commission under grant agreement number H2020-ICT-2016-732737 (ILIAD).

Contents

1 Introduction **1**
- 1.1 Problem Statement and Terminology . . . 2
- 1.2 Motivation . . . 4
- 1.3 Application Domains . . . 6
  - 1.3.1 Service Robots . . . 6
  - 1.3.2 Self-driving Vehicles . . . 7
  - 1.3.3 Surveillance . . . 7
- 1.4 Research Question and Contributions . . . 8
- 1.5 Publications . . . 12
- 1.6 Dissemination . . . 13
- 1.7 Outline . . . 14
- 1.8 Ethical Considerations . . . 15

2 Motion Prediction Review **17**
- 2.1 Introduction . . . 17
- 2.2 Related Reviews . . . 19
- 2.3 Taxonomy . . . 21
  - 2.3.1 Modeling Approach . . . 22
  - 2.3.2 Contextual Cues . . . 23
  - 2.3.3 Classification Rules . . . 24
- 2.4 Physics-based Approaches . . . 26
  - 2.4.1 Single-model Approaches . . . 26
  - 2.4.2 Multi-model Approaches . . . 32
- 2.5 Pattern-based Approaches . . . 34
  - 2.5.1 Sequential Models . . . 36
  - 2.5.2 Non-sequential Models . . . 41
- 2.6 Planning-based Approaches . . . 43
  - 2.6.1 Forward Planning Approaches . . . 43
  - 2.6.2 Inverse Planning Approaches . . . 46
- 2.7 Contextual Cues . . . 49
  - 2.7.1 Cues of the Target Agent . . . 49

2.7.2 Cues of Other Dynamic Agents . . . 50
2.7.3 Cues of the Static Environment . . . 52
2.8 Motion Prediction Evaluation . . . 53
2.8.1 Performance Metrics . . . 53
2.8.2 Datasets . . . 58
2.9 Discussion . . . 62
2.9.1 Benchmarking . . . 62
2.9.2 Modeling Approaches . . . 64
2.9.3 Application Domains . . . 66
2.10 Conclusions and Outlook . . . 69

3 Interaction-aware Planning-based Trajectory Prediction 71
3.1 Introduction . . . 72
3.1.1 Contribution . . . 73
3.1.2 Outline . . . 75
3.2 Joint sampling MDP for Motion Prediction . . . 75
3.2.1 Problem Formulation . . . 75
3.2.2 Markov Decision Process Notation . . . 76
3.2.3 MDP for Global Motion Prediction . . . 76
3.2.4 Joint Human Motion Prediction with Group Social Forces 78
3.2.5 Stochastic Policy Sampling Using Random Walks . . . 82
3.2.6 Complexity Analysis . . . 85
3.2.7 Implementation Details . . . 86
3.3 Experiments . . . 87
3.3.1 Environments with no Groups . . . 88
3.3.2 Experiments with Groups . . . 89
3.4 Conclusion and Outlook . . . 98
3.4.1 Semantic Context-awareness . . . 98
3.4.2 Combination with a Pattern-based Interaction Model . . 99
3.4.3 Robot Motion Planning Using Predictions . . . 99

4 Occupancy Priors of Human Motion in Urban Environments 101
4.1 Introduction . . . 101
4.1.1 Contribution . . . 103
4.1.2 Outline . . . 103
4.2 Related Work . . . 104
4.3 IOC and CNN Approaches for Occupancy Priors Estimation . . 105
4.3.1 Inverse Optimal Control on Multiple Maps (IOCMM) . 105
4.3.2 Semantic Map-Aware Pedestrian Prediction (semapp) . . 109
4.4 Experiments . . . 111
4.4.1 Datasets . . . 111
4.4.2 Training and Evaluation . . . 113
4.4.3 A Remark on Semantic Segmentation . . . 115

4.5 Results . . . 115
4.6 Conclusions and Outlook . . . 116

5 Data Collection for Motion Prediction 119
5.1 Introduction . . . 119
5.1.1 Contribution . . . 122
5.1.2 Outline . . . 122
5.2 Related Work . . . 122
5.3 Data Collection Procedure . . . 123
5.3.1 System Setup . . . 124
5.3.2 Scenario Description and Participants' Priming . . . 126
5.4 Results and Analysis . . . 128
5.4.1 Data Description . . . 128
5.4.2 Baselines and Metrics . . . 129
5.4.3 Results . . . 130
5.5 Conclusions and Outlook . . . 131

6 Benchmarking Human Motion Prediction Methods 137
6.1 Introduction . . . 137
6.1.1 Contribution . . . 138
6.1.2 Outline . . . 139
6.2 Background . . . 139
6.3 Our Benchmark Description . . . 142
6.3.1 Datasets . . . 143
6.3.2 Preprocessing . . . 144
6.3.3 Prediction . . . 144
6.3.4 Evaluation . . . 145
6.3.5 Experiments . . . 146
6.4 Case Study: Benchmarking Local Interaction Models . . . 148
6.4.1 Predictive Social Force Model [356] . . . 148
6.4.2 Predictive Collision Avoidance Model [141] . . . 150
6.4.3 Results and Discussion . . . 153
6.5 Conclusions and Outlook . . . 154

7 Conclusions 157
7.1 Contributions . . . 157
7.2 Open Challenges and Future Research Directions . . . 160
7.2.1 Use of enhanced contextual cues . . . 160
7.2.2 Robustness and Integration . . . 161
7.3 Ongoing and Future Work . . . 162
7.3.1 Benchmarking Trajectory Forecasting Methods . . . 162
7.3.2 Hierarchical Predictive Planning System . . . 163

References 165

List of Figures

1.1 Basic elements of a motion prediction system 4
1.2 Application domains of human motion prediction 6
1.3 Prediction application scenario in the airport environment . . . 8
1.4 Application of motion prediction in intralogistics settings 9

2.1 Overview of the categories in our taxonomy 19
2.2 Publications trends in the literature on motion prediction 20
2.3 Basic working principle of the modeling approaches 22
2.4 Dynamic environment cues 24
2.5 Static environment cues 25
2.6 Physics-based approaches to model motion 27
2.7 Pattern-based approaches to model motion 35
2.8 Planning-based approaches to model motion 44

3.1 Prediction results in the ATC shopping center dataset 73
3.2 Illustrating example of group motion in a shopping mall 74
3.3 Social force model parameters 79
3.4 Groups social force model parameters 80
3.5 Summary of our MDP-based approach for motion prediction . . 81
3.6 Social interaction prediction with our random walk algorithm . 83
3.7 Illustration of the random walk stochastic policy sampling . . . 85
3.8 Prediction results in simulated scenarios with no groups (part 1) 91
3.9 Prediction results in simulated scenarios with no groups (part 2) 92
3.10 NLP and MHD evaluation results in scenarios with no groups . 93
3.11 Runtime of our algorithm in scenarios with no groups 93
3.12 Prediction results in a simulated scenario with groups 94
3.13 Prediction results in a challenging simulated scenario with groups 95
3.14 NLP and MHD evaluation results in scenarios with groups . . . 96
3.15 Runtime of our algorithm in scenarios with groups 97
3.16 Challenging changes of motion dynamics in the ATC dataset . . 98
3.17 Predictive planning example 100

4.1 Predicting occupancy priors in urban environments 102
4.2 Structure of the *semapp* network 109
4.3 CNN training and inference using crops from larger maps . . . 110
4.4 Training examples from the U4 dataset 113
4.5 Semantic segmentation in the Stanford Drone Dataset 114
4.6 Comparison of the occupancy prior prediction results 117
4.7 Optimal θ costs of various semantic classes, learned by IOCMM 118
4.8 Runtime and performance of IOCMM and semapp 118

5.1 THÖR recording environment configuration 120
5.2 Overview of the data collection environment 124
5.3 Equipment used for data collection 125
5.4 Roles of the participants and their expected motion patterns . . 127
5.5 Recorded trajectories 132
5.6 Social interactions in the THÖR dataset (1) 133
5.7 Social interactions in the THÖR dataset (2) 134
5.8 Social interactions in the THÖR dataset (3) 135

6.1 Example of data pre-processing in the Atlas benchmark 141
6.2 Altas benchmark design 143
6.3 Data pre-processing in the Atlas benchmark 145
6.4 Synthetic testing scenarios in the Atlas benchmark 146
6.5 A testing scenario from the ATC dataset 147
6.6 The force projection principle in Zanlungo et al. [356] 149
6.7 The distance-force relation in the model by Karamouzas et al. [141] 151
6.8 Observation length variation experiment in the ATC dataset . . 153
6.9 Observation length variation experiment in the THÖR dataset . 154
6.10 Robustness experiment in the ETH dataset 156
6.11 Robustness experiment in the THÖR dataset 156

7.1 Aspects in benchmarking trajectory prediction 163
7.2 Hierarchical predictive planning system 164

List of Tables

2.1 Metrics to evaluate motion prediction 56
2.2 Overview of the motion trajectories datasets (part 1) 59
2.3 Overview of the motion trajectories datasets (part 2) 60
2.4 Additional motion trajectories datasets 61

3.1 Estimated hyperparameters in experiments with no groups . . . 89
3.2 Estimated hyperparameters in experiments with groups 90

4.1 Datasets summary 111
4.2 IOCMM and semapp training and inference parameters 112
4.3 Average KL-div in the U4 and Stanford Drone datasets 116

5.1 Contextual cues in the datasets of human motion trajectories . . 121
5.2 Details of the data recording 128
5.3 Comparison of the datasets 130

6.1 Benchmarks for human motion prediction 140
6.2 ADE in the ETH dataset with different prediction horizons . . . 152
6.3 FDE in the ETH dataset with different prediction horizons . . . 152
6.4 ADE in the THÖR dataset with different prediction horizons . . 152
6.5 FDE in the THÖR dataset with different prediction horizons . . 153
6.6 ADE measured in the transfer experiments on different datasets 155
6.7 FDE measured in the transfer experiments on different datasets . 155

List of Algorithms

1 Joint Random Walk Stochastic Policy Sampling 84
2 Group Social Force MDP Motion Prediction 85
3 Joint Stochastic Policy Sampling Operation Analysis 86
4 Inverse Optimal Control: Backward pass 106
5 Inverse Optimal Control: Forward pass 107
6 Inverse Optimal Control on Multiple Maps (IOCMM) 108

List of papers

Paper I A. Rudenko, L. Palmieri, and K. O. Arras. Predictive planning for a mobile robot in human environments. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA), Workshop on AI Planning and Robotics*, 2017

Paper II A. Rudenko, L. Palmieri, and K. O. Arras. Joint prediction of human motion using a planning-based social force approach. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 1–7, 2018

Paper III A. Rudenko, L. Palmieri, A. J. Lilienthal, and K. O. Arras. Human motion prediction under social grouping constraints. In *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*, 2018

Paper IV A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras. Human motion trajectory prediction: A survey. *Int. J. of Robotics Research*, 39(8):895–935, 2020

Paper V A. Rudenko, T. P. Kucner, C. S. Swaminathan, R. T. Chadalavada, K. O. Arras, and A. J. Lilienthal. THÖR: Human-robot navigation data collection and accurate motion trajectories dataset. *IEEE Robotics and Automation Letters*, 5(2):676–682, 2020

Paper VI A. Rudenko, T. Kucner, C. Swaminathan, R. Chadalavada, K. O. Arras, and A. J. Lilienthal. Benchmarking human motion prediction methods. In *Proc. of the ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI), Workshop on Test Methods and Metrics for Effective HRI in Real World Human-Robot Teams*, 2020

Paper VII A. Rudenko, L. Palmieri, J. Doellinger, A. J. Lilienthal, and K. O. Arras. Learning occupancy priors of human motion from semantic maps of urban environments. *IEEE Robotics and Automation Letters*, 6(2):3248–3255, 2021

Chapter 1
Introduction

Eppur si muove!

Galileo Galilei

The research in robotics, automation and artificial intelligence has greatly accelerated in the 21st century, taking a well-established shape in the higher education programs, university laboratories and industrial R&D departments alike. Several key directions include perception, localization and mapping, control and motion planning, grasping and manipulation, human-robot interaction. Overarching the entirety of robotics research is the figure of the human – the customer, the operator, the teammate and the central object of service robotics.

The role of humans, and consequently human-oriented robotics research, is steadily increasing. It is hardly surprising, as the more advanced robotic technology gets, the more enabled and economically justified practical application of service robots becomes. When robots become products and service providers, they leave the structured and protected laboratories, and their behavior is no longer overwatched by expert engineers. Autonomous operation in social environments puts enormous expectations on the performance of the robots in terms of their safety and efficiency.

Understanding human behavior is a key skill for intelligent systems to coexist and interact with humans. It involves aspects in representation, perception and motion analysis. Prediction plays an important part in human motion analysis: foreseeing how a scene involving multiple agents will unfold over time allows to incorporate this knowledge in a pro-active manner, i.e. allowing for enhanced ways of active perception, predictive planning, model predictive control, or human-robot interaction.

Human motion in robotics comes in many forms: articulated full-body motion, gestures and facial expressions, movement through space by walking, using a mobility device or driving a vehicle. Working with these forms of human

motion reveals many similarities, as they require to define, abstract, model, validate and integrate spatial and temporal aspects of human motion in a complex dynamic world. As such, this domain can be roughly enveloped under the term *human motion prediction*.

Still a young branch of robotics research, prediction of human motion is actively shaping today. It is placed in the hot spot between perception and planning, between autonomous vehicles and crowd-navigating robots, between social proxemics theories, biomechanical understanding of motion and machine learning of motion patterns. Driven by the remarkable success of the deep learning methods in perception, advances in automated driving technology and growing interest to manufacturing automation, robotic systems are becoming more reliable to be deployed in human environments. The focus shifts from purely technical aspects of robot operation to how they should *interact* and *behave* in complex human environments. This is where motion prediction and behavior understanding become critical components of an autonomous system.

The challenge of making accurate predictions of human motion arises from the complexity of human behavior and the variety of its internal and external stimuli. Motion behavior may be driven by the target agent's own goal intent, the presence and actions of surrounding agents, social relations between agents, social rules and norms, or the environment with its topology, geometry, affordances and semantics. Most factors are not directly observable and need to be inferred from noisy perceptual cues or modeled from context information. Furthermore, to be effective in practice, motion prediction should be robust and operate in real-time.

This thesis makes a broad attempt to unify and offer a holistic view on several key aspects of motion prediction in robotics. It explores the fundamentals of human motion prediction for autonomous systems, ranging from surveying the complete methodology, tasks and application scenarios, aspects in data collection and method development, to evaluation, benchmarking and integration. Building a frame of reference in the motion prediction domains, it studies the state of the art against several research questions, and outlines the trends and open research questions. Furthermore, it presents several contributions to these open questions in methodology, data collection, benchmarking and semantic awareness of the intelligent systems. Eventually, this work lays foundations for the spanning tree of future research directions.

In the remainder of this chapter we define and discuss the problem in more detail, present the application areas, define the research question and outline the contributions.

1.1 Problem Statement and Terminology

In the most general form, the problem of motion prediction can be formulated as follows:

Given the current state of the target agent and its environment, motion prediction makes a hypothesis about the future state of the agent.

On the highest level of abstraction, the motion prediction problem contains the following three elements (Fig. 1.1):

- *Stimuli:* Internal and external stimuli that determine motion behavior include the agents' motion intent and other directly or indirectly observable influences. Most prediction methods rely on observed partial trajectories, or generally, sequences of agent state observations such as positions, velocities, body joint angles or attributes. Often, this is provided by a target tracking system and it is common to assume correct track identity over the observation period. Other forms of inputs include contextual cues from the environment such as scene geometry, semantics, or cues that relate to other moving entities in the surrounding. End-to-end approaches rely on sequences of raw sensor data.

- *Modeling approach:* Approaches to human motion prediction differ in the way they represent, parametrize, learn and solve the task. This thesis in part focuses on finding and analyzing useful categories, hidden similarities, common assumptions and best evaluation practices in the growing body of literature.

- *Prediction:* Different methods produce various parametric, non-parametric or structured forms of predictions, such as Gaussians over agent states, probability distributions over grids, singular or multiple trajectory samples or motion patterns using graphical models.

Throughout this thesis, various aspects of motion prediction are discussed. We use the term *agent* to denote dynamic objects of interest such as robots, pedestrians, cyclists, cars or other human-driven vehicles. The *target agent* is the dynamic object for which we make the actual motion prediction. We assume the agent behavior to be non-erratic and goal-directed with regard to an optimal or near-optimal expected outcome. This assumption is typical as the motion prediction problem were much harder or even ill-posed otherwise. We call the underlying hidden motivation of the agent a *goal*, *intent* or *intention*, and use the same term to describe the terminal state of the trajectory, or the target region in space.

By *prediction* or *forecast* we understand any hypothesis about the future state/location/configuration of the agent and/or their intent. *Prediction horizon* is the future point in time until which predictions are made starting from the current time instance. Based on the length of the prediction horizon, we informally distinguish *short-term* (up to 1-2 seconds ahead) and *long-term* (up to 20 seconds ahead) predictions. *Observed trajectory*, *track* or *tracklet* of the agent

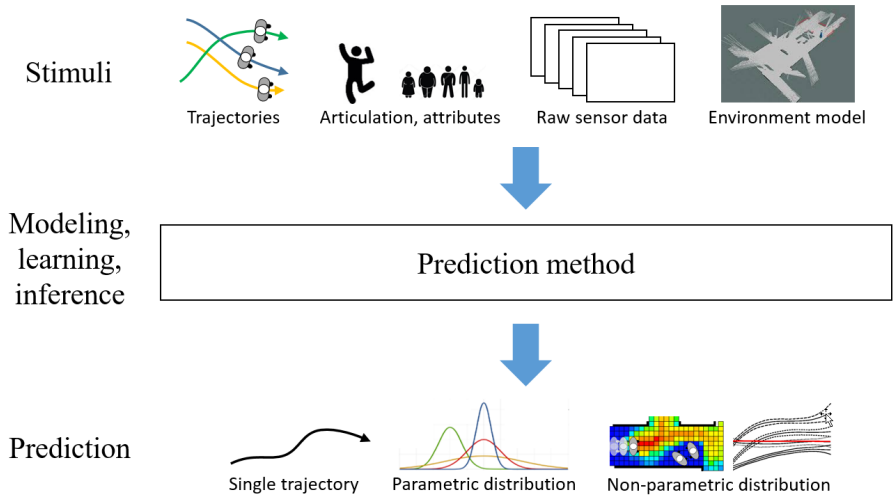


Figure 1.1: Typical elements of a motion prediction system: internal and external stimuli that influence motion behavior, the method itself and the different parametric, non-parametric or structured forms of predictions.

is the sequence of states and/or actions which the agent performed until the current time instance. The term *joint* predictions is used when motion of the target agent is assumed to be cross-influenced by other agents nearby.

We define a *path* to be a sequence of positions, often two-dimensional over the ground plane, and a *trajectory* to be a path combined with a timing law or a velocity profile. Formally, we denote $\mathbf{s}_t$ as the state of an agent at time t, $\mathbf{u}_t$ as the action that the agent takes at time t, $\mathbf{o}_t \in \mathcal{O}$ as the observations of the agent's state at time t, and use ζ to denote trajectories. We refer to a history of several states, actions or observations from time t to time T using subscripts $t:T$.

1.2 Motivation

The formulation of the motion prediction problem, presented in Sec. 1.1, is general enough to apply to many relevant forms of motion and types of agents in various domains, for instance:

- a social robot, predicting the future trajectories of surrounding people to avoid collisions,
- an automated surveillance system, predicting future positions to re-identify tracked objects between the fields of view of the sensors,

- a stationary manufacturing robot, predicting the presence of human co-workers from distributed sensors to safely control and limit the maximum permitted velocity,
- an autonomous vehicle, predicting the intentions, maneuvers and positions of other driving vehicles,
- a virtual reality helmet, predicting the head motions and preemptively processing the virtual image.

Some of these applications benefit greatly from motion prediction, while others are unimaginable without it. For example, next state prediction has long been an integral part of the people tracking systems to assign new observations to the existing tracks. Similarly, a fully autonomous vehicle, such as the one depicted in Fig. 1.2 (top left), won't be able to progress without some form of reasoning on whether the pedestrian is intending to cross the road. Mobile robots, on the other hand, often treat the motion of people as unmodeled uncertainty, projected onto the dynamic occupancy map, and execute a trajectory replanning cycle once the updated positions of the people are available. This approach, apart from discarding any possibility of pre-planned social behavior on the robot's part, is limited in two key problems: "freezing" and "dancing" behaviors of the robot [313]. Switching between the often homotopically distinct optimal trajectories through the crowd (so-called "dancing"), the approach eventually fails to plan a safe path once the crowd density exceeds a certain threshold (thus "freezing" in place).

Designing a motion predictor for any of the tasks listed above raises similar problems: how to formulate the *state* of the target agent? What is the available range of *actions*, what are the mechanical and biological constraints of that agent? What is the *static* and *dynamic environment* in which the agent operates? What is the likely *goal* of the observed motion? What aspects of the motion model can be learned from observations? How to validate and compare different prediction methods? How to integrate the motion prediction into the decision making and control pipeline? Typically, the assumptions, approaches and validation methods can be transferred and adapted between these tasks.

This thesis includes an analysis and discussion for the range of problems in motion prediction, described above, and many of the findings are relevant for a broad scope of tasks and application areas. For instance, Chapter 5 presents a procedure design to collect accurate and diverse data to study human-robot interaction in intralogistics settings. This procedure can be easily adapted for other scenarios, with both a stationary or a moving robot. The categorizing taxonomy for the motion modeling and prediction approaches, presented in Chapter 2, is relevant for both service robots and autonomous vehicles. And the semantically-informed method for learning occupancy priors in urban environments from Chapter 4 is useful for mobile robots, vehicles and autonomous surveillance systems alike.

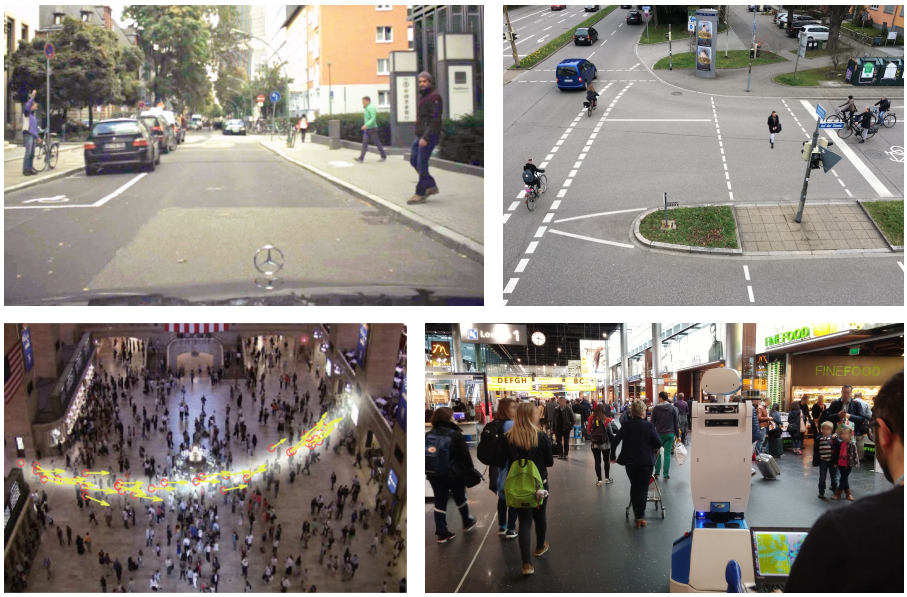


Figure 1.2: Application domains of human motion prediction. **Top left:** Will the pedestrian cross? Self-driving vehicles have to quickly reason about intentions and future locations of other traffic participants, such as pedestrians (Illustration from [157]). **Top right:** Advanced traffic surveillance systems can provide real-time alerts of pending collisions using communication technology. **Bottom left:** Advanced surveillance systems analyze human motion in public spaces for suspicious activity detection or crowd control (Illustration from [367]). **Bottom right:** Robot navigation in densely populated spaces requires accurate motion prediction of surrounding people to safely and efficiently move through crowds (Illustration featuring SPENCER robot [316]).

Following this motivation, in the next section we take a deeper look into the problem of motion prediction and specific requirements to the methods in several key application domains.

1.3 Application Domains

Motion prediction is a key task for service robots, self-driving vehicles, and advanced surveillance systems (see Fig. 1.2).

1.3.1 Service Robots

Mobile service robots increasingly operate in open-ended domestic, industrial and urban environments shared with humans. Anticipating motion of surrounding agents is an important prerequisite for safe and efficient motion planning

and human-robot interaction. Limited on-board resources for computation and first-person sensing make this a challenging task.

For example, consider SPENCER – an airport passenger guide robot, developed in the EU-funded FP7 robotics research project [316] for the Schiphol airport, see Fig. 1.2 (bottom right) and 1.3. The task of Spencer is to guide groups of people towards their destinations, e.g. the departure gates or passenger control points. The airport environment poses a serious challenge for robot navigation due to its scope, complex topology, dynamics and high-density crowds. In absence of prediction, the robot may fail to move through the crowd due to the safety constraints and uncertainty in the motion of people.

Another application example from the industry perspective is ILIAD – the EU-funded research project, aiming to explore warehouse automation with scalable fleets of intralogistic systems for environments shared with humans (see Fig 1.4). In particular, the project develops integration and operation solutions for automated industrial vehicles of various sizes, from the relatively small pallet trucks to the massive bulk loaders. This warehouse navigation scenario includes varied presence of people, complex obstacle layouts and semantically meaningful areas of the environment, such as the picking and walking areas drop-off zones, etc. The size and mass of the industrial robots introduce exceptional safety concerns, therefore perception and prediction of human movements and activity are critical.

1.3.2 Self-driving Vehicles

The ability to anticipate motion of other road users is essential for automated driving. Similar challenges apply as in the service robot domain, although they are more pronounced given the higher masses and velocities of vehicles and the resulting larger harm that can potentially be inflicted, especially towards vulnerable road users (i.e. pedestrians and cyclists). Furthermore, vehicles need to operate in rapidly changing, semantically rich outdoor traffic settings with complex iterations of heterogeneous agents. Finally, they need to comply with the hard real-time operating constraints. Knowledge of the traffic infrastructure (location of lanes, curbside, traffic signs, traffic lights, other road markings such as zebras), the traffic rules and specific dynamical properties of vehicles can help in the motion prediction.

1.3.3 Surveillance

Visual surveillance of vehicular traffic or human crowds relies on the ability to accurately track a large number of targets across distributed networks of stationary cameras. Long-term motion prediction can support a variety of surveillance tasks such as person retrieval, perimeter protection, traffic monitoring, crowd management or retail analytics by further reducing the number of false-

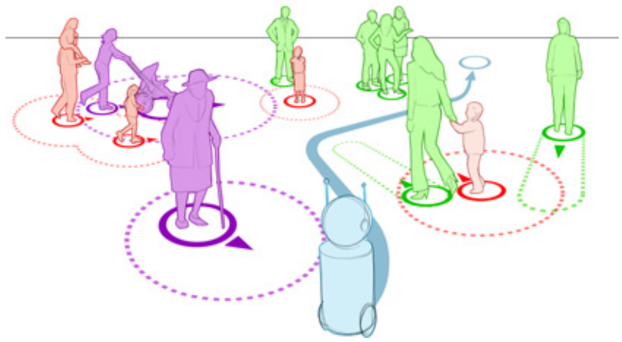


Figure 1.3: A mobile robot concept, developed in the EU-funded FP7 project SPENCER. This image shows the application of motion prediction in the airport settings for the SPENCER passenger guide robot. Moving through the crowd requires tracking and prediction of people movement for safe and efficient motion planning.

positive tracks and track identifier switches, particularly in dense crowds or across non-overlapping fields of views.

1.4 Research Question and Contributions

In the settings of service robotics and intelligent autonomous systems, described above, the research question on this thesis is formulated as follows:

How to design and validate a system to predict trajectories of people in dynamic, unstructured social environments?

Sec. 1.2 outlined a whole range of problems, which stem from this research question, on which we elaborate in this thesis.

We start our discussion with the central problem of motion prediction: how to model motion? The existing works feature a large selection of methods to this end. The choice of the model depends on many factors: the task and application scenario, required prediction horizon, relevant contextual cues, available training data and expert knowledge. The cross-disciplinary organization and detailed discussion of the existing models from the applications-centered perspective is the first contribution of this thesis.

Contribution 1: A thorough survey of motion prediction methods in which we review in detail and organize three decades of research in a taxonomy along two characteristic axes: the approach for motion modeling and the level of contextual cues awareness.

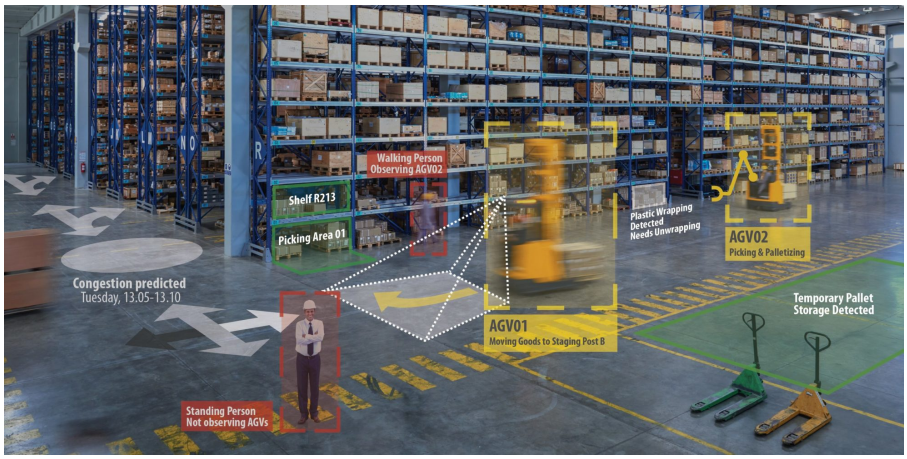


Figure 1.4: ILIAD project application scenario for motion prediction in intralogistics settings. Warehouse facilities introduce a rich contextual layout with semantically-significant areas, temporally distributed events and recurring dynamical patterns.

Our taxonomy includes three classes of motion modeling approaches: physics-based, pattern-based and planning-based. In the physics-based models, motion is represented with a set of explicitly defined dynamics equations that follow a physics-inspired model. Forward simulating these models yields prediction. Pattern-based approaches approximate an arbitrary dynamics function from training data. These approaches are able to discover statistical behavioral patterns in the observed motion trajectories. Planning-based approaches explicitly reason about the agent's long-term motion goals and compute policies or path hypotheses that enable an agent to reach those goals. All three classes can be informed with various cues from the target agent, static and dynamic environment. The level of context awareness makes up the second classification criteria in our taxonomy.

Having discussed the motion modeling approaches, we move on to a practical case of designing a highly context-aware long-term prediction method for a mobile robot.

Contribution 2: A planning-based interaction-aware approach to predict motion that combines the benefits of the two major modeling classes from our taxonomy. The proposed approach is based on Markov Decision Processes to predict global map-aware motion paths in arbitrary environments. To account for social contextual cues, our approach biases the obtained global motion policies using Group Social Forces.

As a mobile robot is expected to perform in previously unseen environments, which may be crowded and filled with obstacles, we build our approach ac-

cording to these requirements. Our method has high generalizability to new environments with obstacles, walking and interacting people. Optionally, our method takes a small amount of training data for hyperparameter tuning. It has a high level of context-awareness, supporting semantic maps and social grouping cues. Finally, the method actively reasons about the navigation goals of the observed people, producing multi-modal and uncertainty-aware predictions of the possible paths towards the goals. Without strong dependence on specific training data, our method offers reliable and certifiable performance.

Among the contextual cues, relevant for motion prediction, semantics plays an important role, especially in urban environments. Prediction there is not possible without properly understanding the different walkable surfaces, such as sidewalks, roads, crosswalks, unpaved areas and greenspaces. Not only the surface properties, but also their layout, relative to each other, should be considered when making hypotheses on walking preferences of the people. Building on this insight, the next contribution of this thesis is:

Contribution 3: A method to learn occupancy priors from semantic maps which brings an increased level of semantics-awareness to an autonomous system in urban settings. Our method uses a Convolutional Neural Network to learn preferences of walking people, taking the global context and topological connectivity of the environment into account.

The role of semantics in urban movement understanding is underappreciated, as we show in the first contribution. Semantic maps in the advent of automated driving are recognized increasingly often as a relevant cue for motion prediction, but methods to utilize such input are still few and far between. We take the next step in this development, from using semantic maps for trajectory prediction to creating an independent relation between semantics and human motion. This relation allows assessing the semantically-rich environment regardless of the specific observed and tracked pedestrians. It enables an autonomous system to *anticipate* the dynamics in a specific area. For example, a cleaning robot could infer more heavily used areas, or a service robot could better find people to assist. One particularly interesting application example for this method, explored in our experiments, is finding "illegal crosswalks" – such places, where the possibility of crossing the road is high due to the topology of the environment.

Benchmarking and evaluation are fundamental to any method research. The data, experiments and metrics used for evaluation have the potential to provide insight on the methods' benefits and drawbacks under a variety of conditions, expose the structural limitations, and guide the directions of future research. The following two contributions are related to benchmarking the motion prediction methods.

One question, critical to all prediction methods, is the one of data. Data is useful for training, calibration and, in some cases, for pattern extraction.

There are many factors to consider in a dataset: what is the recording location, how accurate are the annotations, what are the available cues, how diverse are the interactions between people. Existing datasets are often limited in terms of information content, annotation quality or variability of human behavior. With this motivation, the next contribution of this thesis is:

Contribution 4: A procedure design to collect diverse and accurate motion data in the interactive, weakly scripted setup in a controlled environment, which includes obstacles, goals, dynamically allocated tasks and various social roles of the participating people. We instantiate this procedure in a new dataset of motion trajectories, called THÖR.

THÖR contains over 60 minutes of human motion in 395k frames, recorded at 100 Hz, 2531k people detections and over 600 individual and group trajectories between multiple resting points. In addition to the video stream from one of the eye tracking headsets, the data includes 3D LiDAR scans and a video recording from stationary sensors. On top of the recording design, we propose a set of quantitative metrics to analyze the trajectory datasets, such as tracking duration, perception noise, curvature and speed variation of the trajectories.

Proper development of prediction methodology is not possible without benchmarking. With a multitude of new methods proposed by different communities, the lack of standardized benchmarking and objective comparison between them has been a major limitation for assessing the capabilities of the state-of-the-art systems. The few existing benchmarks do not cover the full spectrum of important experiments and do not include necessary contextual cues, excluding a large portion of prediction models from evaluation. To advance the state of benchmarking, this thesis presents:

Contribution 5: A benchmark design for motion prediction methods built for thorough evaluation and comparison in automated repeatable experiments with a systematic variation of the several key prediction parameters. The benchmark offers tools, such as metrics, data preparation and filtering, calibration and visualization, and includes a large variety of heterogeneous datasets, representing usual human motion behaviors in different places and cultures.

The benchmark currently includes three experiments. In the accuracy experiment, the metric values are conditioned on the prediction horizon and observation length, allowing to gain insight into the effective range of operation and sensitivity to the input length. In the transfer experiment we study the performance decrease which occurs when a method is applied outside the training/validation dataset. Finally, the robustness experiment tests the method performance in presence of perception noise. Using this benchmark, we evaluate several local interaction models with theoretically better performance than the one used in our MDP-based predictor.

1.5 Publications

Most of the work presented in this thesis has been published in peer-reviewed conferences and journals.

- A. Rudenko, L. Palmieri, and K. O. Arras. Predictive planning for a mobile robot in human environments. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA), Workshop on AI Planning and Robotics*, 2017

  Part of Chapter 3

- A. Rudenko, L. Palmieri, and K. O. Arras. Joint prediction of human motion using a planning-based social force approach. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 1–7, 2018

  Part of Chapter 3

* A. Rudenko, L. Palmieri, A. J. Lilienthal, and K. O. Arras. Human motion prediction under social grouping constraints. In *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*, 2018

  Main part of Chapter 3

- A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras. Human motion trajectory prediction: A survey. *Int. J. of Robotics Research*, 39(8):895–935, 2020

  Main part of Chapter 2

- A. Rudenko, T. P. Kucner, C. S. Swaminathan, R. T. Chadalavada, K. O. Arras, and A. J. Lilienthal. THÖR: Human-robot navigation data collection and accurate motion trajectories dataset. *IEEE Robotics and Automation Letters*, 5(2):676–682, 2020

  Main part of Chapter 5

- A. Rudenko, T. Kucner, C. Swaminathan, R. Chadalavada, K. O. Arras, and A. J. Lilienthal. Benchmarking human motion prediction methods. In *Proc. of the ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI), Workshop on Test Methods and Metrics for Effective HRI in Real World Human-Robot Teams*, 2020

  Part of Chapter 6

* Nominated for the Best Paper Award in Safety, Security and Rescue Robotics at IROS 2018

- A. Rudenko, L. Palmieri, J. Doellinger, A. J. Lilienthal, and K. O. Arras. Learning occupancy priors of human motion from semantic maps of urban environments. *IEEE Robotics and Automation Letters*, 6(2): 3248–3255, 2021

  Main part of Chapter 4

In each of these papers I have contributed to idea development, method design, software implementation, validation and results analysis. In addition, Chapter 6 includes material developed together with my M.Sc. student Wanting Huang during the work on her thesis. To my inputs here belong the design of the benchmark and the experiments, as well as the proposed methods in the comparison, while Wanting carried out the implementation and result analysis.

1.6 Dissemination

One key concept, which threads throughout this thesis, is the one of creating interdisciplinary connections. As the task of motion modeling and prediction has been the subject of research in many diverse applications, innovative solutions and insights are dispersed in time and among communities. Such area of research benefits greatly from a common discussion platform, which fosters inspiration, knowledge exchange and leads to better method development. As part of the effort to reach out to and consolidate the prediction world, I was actively engaged in multiple workshops organization and editorial work. With a careful selection of invited speakers from academic and industry background, balancing topics in human perception, prediction, human-aware planning, full-body motion and human-robot interaction, and an expert team of Program Committee members and guest reviewers, our dissemination efforts have established a new standard of communication in the motion prediction domain.

- A. Rudenko, L. Palmieri, A. Alahi, J. Mainprice, and K. O. Arras. 2nd Workshop on Long-term Human Motion Prediciton (LHMP 2020). In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2020. URL `https://motionpredictionicra2020.github.io`

- L. Palmieri, A. Rudenko, J. Mainprice, and K. O. Arras. Special Issue on Long-term Human Motion Prediction. In *IEEE Robotics and Automation Letters*, 2020

- A. Alahi, L. Ballan, P. Coscia, L. Palmieri, and A. Rudenko. Workshop on Benchmarking Trajectory Forecasting Models (BTFM 2020). In *Proc. of the Europ. Conf. on Comp. Vision (ECCV)*, 2020. URL `https://sites.google.com/view/btfm2020`

- A. Rudenko, L. Palmieri, K. O. Arras, A. Bajcsy, A. Alahi, and A. J. Lilienthal. 3rd Workshop on Long-term Human Motion Prediciton (LHMP

2021). In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2021. URL `https://motionpredictionicra2021.github.io`

1.7 Outline

The rest of this thesis is structured as follows:

Chapter 2 presents a large-scale review of the motion prediction methodology, covering three decades of research. The ambitious aim of this chapter is to introduce, bring together and organize a larger world of motion prediction methods, which is historically very fragmented: insightful ideas are spread sparsely over time, across communities and domains. In this chapter we review methods, datasets, metrics and application areas, formulate current trends and open challenges, and lay the foundations for the contributions of the following chapters.

Chapter 3 introduces the novel combination of the two major classes of approaches in one powerful method to predict human motion in environments both crowded and cluttered. We present an MDP-based predictor, which is obstacle- and uncertainty-aware, integrates reasoning on the goals of the observed people and dynamically adapts the prediction to the perceived velocity of the person. An interaction component, based on the social forces, adds active collision avoidance and group-awareness to the predicted trajectories.

Chapter 4 explores beyond the prediction horizon, introducing a method to learn occupancy patterns in semantic maps of urban environments. Our solution uses a Convolutional Neural Network to infer the probability to find a person in any state in the environment, taking the topology and semantic context into account.

Chapter 5 explores in detail the data collection for training and validation of the human-aware autonomous systems. In this regard, we present a general experiment design and methodology to collect diverse and accurate data of human motion, which is rich in relevant motion cues. Our collected dataset THÖR includes one hour of interactive human motion with accurate ground truth for position, head orientation, gaze direction, social grouping, obstacles map and goal coordinates.

Chapter 6 presents an automated benchmark for a thorough evaluation of the motion prediction methods. Using this benchmark, we revisit our choice for the local collision avoidance method used in Chapter 3, and evaluate a predictive social force method as a prospective alternative to it.

Chapter 7 concludes the thesis with a summary of the contributions, a review of the open challenges and an outlook on the ongoing work: an integrated

hierarchical motion planner with multiple levels of prediction, and further formalization of benchmarking with more powerful experiments.

1.8 Ethical Considerations

The work on this thesis has been concluded in 2020, the year when the global pandemic of the SARS-CoV-19 virus stroke our lives. A deeply traumatic experience for all people on the planet, it has shown the tremendous capabilities of science and technology to provide medical research of astonishing scale and velocity, ease the forced social isolation, retain productivity in home offices and minimize the necessary lockdown measures. Among other things, we have seen the importance of mobility tracking for cutting the chains of infections, as well as crowd density control (e.g. in shops and supermarkets) to enable safe social distancing of the customers.

Human motion understanding, modeling and prediction can help us better understand these mobility patterns, and design social and urban spaces in a way that would minimize the spread of airborne viruses. This technology is not harmful by itself, but it can be exploited for people tracking and privacy violation. It is our duty as scientists to understand and acknowledge the potential threats and provide timely warnings, educating the general public and helping the policymakers to prevent malevolent exploits.

Chapter 2
Motion Prediction Review

If you can look into the seeds of time, and say which grain will grow and which will not, speak then to me.

Macbeth
William Shakespeare

With growing numbers of intelligent autonomous systems in human environments, the ability of such systems to perceive, understand and anticipate human behavior becomes increasingly important. Specifically, predicting future positions of dynamic agents and planning considering such predictions are key tasks for self-driving vehicles, service robots and advanced surveillance systems.

This chapter provides a survey of human motion trajectory prediction. We review, analyze and structure a large selection of work from different communities and propose a taxonomy that categorizes existing methods based on the motion modeling approach and level of contextual information used. We provide an overview of the existing datasets and performance metrics. We discuss limitations of the state of the art and outline directions for further research.

2.1 Introduction

In the previous chapter we have detailed and motivated the interest for accurate prediction of future trajectories, arising in many tasks and application domains. These include self-driving vehicles, service robots, and advanced surveillance systems, as shown in Fig. 1.2, but also many others, such as AI in gaming, building design, evacuation and panic modeling, urban spaces planning. Indeed, this broad motivation has stimulated research in the robotics, automation and simulation communities already three decades ago. The recent burst of attention to the surrounding dynamics of people and vehicles from the service robots and automated vehicles has sparked the publication rates in motion prediction,

as we show in Fig. 2.2. A rapidly developing field benefits greatly from intrinsic organization and common frame of discussion, which we propose in this chapter.

The scope of the survey in this chapter is human motion trajectory prediction. Specifically, we focus on ground-level 2D trajectory prediction for pedestrians and also consider the literature on cyclists and vehicles. Prediction of video frames, articulated motion, or human actions or activities is out of scope although many of those tasks rely on the same motion modeling principles and trajectory prediction methodology considered here. Within this scope, we survey a large selection of works from different communities and propose a novel taxonomy based on the motion modeling approaches and the contextual cues. We categorize the state of the art and discuss typical properties, advantages and drawbacks of the categories as well as outline open challenges for future research. Finally, we raise three questions:

Q1: are the evaluation techniques to measure prediction performance good enough and follow best practices?

Q2: have all prediction methods arrived on the same performance level and the choice of the modeling approach does not matter anymore?

Q3: is motion prediction solved?

The chapter is structured as follows: we present the taxonomy in Sec. 2.3, review and analyze the literature on human motion prediction first by the modeling approaches in Sec. 2.4 – Sec. 2.6, and then by the contextual cues in Sec. 2.7. In Sec. 2.8 we review the benchmarking of motion prediction techniques in terms of commonly used performance metrics and datasets. In Sec. 2.9 we discuss the state of the art with respect to the above three questions and outline open research challenges. Finally, Sec. 2.10 concludes the chapter.

This chapter includes a review and categorization of over 200 methods, covering three decades of research. The methods come from a variety of communities, their assumptions and approaches to motion prediction sometimes vary greatly from something as simple as a Kalman filter to elaborate imitation learning and game-theoretic frameworks, featuring complexly designed competing agents. Clearly, a review of this sort may be overwhelming to an unprepared reader. With that in mind, this chapter is designed in a modular structure to allow for both quick overview and deep dive into the details. In particular, we recommend Sec. 2.2, 2.3, Fig. 2.6–2.8 and Sec. 2.9 as a coarse overview of the motion prediction methodology for a general reader. A practitioner may find value in the review of the datasets and metrics in Sec. 2.8. Finally, the thorough analysis of the literature in Sec. 2.4–2.7 is recommended for expert readers.

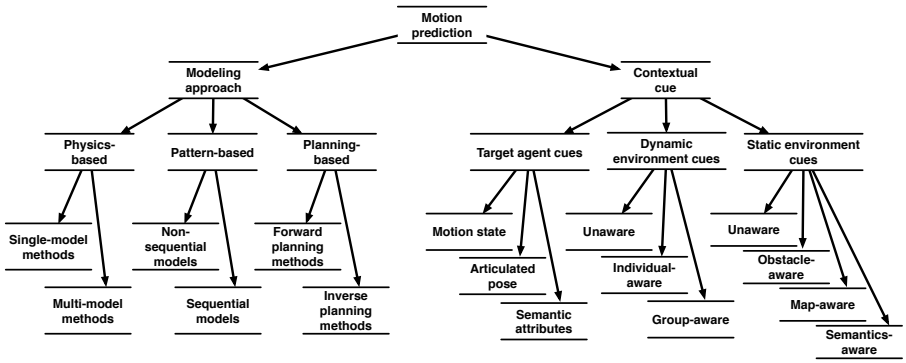


Figure 2.1: Overview of the categories in our taxonomy

2.2 Related Reviews

In this section, we detail related surveys from different scientific communities, i.e. robotics [59, 164, 174], intelligent vehicles [44, 180, 260], and computer vision [113, 215, 220].

Kruse et al. [164] provide a survey of approaches for wheeled mobile robots and categorize human-aware motion based on comfort, naturalness and sociability features. Motion prediction is seen as part of a human-aware navigation framework and categorized into *reasoning-based* and *learning-based* approaches. In reasoning-based methods, predictions are based on simple geometric reasoning or dynamic models of the target agent. Learning-based approaches make predictions via motion patterns that are learned from observed agent trajectories.

A short survey on frameworks for socially-aware robot navigation is provided by Chik et al. [59]. The authors discuss key components of such frameworks including several planners and human motion prediction techniques.

Lasota et al. [174] survey the literature on safe human-robot interaction along the four themes of safety through control, motion planning, prediction and psychological factors. In addition to wheeled robots, they also include related works on manipulator arms, drones or self-driving vehicles. The literature on human motion prediction is divided into methods based on *goal intent* or *motion characteristics*. Goal intent techniques infer an agent's goal and predict a trajectory that the agent is likely to take to reach that goal. The latter group of approaches does not rely explicitly on goals and makes use of observations about how humans move and plan natural paths.

Lefèvre et al. [180] survey vehicular motion prediction and risk assessment in an automated driving context. The authors discuss the literature based on the semantics used to define motion and risk and distinguish *physics-based*,

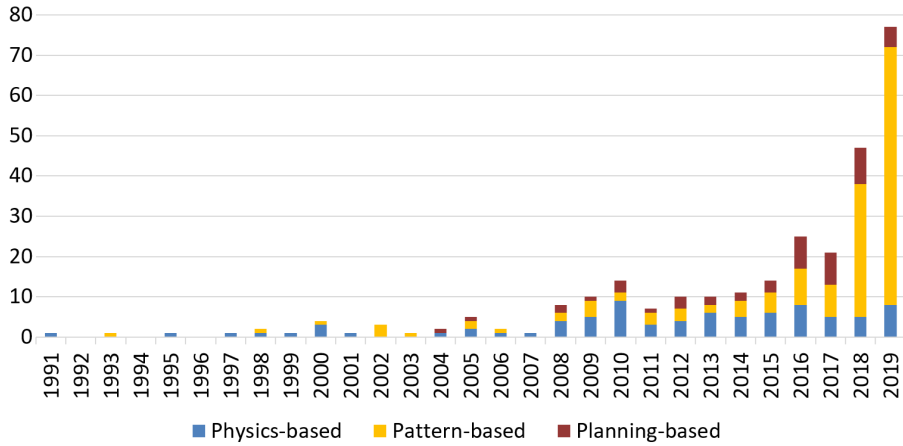


Figure 2.2: Publications trends in the literature reviewed in this chapter, color-coded by modeling approach.

maneuver-based and *interaction-aware* models for prediction. Physics-based methods predict future trajectories via forward simulation of a vehicle model, typically under kinodynamic constraints and uncertainties in initial states and controls. Maneuver-based methods assume that vehicle motion is a series of typical motion patterns (maneuvers) that have been acquired a priori and can be recognized from observed partial agent trajectories. Intention-aware methods make joint predictions that account for inter-vehicle interactions, also considering that such interactions are regulated by traffic rules.

Brouwer et al. [44] review and compare pedestrian motion models for vehicle safety systems. According to the cues from the environment used as input for motion prediction, authors distinguish four classes of methods: *dynamics-based models* which only use the target agent's motion state, methods which use *psychological knowledge of human behavior* in urban environments (e.g. probabilities of acceleration, deceleration, switch of the dynamical model), methods which use *head orientation* and *semantic map* of the environment. This categorization is extended by Ridel et al. [260] to review pedestrian crossing intention inference techniques.

Morris and Trivedi [215] survey methods for trajectory learning and analysis for visual surveillance. They discuss similarity metrics, techniques and models for learning prototypical motion patterns (called activity paths) and briefly consider trajectory prediction as a case of online activity analysis. Murino et al. [220] discuss group and crowd motion analysis as a multidisciplinary problem that combines insights from the social sciences with concepts from computer vision and pattern recognition. The authors review several recent methods for tracking and prediction of human motion in crowds. Hirakawa et al. [113] sur-

vey video-based methods for semantic feature extraction and human trajectory prediction. The literature is divided based on the motion modeling approach into *Bayesian models*, *energy minimization methods*, *deep learning methods*, *inverse reinforcement learning methods* and *other* approaches.

Related to our discussion of the benchmarking practices, several works survey the datasets of motion trajectories [113, 244, 260] and metrics for prediction evaluation [249]. Poiesi and Cavallaro [244] and Hirakawa et al. [113] describe several datasets of human trajectories in crowded scenarios, used to study social interactions and evaluate path prediction algorithms. Ridel et al. [260] discuss available datasets of pedestrian motion in urban settings. Quehl et al. [249] review several trajectory similarity metrics, applicable in the motion prediction context.

Unlike these surveys, this chapter reviews and analyze the literature across multiple application domains and agent types. The presented taxonomy offers a novel way to structure the growing body of literature, containing the categories proposed by Kruse et al. [164], Lasota et al. [174] and Lefèvre et al. [180] and extending them with a systematic categorization of contextual cues. In particular, we argue that the modeling approach and the contextual cues are two fundamentally different aspects underlying the motion prediction problem and should be considered separate dimensions for the categorization of methods. This allows, for example, the distinction of physics-based methods that are unaware of any external stimuli from methods in the same category that are highly situational aware accounting for road geometry, semantics and the presence of other agents. This is unlike previous surveys whose categorizations are along a single dimension based on both different modeling approaches and increasing levels of contextual awareness.

This chapter extends the existing reviews of the benchmarking and evaluation efforts for motion prediction [113, 244, 249, 260] with additional datasets, probabilistic and robustness metrics, and a principled analysis of existing benchmarking practices. Furthermore, we give an up-to-date discussion of the current state of the art and conclude with recommendations for promising directions of future research.

2.3 Taxonomy

In this section we describe our taxonomy to decompose the motion prediction problem based on the modeling approach and the type of contextual cues, see Fig. 2.1 for an overview. In Sec. 2.3.1 and 2.3.2 we detail the categories and give representative papers as examples of each category, and in Sec. 2.3.3 we describe the rules for classifying the methods.

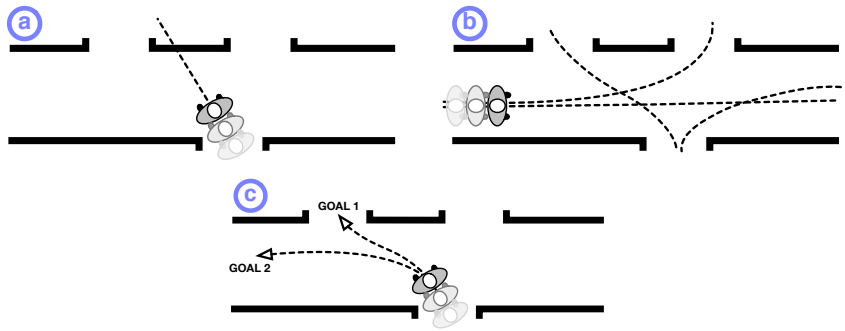


Figure 2.3: Illustration of the basic working principle of the modeling approaches: (**a**) physics-based methods project the motion state of the agent using explicit dynamical models based on Newton's law of motion. (**b**) pattern-based methods learn prototypical trajectories from observed agent behavior to predict future motion. (**c**) planning-based methods include some form of reasoning about the likely goals and compute possible paths to reach those goals. In order to incorporate internal and external stimuli that influence motion behavior, approaches can be extended to account for different contextual cues.

2.3.1 Modeling Approach

The motion modeling category subdivides the prediction approaches based on how they represent human motion and formulate the causes thereof. *Physics-based methods* define an explicit dynamical model based on Newton's law of motion. *Pattern-based methods* learn motion patterns from data of observed agent trajectories. *Planning-based methods* reason on motion intent of rational agents (see Fig. 2.3). The categorization can be seen to differ also in the level of cognition typically involved in the prediction process: physics-based methods follow a reactive sense-predict scheme, pattern-based methods follow a sense-learn-predict scheme, and planning-based methods follow a sense-reason-predict scheme in which agents reason about intentions and possible ways to the goal.

1. **Physics-based methods** (Sense – Predict): motion is predicted by forward simulating a set of explicitly defined dynamics equations that follow a physics-inspired model. Based on the complexity of the model, we recognize the following subclasses:

    1.1. **Single-model methods** define a single dynamical motion model, e.g. [15, 66, 82, 198, 236, 239, 347, 358]

    1.2. **Multi-model methods** include a fixed or on-line adaptive set of multiple dynamics models and a mechanism to fuse or select the individual models, e.g. [3, 8, 98, 135, 157, 246]

2. **Pattern-based methods** (Sense – Learn – Predict) approximate an arbitrary dynamics function from training data. These approaches are able to discover statistical behavioral patterns in the observed motion trajectories and are separated into two categories:

    2.1. **Sequential methods** learn conditional models over time and recursively apply learned transition functions for inference, e.g. [4, 14, 99, 143, 163, 166, 190, 327]

    2.2. **Non-sequential methods** directly model the distribution over full trajectories without temporal factorization of the dynamics, e.g. [30, 136, 143, 200, 309, 313, 339]

3. **Planning-based methods** (Sense – Reason – Predict) explicitly reason about the agent's long-term motion goals and compute policies or path hypotheses that enable an agent to reach those goals. We classify the planning-based approaches into two categories:

    3.1. **Forward planning methods** make an explicit assumption regarding the optimality criteria of an agent's motion, using a pre-defined reward function, e.g. [35, 46, 95, 142, 264, 266, 324, 340, 353]

    3.2. **Inverse planning methods** estimate the reward function or action model from observed trajectories using statistical learning techniques, e.g. [63, 119, 153, 167, 179, 241, 255, 292, 331, 370]

Figure 2.2 shows the publications trends over the last years, color-coded by modeling approach. The number of related works is strongly increasing during the last two years in particular for the pattern-based methods.

2.3.2 Contextual Cues

We define contextual cues to be all relevant internal and external stimuli that influence motion behavior and categorize them based on their relation to the target agent, other agents in the scene and properties of the static environment, see Fig. 2.4 and Fig. 2.5.

1. Cues of the **target agent** include

    1.1. **Motion state** (position and possibly velocity), e.g. [30, 31, 80, 91, 142, 153, 157, 166, 167, 236, 313, 370]

    1.2. **Articulated pose** such as head orientation [109, 156, 157, 265, 317] or full-body pose [212, 250]

    1.3. **Semantic attributes** such as the age and gender [202], personality [32], and awareness of the robot's presence [157, 226]

2. With respect to the **dynamic environment** we distinguish

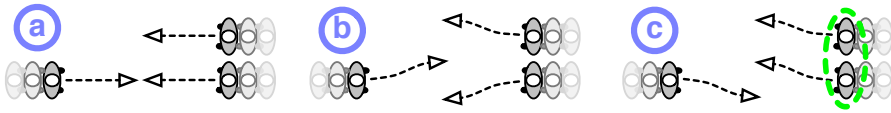


Figure 2.4: Dynamic environment cues: (**a**) unaware, (**b**) individual-aware, (**c**) group-aware (accounting for social grouping cues, in green).

2.1. **Unaware methods**, which compute motion predictions for the target agent not considering the presence of other agents, e.g. [30, 30, 82, 83, 149, 149, 165, 165, 311, 311, 334, 334, 368]

2.2. **Individual-aware methods**, which account for the presence of other agents, e.g. [4, 80, 91, 157, 167, 198, 313, 327]

2.3. **Group-aware methods**, which account for the presence of other agents as well as social grouping information. This allows to consider agents in groups, formations or convoys that move differently than independent agents, e.g. [140, 237, 248, 262, 288, 295, 347]

3. With respect to the **static environment** we distinguish

3.1. **Unaware methods**, which assume an open-space environment, e.g. [29, 81, 88, 94, 129, 163, 200, 283, 317, 325]

3.2. **Obstacle-aware methods**, which account for the presence of individual static obstacles, e.g. [4, 9, 31, 80, 91, 254, 313, 327]

3.3. **Map-aware methods**, which account for environment geometry and topology, e.g. [56, 62, 63, 100, 111, 124, 157, 190, 241, 246, 264, 266, 268, 324, 352, 370]

3.4. **Semantics-aware methods**, which additionally account for environment semantics or affordances such as no-go-zones, crosswalks, sidewalks, or traffic lights, e.g. [20, 66, 142, 153, 170, 179, 202, 255, 363]

In the following Sections 2.4, 2.5 and 2.6 we survey the different classes of the motion model category. We detail contextual cues categories in Section 2.7. In each section we review methods in the order of increasing complexity, considering inheritance of ideas and grouped by the similarity of the motion modeling techniques. A discussion of the strengths and limitations of the modeling approaches follows in Sec. 2.9.

2.3.3 Classification Rules

Some of the surveyed papers may not fall univocally into a single class of our taxonomy, especially those using a mixture of different approaches, e.g. the

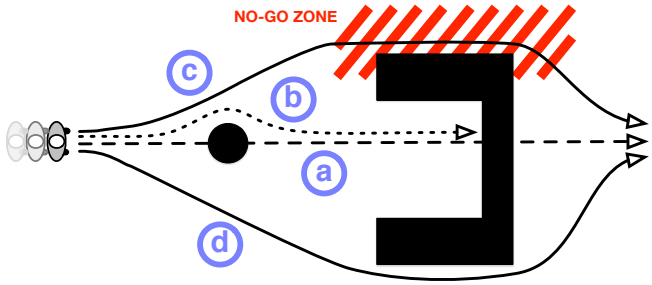


Figure 2.5: Static environment cues: (**a**) unaware (ignoring any static objects, dashed line), (**b**) obstacle-aware (accounting for unmodeled obstacles, dotted line), (**c**) map-aware (accounting for a topometric environment model avoiding local minima, solid line), (**d**) semantics-aware (solid line).

work by Bennewitz et al. [30] which combines a non-sequential clustering approach with sequential HMM inference. For those borderline cases, we adopt the following rules:

i) We classify methods primarily in the category that best describes the modelling approach over the inference method, e.g. for [30] we give more weight to the clustering technique used for modelling the prototypical human motion behavior.

ii) Some approaches add sub-components from other categories in their main modeling approach, e.g. planning-based approaches using physics-based transition functions [267, 319], physics-based methods tuned with learned parameters [91], planning-based approaches using inverse reinforcement learning to recover the hidden reward function of human behaviors [153, 370]. We classify such approaches based on their main modeling method.

iii) Methods that use behavior cloning (imitation of human behaviors with supervised learning techniques), i.e. learn/recover the motion model directly from data, are classified as pattern-based approaches [281, 363]. In contrast to that, imitation learning techniques that reason on policies (e.g. using generative adversarial imitation learning [188]) are classified as planning-based methods.

Furthermore, a single work is categorized into all three classes of contextual awareness with respect to its perception of the target agent, static and dynamic contextual cues.

2.4 Physics-based Approaches

Physics-based models generate future human motion considering a hand-crafted, explicit dynamical model f based on Newton's laws of motion. A common form for f is $\dot{\mathbf{s}}_t = f(\mathbf{s}_t, \mathbf{u}_t, t) + w_t$ where $\mathbf{u}_t$ is the (unknown) control input and w_t the process noise. In fact, motion prediction can be seen as inferring $\mathbf{s}_t$ and $\mathbf{u}_t$ from various estimated or observed cues.

A large variety of physics-based models have been developed in the target tracking and automatic control communities to describe motion of dynamic objects in ground, marine, airborne or space applications, typically used as building blocks of a recursive Bayesian filter or multiple-model algorithm. These models differ in the type of motion they describe such as maneuvering or non-maneuvering motion in 2D or 3D, and in the complexity of the target's kinematic or dynamic model and the complexity of the noise model. See [185, 187] for a survey on physics-based motion models for target tracking.

We subdivide physics-based models into (1) *single-model approaches* that rely on a single dynamical model f and (2) *multi-model approaches* that involve several modes of dynamics (see Fig. 2.6).

2.4.1 Single-model Approaches

Early works and basic models

Many approaches to human motion prediction represent the motion state of target agents as position, velocity and acceleration and use different physics-based models for prediction. Among the simplest ones are kinematic models without considering forces that govern the motion. Popular examples include the constant velocity model (CV) that assumes piecewise constant velocity with white noise acceleration, the constant acceleration model (CA) that assumes piecewise constant acceleration with white noise jerk, the coordinated turn model (CT) that assumes constant turn rate and speed with white noise linear and white noise turn acceleration or the more general curvilinear motion model by Best and Norton [36]. The bicycle model is an often used as an approximation to model the vehicle dynamics (see e.g. [286]).

A large number of works across all application domains rely on kinematic models for their simplicity and acceptable performance under mild conditions such as tracking with little motion uncertainty and short prediction horizons. Examples include [213] for hazard inference from linear motion predictions of pedestrians or [82] for Kalman filter-based (KF) prediction of dynamic obstacles using a constant acceleration model. Barth and Franke [24] use the coordinated turn model for one-step ahead prediction in an Extended Kalman Filter (EKF) to track oncoming vehicles from point clouds generated by an in-car stereo camera. Batz et al. [27] use a variant of the coordinated turn model for one-

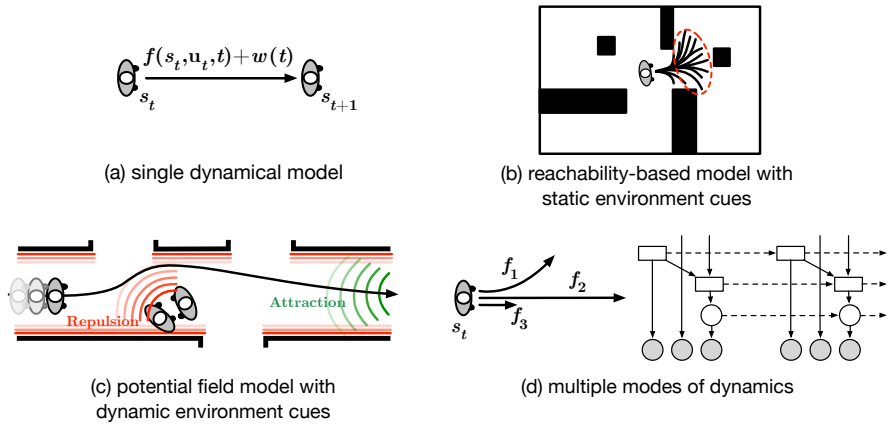


Figure 2.6: Examples of the physics-based approaches: (**a**) a method with a single dynamical model, (**b**) a reachability-based method, which accounts for all possible transitions from the given motion state, (**c**) an attraction-repulsion approach, which accounts for dynamic environment cues, (**d**) a multi-model method with several modes of dynamics and the DBN switching mechanism.

step motion prediction of vehicles within an Unscented KF to detect dangerous situations based on predicted mutual distances between vehicles.

Dynamic models account for forces which, following Newton's laws, are the key descriptor of motion. Such models can become complex when they describe the physics of wheels, gearboxes, engines, or friction effects. In addition to their complexity, forces that govern the motion of other agents are not directly observable from sensory data. This makes dynamic models more challenging for motion prediction. Zernetsch et al. [358] use a dynamic model for trajectory prediction of cyclists that contains the driving force and the resistance forces from acceleration, inclination, rolling and air. The authors show experimentally that long-term predictions up to 2.5 sec ahead are geometrically more accurate when compared to a standard CV model.

Autoregressive models (ARM) that, unlike first-order Markov models, account for the history of states have also been used for motion prediction. Elnagar and Gupta [83] employ a third-order ARM to predict the next position and orientation of moving obstacles using maximum-likelihood estimation of the ARM parameters. Cai et al. [50] use a second-order ARM for single step motion prediction within a particle filter for visual target tracking of hockey players. The early work by Zhu [368] uses an autoregressive moving average model as transition function of a Hidden Markov Model (HMM) to predict occupancy probabilities of moving obstacles over multiple time steps with applications to predictive planning.

Physics-based models are used for motion prediction by recursively applying the dynamics model f to the current state of the target agent. So far, with the exception of [368], the works described above make only one-step ahead predictions and ignore contextual cues from the environment. To account for context, the dynamics model f can be extended by additional forces, model parameters or state constraints as discussed hereafter.

Models with map-based contextual cues

A number of approaches extend physics-based models to account for information from a map, particularly for the task of tracking ground vehicles on roads. The methods developed to this end differ in how road constraints are derived and incorporated into the state estimation problem, see the survey by Simon [294]. Yang and Blasch [350], for example, use a regular KF and project the unconstrainted state estimate onto the constrained surface for tracking on-road ground vehicles with a surveillance radar. Yang et al. [351] use the technique to reduce the system model parametrization to the constrained surface. They reduce vehicle motion to a 1D curvilinear road representation for filtering. Batkovic et al. [26] predict pedestrian motion along a graph with straight line edges centered on side- and crosswalks. Using a unicycle model and a control approach to keep the predictions along the edges, they evaluate long-term predictions up to 10 sec ahead. When there are several possible turns at a node, i.e. at bifurcations, predictions are propagated along all outgoing edges. Another class of techniques uses the road information as pseudo measurements, pursued e.g. by Petrich et al. [239] who use a kinematic bicycle model for f and pseudo measurements from the centerlines of lanes to predict future vehicle trajectories several seconds ahead. When there are several possible turns, e.g. at intersections, the approach generates new motion hypothesis for each relevant lane by using an EKF.

When agents move freely, e.g. do not comply with road constraints, we need different ways to represent free space and account for map information. To this end, several authors propose grid-based [66, 199, 254] and more general graph-based space discretizations [15, 159]. Luber et al. [199] use 2D laser data to track people from a mobile robot and learn a so called spatial affordance map, a grid-based spatial Poisson process from which a walkable area map of the environment can be derived. They predict future trajectories of people during lengthy occlusion events using an auxiliary PF with look-ahead particles obtained by forward-simulation of the curvilinear motion model proposed by Best and Norton [36]. This way, long-term predictions (up to 50 steps ahead) stay focused on high-probability regions with the result of improved tracking performance. Rehder and Klöden [254] also choose a regular grid to represent the belief about pedestrian locations in a linear road scenario. They propose a variant of a Bayesian histogram filter to achieve map-aware predictions 3 seconds

ahead by combining forward propagation of an unicycle pedestrian model from the start and in backward direction from the goal with prior place-dependent knowledge of motion learned from previously observed trajectories. Similarly, Coscia et al. [66] use polars grids, centered at the currently predicted agent position to represent four different local influences: a CV motion model, prior motion knowledge learned from data, semantic map annotations like "road" or "grass" and direction to goal. The next velocity is then obtained from the normalized product of the four polar distributions and forward propagated for long-term prediction of pedestrians and cyclists in urban scenarios. Like [254], no planning is involved and the learned prior knowledge is place-dependent. Koschi et al. [159] exploit information on road segments connectivity and semantic regions to compute reachability-based predictions of pedestrians, similarly to [254]. The authors formalize several relevant traffic rules, e.g. pedestrian crossing permission on the green light, as additional motion constraints. Aoude et al. [15] grow a tree of future trajectories for each target agent using a closed-loop RRT algorithm that samples the controls of a bicycle motion model [172] avoiding obstacles in the map. Based on agent's recognized intentions using an SVM classifier and features from observed trajectories, they bias the tree growth towards areas that are more likely for the agent to enter and determine the best evasive maneuver for the ego-vehicle to minimize threat at intersection scenarios. A reachibility-based model, such as [15, 159, 254], is illustrated in Fig. 2.6 (b).

So far, we discussed extensions to physics-based motion models that embed different types of map information. All those works, however, consider only a single target agent and neglect local interactions between multiple agents. Hereafter, we will discuss methods that add social situation awareness, predicting several target agents jointly.

Models with dynamic environment cues

There are several ways to incorporate local agent interaction models into physics-based approaches for prediction, one popular example being the social force (SF) model by Helbing and Molnar [110], see Fig. 2.6 (c). Developed for the purpose of crowd analysis and egress research, the model superimposes attractive forces from a goal with repulsive forces from other agents and obstacles. Several works extend the dynamics model f to include social forces e.g. for improved short-term prediction for pedestrian tracking in 2D laser data [198] or image data [236].

Elfring et al. [80] combine the HMM-based goal estimation method introduced by Vasquez et al. [325] with the basic SF-based human motion prediction by Luber et al. [198]. For intention estimation, the observed people trajectories are summarized in a sparse topological map of the environment. Each node of the map encodes a state–destination pair, and the goal inference using the

observed trajectory is carried out in a maximum-likelihood manner. Ferrer and Sanfeliu [91] estimate the interaction parameters of the SF for each two people in the scene individually. For this purpose several *behaviors* (i.e. sets of SF parameters) are learned offline, and the observed interaction between any two people is associated to the closest "behavior". The approach by Oli et al. [226] defines the robot operating in social spaces as an interacting agent, affected by the social forces. Each human is flagged as either aware or unaware of the robot, which defines the repulsive force the robot exerts on that person. Such awareness is inferred using visual cues (gaze direction and past trajectory).

In order to achieve more realistic behaviors, several extensions to the social force model are proposed. Yan et al. [348] present a model that embeds social relationships in the linear combination of predefined basic social effects (attraction, repulsion and non-interaction). The motion predictor maintains several hypothesis over the social modes, in which the pedestrians are involved. Predictive collision avoidance behavior of the SF agents is introduced by Karamouzas et al. [141] and Zanlungo et al. [356]. In particular, Karamouzas et al. [141] models each agent to adapt their route as early as possible, trying to minimize the amount of interactions with others and the energy required to solve these interactions. To this end an evasion force, that depends on the predicted point of collision and the distance to it, is applied to each agent. Updates to the SF model to consider also group motion are proposed by Moussaïd et al. [216] and Farina et al. [86].

Other agent interaction models, not based on the social forces, for example for road vehicles, have also been used. An interactive kinematic motion model for vehicles on a single lane has been proposed by Treiber et al. [315] to predict the longitudinal motion of a target vehicle in the presence of preceding vehicles. The model, called Intelligent Driver Model (IDM), was used e.g. by Liebner et al. [191] for driver intent inference at urban intersections. Hoermann et al. [115] learn the driving style of preceding vehicles by on-line estimating the IDM parameters using particle filtering and near- and far-range radar observations. Prediction of longitudinal motion of preceding vehicles, in the experiments up to 10 seconds ahead, is then obtained by forward propagation of the model.

Several approaches exploit the *reciprocal velocity obstacles* (RVO) model [318] for jointly predicting human motions. Kim et al. [150] use the Ensemble Kalman filtering technique together with the Expectation-Maximization algorithm to estimate and improve the human motion model (i.e. RVO parameters). Bera et al. [31] propose a method that dynamically estimates parameters of the RVO function for each pedestrian, moving in a crowd, namely current and preferred velocities per agent and global motion characteristics such as entry points and movement features. A follow-up work [32] also introduces online estimation of personality traits. Each pedestrian's behavior is characterized as a weighted combination of six personality traits (aggressive, assertive, shy, active,

tense and impulsive) based on the observations, thus defining parameters of the RVO model for this person.

Other approaches instead compute joint motion predictions based on the time of possible collision between pairs of agents. Paris et al. [233] propose a method for modeling predictive collision avoidance behavior in simulated scenarios. For each pedestrian current velocities of their neighbors are extrapolated in the 3D (x, y, t) space, and all actions that result in collision with dynamic and static obstacles are excluded. A similar problem is addressed by Pettré et al. [240], who evaluate real people trajectories in an interactive experiment and design a predictive collision avoidance approach, capable of reproducing realistic joint maneuvers, such as giving way and passing first.

Other methods propose to compute joint motion prediction based on the expected point of closest approach between pedestrians. Pellegrini et al. [236] is the first to propose such approach called *Linear Trajectory Avoidance* (LTA): the method firstly computes the expected point of closest approach between different agents, and then uses it as driving force to perform avoidance between the agents. Based on the LTA, Yamaguchi et al. [347] formulate a human motion prediction approach as an energy minimization problem. The energy function considers different properties of people motion: damping, speed, direction, attraction, being in a group, avoiding collisions. The approach of Yamaguchi is further improved by Robicquet et al. [262] by considering several different sets of the energy functional parameters, learned from the training data. Each set of parameters represents a distinct behavior (navigation style of the agent).

Local interaction modeling methods, as well as approaches for predicting motion in crowds, usually benefit from detecting and considering groups of people who walk together. For example, Pellegrini et al. [237] propose an approach to model joint trajectories of people, taking group relations into account. The proposed framework operates in two steps: first, it generates possible trajectory hypotheses for each person, then it selects the best hypothesis that maximize a likelihood function, taking into account social factors, while at the same time estimating group membership. People and relations are modeled with Conditional Random Fields (CRF). Choi and Savarese [61] propose an interaction model that incorporates linear motion assumption, repulsion of nearby people and group coherence via synchronization of velocities. Further group motion models, e.g. [140, 248, 288, 295], developed in the simulation and visualization communities, typically address the groups cohesion with additional forces to attract members to each other, assigning leader's and follower's roles or imposing certain group formation.

A recent reachability-based pedestrian occupancy prediction method, presented by Zechel et al. [357], accounts both for dynamic objects and semantics of the static environment. The authors first use a physical model to determine reachable locations of a person, and then reduce the area based on the intersections with static environment and presence probabilities of other dynamic

agents. Similarly Luo and Cai [201] compute future agents predictions based on an optimization approach that handles physical constraints, i.e. kinematics and geometry of the agents, and behavioral constraints, i.e. intention, attention and responsibility.

2.4.2 Multi-model Approaches

Complex agent motion is poorly described by a single dynamical model f. Although the incorporation of map information and influences from multiple agents render such approaches more flexible, they remain inherently limited. A common approach to modeling general motion of maneuvering targets is the definition and fusion of different prototypical motion modes, each described by a different dynamic regime f. Modes may be linear movements, turn maneuvers, or sudden accelerations, that over time, form sequences able to describe complex motion behavior. Since the motion modes of other agents are not directly observable, we need techniques to represent and reason about motion mode uncertainty. The primary approach to this end are multi-model (MM) methods [186] and hybrid estimation [116]. MM methods maintain a hybrid system state $\xi = (\mathbf{x}, s)$ that augments the continuous valued $\mathbf{x}$ by a discrete-valued modal state s. Following [186], MM methods generally consist of four elements: a fixed or on-line adaptive model set, a strategy to deal with the discrete-valued uncertainties (e.g. model sequences under a Markov or semi-Markov assumption), a recursive estimation scheme to deal with the continuous valued components conditioned on the model, and a mechanism to generate the overall best estimate from a fusion or selection of the individual filters. For prediction, MM methods are used in several ways, to represent more complex motion, to incorporate context information from other agents and context information from the map. A naive MM approach, presented by Pool et al. [246], predicts future motion of cyclists using a uniform mixture of five Linear Dynamic Systems (LDS) dynamics-based motion strategies: go on straight, turn 45° or 90° left or right. Probability of each strategy is set to zero if the predicted path does not comply with the road topology in the place of prediction.

The interactive multiple model filter (IMM) is a widely used inference technique applied on MM models with numerous applications in tracking [209] and predictions. For instance, Kaempchen et al. [135] propose a method for future vehicle states estimation that switches between constant acceleration and simplified bicycle dynamical models. Uncertainty in the next transition is explicitly modeled with Gaussian noise. Schneider and Gavrila [283] introduce an IMM for pedestrian trajectory prediction which combines several basic motion models (constant velocity, constant acceleration and constant turn). Also Schulz and Stiefelhagen [287] propose a method for predicting the future path of a pedestrian using an IMM framework with constant velocity, constant position and

coordinated turn models. In this work, model transitions are controlled by an intention recognition system based on Latent-dynamic Conditional Random Fields: based on the features of the person's dynamics (position and velocity) and situational awareness (head orientation), intention is classified as crossing, stopping or going in the same direction. Joint vehicle trajectory estimation also using IMMs is considered by Kuhnt et al. [169, 170] in a method which adopts pre-defined environment geometry to estimate possible routes of each individual vehicle. Contextual interaction constraints are embedded in a Bayesian Network that estimates the evolution of the traffic situation.

Other examples of IMMs techniques are variable-structure IMM for ground vehicles [152, 223, 232, 291] to account for road constraints. In a recent work Xie et al. [341] combined a kinematics-based constant turn rate and acceleration model with IMM-based lane keeping and changing maneuvers mixing. The method is aware of road geometry and produces results for a varying prediction horizon.

An alternative approach to hybrid estimation problems are dynamic Bayesian networks (DBN) which inherit the broad variety of modeling schemes and large corpus of exact and approximate inference and learning techniques from probabilistic graphical models [155]. An example of a DBN-based multi-model approach is given in Fig. 2.6 (d). The seminal work of Pentland and Liu [238] introduces an approach to model human behaviors by coupling a set of dynamic systems (i.e. a bank of Kalman filters (KF)) with an HMM, which is a special case of the DBNs. The authors introduce a dynamic Markov system that infers human future behaviors, a set of macro-actions described by a set of KFs, based on measured dynamic quantities (i.e. acceleration, torque). The approach was used to accurately categorize human driving actions. Agamennoni et al. [3] jointly model the agent dynamics and situational context using a DBN. The vehicular dynamics is described by a bicycle model whereas the context is defined by a weighted feature function to account e.g. for closeness between agents or place-dependent information from a map. The model resembles a switched Bayesian filter but considers a more general conditioning of the switch transitions and the case of multiple agents. The authors apply the model for the task of long-term multi-vehicle trajectory prediction of mining vehicles, useful for instance during GPS outages. Kooij et al. [156] propose a context-aware path prediction method for pedestrians intending to laterally cross a street, that makes use of Switching Linear Dynamical Systems (SLDS) to model maneuvering pedestrians that alternate between motion models (e.g. walking straight, stopping). The approach adopts a Dynamic Bayesian Network (DBN) to infer the next pedestrian movements based on the SLDS model. The latent (context) variables relate to pedestrian awareness of an oncoming vehicle (head orientation), the distance to the curbside and the situation criticality. Kooij et al. [157] extend this work to cover a cyclist turning scenario. In another extension of [156], Roth et al. [265] use a second context-based SLDS to model the "brak-

ing" and "driving" behaviors of the ego-vehicle. The two SLDS sub-graphs for modeling pedestrian and vehicle paths are combined into a joint DBN, where the situation criticality latent state is shared. Gu et al. [104] propose a DBN-based motion model with a particle filter inference to estimate future position, velocity and crossing intention of a pedestrian. During inference the approach considers standing, walking and running motion modes of pedestrians. Gindele et al. [98] is jointly modeling future trajectories of vehicles with a DBN, describing the local context of the interaction between multiple drivers with a set of numerical features. These features are used to classify the current situation of each driver and reason on available behaviors, such as "follow", "sheer in" or "overtake", represented as Bézier curves. Blaiotta [41] also proposes a DBN for pedestrian prediction with two motion modes (walking and standing), contextual awareness flag for the oncoming vehicle and social force-based motion dynamics for pedestrians.

Techniques derived by the stochastic reachability analysis theory [7] form another class of hybrid approaches to compute human motion prediction. In general, those methods model agents as hybrid systems (with multiple modes) and infer agents' future motions by computing stochastic reachable sets. The approach by Althoff et al. [9] generates the stochastic reachable sets for interacting traffic participants using Markov chains, where each chain approximates the behavior of a single agent. Each vehicle has its own dynamics with many modes (e.g. acceleration, deceleration, standstill, speed limit), and its goal is assumed to be known. Althoff et al. [10] further extend [9] with the over-approximative estimation of the occupancy sets. The method is particularly framed for hybrid dynamics (mixed discrete and continuous) where computing the exact reachability sets could be computationally unfeasible. To overcome this issue, the method proposes to intersect different occupancy sets for different abstractions of the dynamical model. The work by Bansal et al. [23] also uses a reachability approach for solving the prediction problem for multi-models systems. The approach rather than using a probability distribution over human next actions, it uses a deterministic set of allowable human actions. This reduces the complexity of the predictor and allows for an easy certification process.

2.5 Pattern-based Approaches

In contrast to the physics-based approaches which use explicitly defined, parametrized functions of motion dynamics, pattern-based approaches learn the latter from data, following the *Sense - Learn - Predict* paradigm. These methods learn human motion behaviors by fitting different function approximators (i.e. neural networks, hidden Markov models, Gaussian processes) to data. Many of those methods were introduced by the machine learning and computer vision

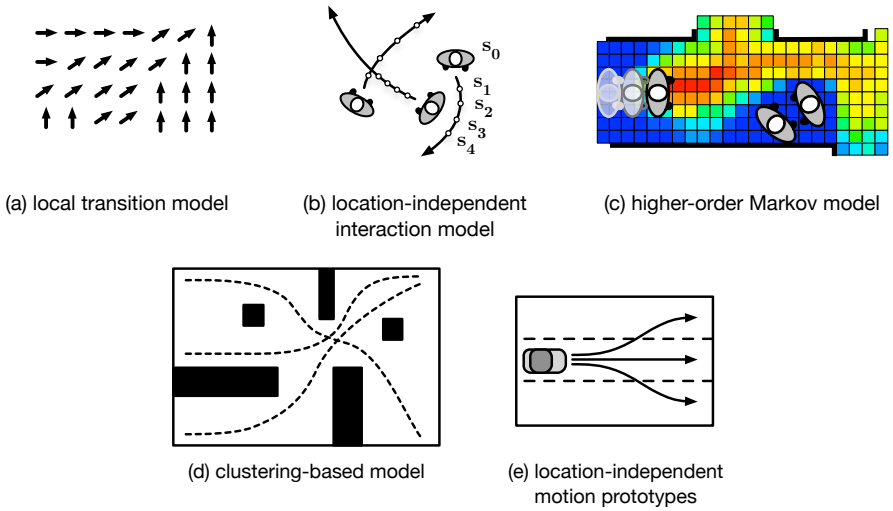


Figure 2.7: Examples of the pattern-based approaches: (**a**) grid-based local transitions learning method, (**b**) sequential location-independent transition model, which accounts for cues from the dynamic environment, (**c**) higher-order sequential Markov model, (**d**) clustering of full trajectories, (**e**) location-independent method which learns long-term transition sequences, i.e. maneuvers.

communities (i.e. for behavior cloning and video surveillance applications), and later applied in robotics and autonomous navigation settings.

In our taxonomy we classify pattern-based approaches into two categories, based on the type of function approximator used:
(1) Sequential methods typically learn conditional models, where it is assumed that the state (e.g. position, velocity) at one time instance is conditionally dependent on some sufficient statistic of the full history of past states. Many of the proposed methods are Markov models, where an N-th order Markov model assumes that a limited state history of N time steps is a sufficient representation of the entire state history. Similarly to many physics-based approaches, sequential methods aim to learn a one-step predictor $\mathbf{s}_{t+1} = f(\mathbf{s}_{t-n:t})$, where the state $\mathbf{s}_{t+1}$ is the one step prediction and the sequence of states $\mathbf{s}_{t-n:t}$ is the sufficient statistic of the history. In order to predict a sequence of state transitions (i.e. a trajectory), consecutive one-step predictions are made to compose a single long-term trajectory.
(2) Non-sequential methods directly model the distribution over full trajectories without imposing a factorization of the dynamics (i.e. Markov assumption) as with sequential models.

2.5.1 Sequential Models

Sequential models are built on the assumption that the motion of intelligent agents can be described with causally conditional models over time. Similarly to the physics-based methods, transition function of sequential models has Markovian property, i.e. information on the future motion is confined in the current state of the agent. Differently, the function, often non-parametric (e.g. Gaussian Processes, vector fields), is learned from statistical observations, and its parameters cannot be directly interpreted as for many of the physics-based methods.

Local transition patterns

Learning local motion patterns, such as probabilities of transitions between cells on a grid-map (Fig. 2.7 (a)), is a simple, commonly used technique for making sequential predictions [20, 163, 165, 214, 305, 311, 333, 334].

Early examples of local motion patterns include the works of Tadokoro et al. [305] and Kruse and Wahl [163]. Kruse and Wahl [163] build two transition models: a stochastic grid where usual motion patterns of dynamic obstacles are stored, and stochastic trajectory prediction modeled with Poisson processes. Tadokoro et al. [305] include empirical biases to account for context features of the cells in the regions where the observations are sparse, e.g. increasing the probability to move away from the wall, stop near a bookshelf or decrease walking speed at the crossing. More recently, Thompson et al. [311] expand the local motion patterns model by accounting for further transitions for several steps into the future. Their method maps the motion state of the person to a series of local patches, describing where the person might be in the future. Besides the current motion state, the learned patterns are also conditioned on the final goal or the topological sub-goal in the environment. Wang et al. [333] model local transition probabilities with an Input-Output HMM. Transition in each cell is conditioned both on the direction of cell entrance and the global starting point of the person's movement. Jacobs et al. [129] use nonlinear estimation of pedestrian dynamics with the learned vector-fields to improve the linear velocity projection model. Ballan et al. [20] propose a Dynamic Bayesian Network method to predict not-interacting human motion based on statistical properties of human behavior. To this end a transferable navigation grid-map is learned. It encodes functional properties of the environment (i.e. direction and speed of the targets, crossing frequency for each patch, identification of routing points). Molina et al. [214] address periodic temporal variations in the learned transition patterns, e.g. based on the time of the day.

In contrast to the discrete transition patterns discussed so far, several authors model the transition dynamics as a continuous function of the agent's motion state, using Gaussian Processes and their mixtures [81, 88, 134, 166]. Ellis et al. [81] model trajectory data in the observed environment by regress-

ing relative motion against current position. Predictions are generated using a sequential Monte-Carlo sampling method. Joseph et al. [134] model the multimodal mobility patterns as a mixture of Gaussian processes with a Dirichlet process prior over mixture weights. Ferguson et al. [88] further extends the work of Joseph et al. [134] by including a change-point detection and clustering algorithm which enables quick detection of changes in intent and on-line learning of motion patterns not seen in prior training data. Kucner et al. [166] model multimodal distributions with a Gaussian Mixture Model (GMM) in the joint velocity-orientation space.

Apart from the commonly used grid-cells, local transition patterns can be learned using a higher-level abstraction of the workspace, such as a graph of sub-goals or transition points [108, 124], map of connected position-velocity points Kalayeh et al. [137], Voronoi diagram [190], Instantaneous Topological Map (ITM) [326], semantic-aware ITM [323]. More flexible representation of the workspace topology is achieved this way. Combining the merits of local and global motion patterns (i.e. sequential and non-sequential models), Chen et al. [54] model trajectories in the environment with a set of overcomplete basis vectors. The method breaks down trajectories into a small number of representative partial motion patterns, where each partial pattern consists of a series of local transitions. A follow-up work by Habibi et al. [107] incorporates semantic features from the environment (relative distance to curbside and the traffic lights signals) in the learning process, improving prediction accuracy and generalization to similar environments. Han et al. [108] propose a method to explicitly learn transition points between the local patterns.

Location-independent behavioral patterns

Unlike the local transition patterns, which are learned and applied for prediction only in a particular environment, *location-independent* patterns are used for predicting transitions of an agent in the general free space [14, 93, 250, 290, 312] (see Fig. 2.7 (b)).

Several authors, e.g. Foka and Trahanias [93], Shalev-Shwartz et al. [290], use location-invariant one-step prediction as a part of collision avoidance framework using neural networks. Aoude et al. [14] extend their physics-based approach [15] by introducing location-independent GP-based motion patterns that guide the RRT-Reach to grow probabilistically weighted feasible paths of the surrounding vehicles. Tran and Firl [312] model location-independent motion patterns of vehicles by applying spatial normalization to the trajectories in the learning set. Cartesian coordinates are turned into the relative coordinate system of the road intersection, based on the topology of the lanes.

Keller and Gavrila [143] use optical flow features derived from a detected pedestrian bounding box to predict future motion. Quintero et al. [250] instead extract full-body articulated pose. In both works, body motion dynamics

for walking and stopping are learned using Gaussian Processes with Dynamic Model (GPDM) in a compact low-dimensional latent space. Mínguez et al. [212] extend [250] by considering standing and starting activities as well. A first-order HMM is used to model the transition between the activities.

Several location-independent methods learn socially-aware models of local interactions [13, 327]. Antonini et al. [13] adapt the Discrete Choice Model from econometrics studies to predict local transitions of individuals, given the intended direction, current velocity, locations of obstacles and other people nearby. Vemula et al. [327] reformulates the non-sequential joint human motion prediction approach by Trautman and Krause [313], discussed in Sec. 2.5.2, as sequential inference with Gaussian Processes. They model the local motion of each agent conditioned on relative positions of other people in the surroundings and the person's goal.

Complex long-term dependencies

Several recent sequential methods use neural networks for time series prediction, i.e. assuming higher order Markov property [4, 25, 99, 130, 281, 299, 300, 321, 328, 363], see Fig. 2.7 (c). Such time series-based models are making a natural transition between the first order Markovian methods (e.g. local transition patterns) and non-sequential techniques (e.g. clustering-based). An early method, presented by Sumpter and Bulpitt [299] learns long-term spatio-temporal motion patterns from visual input in a known environment. The simple neural network architecture, based on natural language processing networks, quantizes partial trajectories in location/shape-space: the symbol network categorizes the object shape and locations at any time, and the context network categorizes the order in which they appear. Goldhammer et al. [99] learn usual human motion patterns using an ANN with the multilayer perceptron architecture. This method was adapted to predict motion of cyclists by Zernetsch et al. [358].

Recurrent Neural Networks (RNN) for sequence learning, and Long Short-term Memory (LSTM) networks in particular, have recently become a widely popular modeling approach for predicting human [4, 25, 277, 279, 300, 321, 328], vehicle [6, 71, 147, 234] and cyclist [245] motion. Alahi et al. [4] was the first one to propose a Social-LSTM model to predict joint trajectories in continuous spaces. Each person is modeled by an individual LSTM. Since humans are influenced by nearby people, LSTMs are connected in the social pooling system, sharing information from the hidden state of the LSTMs with the neighbouring pedestrians. The work of Bartoli et al. [25] extends the Social-LSTM, explicitly modeling human-space interactions by defining a "context-aware" pooling layer, which considers the static objects in the neighborhood of a person. Varshneya and Srinivasaraghavan [321] use a Spatial Matching Network, first introduced by Huang et al. [119] (discussed in Sec. 2.6.2), that models the

spatial context of the surrounding environment, predicting the probability of the subject stepping on a particular patch. Sun et al. [300] use LSTM to learn environment- and time-specific human activity patterns in the target environment from long-term observations, i.e. covering several weeks. The state of the person is extended to include contextual information, i.e. the time of the day when the person is observed. Pfeiffer et al. [242] couple obstacle-awareness with an efficient representation of the surrounding dynamic agents using a 1D vector in polar angle space. Bisagno et al. [39] add group coherence information in the social pooling layer. Saleh et al. predict trajectories of pedestrians [279] and cyclists [278], adapting the LSTM architecture for the perspective of a moving vehicle. Numerous other implementations of the LSTM-based predictors offer various improvements, such as increased generalizability to new and crowded environments [293, 346], considering the immediate [360] or long-term [344] intention of the agents, augmenting the state of the person with the head pose [109] or adding a better pooling mechanism with relative importance of each person in the vicinity of the target agent [89, 235, 343]. Huynh and Alaghband [123] apply LSTM-based trajectory prediction in combination with local transition patterns, learned on the fly in a particular scene. Nonlinear motion, historically observed in a coarse grid cell of the environment, informs the LSTM predictor.

Several authors use LSTMs to estimate kinodynamic motion of vehicles, combining the benefits of the physics-based and the pattern-based methods [69, 252]. Raipuria et al. [252] augment the LSTM model with the road infrastructure indicators, expressed in the curvilinear coordinate system, to better predict motion in curved road segments. Deo and Trivedi [69] propose an interaction-aware multiple-LSTM model to compute stochastic maneuver-dependent predictions of a vehicle, and augment it with an LSTM-based maneuver classification and mixing mechanism.

Other approaches use RNN as models of spatio-temporal graphs for problems that require both spatial and temporal reasoning [68, 79, 120, 127, 130, 328]. Jain et al. [130] propose an approach for training sequence prediction models on arbitrary high-level spatio-temporal graphs, whose nodes and edges are represented by RNNs. The resulting graph is a feed-forward, fully differentiable, and jointly trainable RNN mixture. Vemula et al. [328] apply this method to jointly predict transitions in human crowds.

RNN abilities for prediction of time-series is also combined with different neural networks architectures [61, 184, 281, 359, 363]. Schmerling et al. [281] consider a traffic weaving scenario and propose a Conditional Variational Autoencoder (CVAE) with RNN subcomponents to model interactive human driver behaviors. The CVAE characterizes a multi-modal distribution over human actions at each time step conditioned on interaction history, as well as future robot action choices. Zheng et al. [363] describes a hierarchical policy approach that automatically reasons about both long-term and short-term

goals. The model uses recurrent convolutional neural networks to make predictions for macro-goals (intermediate goals) and micro-actions (relative motion), which are trained independently by supervised learning, combined by an attention module, and finally jointly fine-tuned. Zhan et al. [359] extend this approach using Variational RNNs. Choi et al. [60] uses spatial-temporal graphs in combination with CVAE. The spatial-temporal graphs are used to model the relational influence among predicted agents. Conditions of the CVAE are represented by estimated intentions. Also Li et al. [184] propose a hierarchical architecture where an upper level (based on variational RNN) provides predictions of discrete coordination activities between agents and a lower level generates actual geometric predictions (using a Conditional Generative Adversarial Network). The probabilistic framework called *Multiple Futures Predictor* (MFP) [307] models joint behavior of an arbitrary number of agents via a dynamic attention-based state encoder for capturing relationships between agents, a set of stochastic, discrete latent variables per agent to allow for multimodal future behavior, as well as interactive and step-wise parallel rollouts with agent-specific RNNs to model future interactions. Furthermore, there model allows to make hypothetical rollouts under assumptions of behavior for a particular agent.

Several recent works [131, 251, 259, 261, 297, 320, 345, 362] combine the benefits of sequential (e.g. RNN-based) and convolutional approaches for modeling jointly the spatial and temporal relations of the observed agents' motion. Xue et al. [345] introduce a hierarchical LSTM model, which combines inputs on three scales: trajectory of the person, social neighbourhood and features of the global scene layout, extracted with a CNN. Zhao et al. [362] propose the Multi-Agent Tensor Fusion encoding, which fuses contextual image of the environment with sequential trajectories of agents, thus retaining spatial relation between features of the environment and capturing interaction between the agents. This method is applied to both pedestrian and vehicles. Also Rhinehart et al. [259] present a prediction scheme for multi-agent that combines CNNs with a generative model based on RNNs. Moreover the approach conditions the predictions on inferred intentions of the agents. Srikanth et al. [297] propose a novel input representation for learning vehicle dynamics, which includes semantics images, depth information and other agents' positions. This input is projected into top-down view and fed into the autoregressive convolutional LSTM model to learn temporal dynamics. LSTMs have been also used to predict sequence of future human movements based on a learned reward map Saleh et al. [280].

Recently, many authors have applied the GAN architecture to achieve multimodality in the prediction output [11, 105, 158]. For instance, Gupta et al. [105] extend the Social-LSTM by using Generative Adversarial Networks and a novel variety loss which encourages the generative network to produce diverse multi-modal predictions. Kosaraju et al. [158] use Graph Attention Network

in combination with GAN architecture to better capture relative importance of surrounding agents and semantic features of the environment.

2.5.2 Non-sequential Models

Learning motion patterns in complex environments requires the model to generalize across non-uniform, context-dependent behaviors. Specifying causal constraints, e.g. through the Markovian assumption for the sequential models and additionally the particular functional form for the physics-based methods, might be too restrictive for these situations. Alternatively, instead of focusing on the local transitions of the system, *non-sequential approaches* aim to directly learn a distribution over long-term trajectories, that the observed agent may follow in the future, i.e. learn a set of full motion patterns from data.

Most basic non-sequential approaches are based on clustering the observed trajectories, which creates a set of long-term motion patterns [29–32, 57]. This way global structure of the workspace is imposed on top of a sequential model. Clustering-based approaches are illustrated in Fig. 2.7 (d). Bennewitz et al. [29, 30] cluster recorded trajectories of humans into global motion patterns using the expectation maximization (EM) algorithm and build an HMM model for each cluster. For prediction, the method compares the observed track with the learned motion patterns, and reasons about which patterns best explain it. Uncertainty is handled by probabilistic mixing of the most likely patterns. Similarly, Zhou et al. [367] models the global motion patterns in a crowd with Linear Dynamic Systems using EM for parameters estimation. Several authors [207, 243] propose graph structures to efficiently capture the branching of trajectory clusters. Chen et al. [57] propose a method for dynamic clustering of the observed trajectories, assuming that the set of complete motion patterns may mot be available at the time of prediction, e.g. in new environments. Sung et al. [301] propose to represent the agent's states as short trajectories rather than static positions. This higher level of abstraction provides greater flexibility to represent not only position, but also velocity and intention. Suraj et al. [302] directly use a large-scale database of observed trajectories (up to 10 millions) to estimate the future positions of a vehicle given only its position, rotation and velocity. Combining the concepts of local motion patterns and clustering, Carvalho et al. [51] represent each cluster with a piece-wise linear vector field over an arbitrary state-space mesh.

Several approaches use Gaussian processes (GPs) or mixture models as cluster centroids representation [149, 309, 354]. Tay and Laugier [309] introduce an approach to predict motion of a dynamic object in known scenes based on Gaussian mixture models and Gaussian processes. Kim et al. [149] model continuous dense flow fields from a sparse set of vector sequences. Yoo et al. [354] propose to learn most common patterns in the scene and their co-occurrence tendency using topic mixture and Gaussian mixture models. Observed trajec-

tories are clustered into several groups of typical patterns that occur at the same time with high probability. Given a set of observed trajectories, prediction is performed considering the dominant pattern group. Makansi et al. [206] present a Mixture Density Network architecture which generates multiple hypotheses of future position in fixed interval Δt and then fits a mixture of Gaussian or Laplace distributions to these hypothesis.

Clustering-based methods, discussed so far, generalize statistical information in a particular environment. In comparison, location-invariant methods, based on matching the observed partial trajectory to a set of prototypical trajectories, can be used in arbitrary free space [112, 144, 339], see Fig. 2.7 (e). Hermes et al. [112] predict trajectories of vehicles by comparing the observed track to a set of motion patterns, clustered with a rotationally-invariant distance metric. In their Probabilistic Hierarchical Trajectory Matching (PHTM) approach, Keller et al. [144] propose a probabilistic search tree of sample human trajectory snippets to find the corresponding matching sub-sequence. Xiao et al. [339] decompose the set of sample trajectories into pre-defined motion classes, such as wandering or stopping, rotating and aligning them to start from the same point and have the longest span along the same axis. In contrast, skipping the clustering step, Nikhil and Tran Morris [222] propose a simple method to map the input trajectory of fixed length to the full future trajectory using a Convolutional Neural Network.

For interaction-aware non-sequential motion prediction, several authors consider the case with two interacting agents [136, 200]. Käfer et al. [136] propose a method for joint pairwise vehicle trajectory estimation at intersections. Comparing the observed motion pattern to the ones stored in a motion database, several prospective future trajectories are extracted independently for each vehicle. Probability of each pair of possible future trajectories is then estimated. Luber et al. [200] model joint pairwise interactions between two people using social information. Authors learn a set of dynamic motion prototypes from observations of relative motion behavior of humans in public spaces. An unsupervised clustering technique determines the most likely future paths of two humans approaching a point of social interaction.

In contrast to multi-agent clustering, Trautman and Krause [313] use Gaussian Processes for making single-agent trajectory predictions. Then, an interaction potential re-weights the set of trajectories based on how close people are located to each other at every moment. A follow-up work [314] incorporates goal information into the model: the goal position is added as a training point into the GP. Another approach by Su et al. [298] uses a social-aware LSTM-based crowd descriptor, which is later integrated into the deep Gaussian Process to predict a complete distribution over future trajectories of all people.

Recently, several approaches for non-sequential prediction of vehicle motion using CNNs were presented [67, 72, 117]. An uncertainty-aware CNN-based vehicle motion prediction approach is presented by Djuric et al. [72].

Authors use a high-definition map image with projected prior motion of the target vehicle and full surrounding context as an input to the CNN, which produces the short-term trajectory of the target vehicle. The approach is extended by Cui et al. [67] to inferring multi-modal predictions. Hong et al. [117] propose two methods for output representation using multi-modal regression with uncertainty or stacks of grid-map crops. Chai et al. [53] use a fixed set of state-sequence "anchor" trajectories (clustered from training data), which correspond to possible modes of future behavior, as input to a CNN for mid-level scene features inference, and predict a discrete distribution over these anchors. For each anchor, the method regresses offsets from anchor waypoints along with uncertainties, yielding a Gaussian mixture at each time step.

2.6 Planning-based Approaches

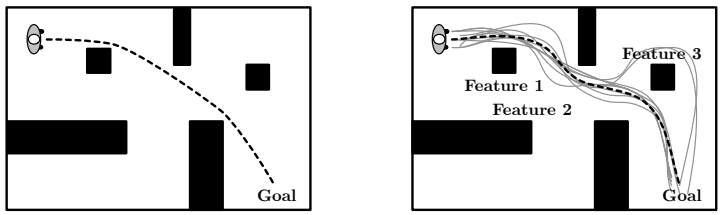
Planning-based approaches solve a sequential decision-making problem by reasoning about the future to infer a model of agent's motion. These approaches follow the *Sense-Reason-Act* paradigm introduced earlier in Sec. 2.3. Unlike the previous two modeling approaches, the planning-based approach incorporates the concept of a rational agent when modeling human motions. By placing an assumption of rationality on the human, the models used to represent human motion must take into account the impact of current actions on the future as part of its model. As a result, much of the work covered in this section use objective functions that minimizes some notion of the total cost of a sequence of actions (motions), and not just the cost of one action in isolation.

Here we classify planning-based approaches into two sub-categories, depicted in Fig. 2.8. *Forward planning-based approaches* (Sec. 2.6.1) use a predefined cost function to predict human motion, and *inverse planning-based approaches* (Sec. 2.6.2) infer the cost (or policy) function from observations of human behavior and then use that cost (or policy) function to predict human motion.

2.6.1 Forward Planning Approaches

Motion and path planning methods

To make basic goal-informed predictions, several methods use optimal motion and path planning techniques with a hand-crafted cost-function [46, 100, 322, 340, 353]. Bruce and Gordon [46] propose to use a path planning algorithm to infer how a person would move towards destinations in the environment. Predictions are performed using a set of learned goals. Gong et al. [100] use multiple long-term goal-directed path hypothesis from different homotopy classes, generated with a modified A* algorithm [37]. Xie et al. [340] describe a Dijkstra-based approach to predict human transitions across *dark energy* fields generated from video data. Every goal location generates an attractive *dark*



(a) MDP-based forward planning model

(b) IRL-based inverse planning model

Figure 2.8: Examples of the planning-based approaches: (**a**) forward planning approach, which uses a predefined cost function (e.g. Euclidean distance), and (**b**) inverse planning approach, which infers the feature-based cost function from observations.

matter Gaussian force field, while every non-walkable location generates a repulsive one. The dark matter functional objects, the map and the goals are inferred on-line using a Monte Carlo Markov Chain technique. For predicting human motion in a crowd, Yi et al. [353] introduce an energy map to model the traveling difficulty of each location in the scene, accounting for obstacles layout, moving people and stationary groups. The energy map is personalized for each observed agent, and the Fast Marching Method (FMM) [289] is used to predict the person's path. Vasishta et al. [322] use A* search over the potential cost-map function for pedestrian trajectory prediction, aiming to recognize illegal crossing intention of the observed agent. The potential field accounts for semantic properties of the urban environment.

Other methods model the probabilities of future motion based on cost-to-go value estimates [35, 142, 266, 324, 352]. Yen et al. [352] propose a probabilistic goal-directed motion model that accounts for several goals in the environment. The method computes the cost-to-go function for each goal and evaluates the probabilities of feasible transitions in each state. A person's trajectory is predicted using a particle filter with Monte-Carlo sampling. Best and Fitch [35] propose a Bayesian framework that exploits the set of path hypotheses to estimate the intended destination and the future trajectory. To this end, a probabilistic dynamical model is used, which evaluates next states of the agent based on the decrease of the distance to the intended goal. Hypothesis are generated from the Probabilistic Roadmap (PRM). Karasev et al. [142] solve the prediction problem using a jump-Markov Decision Process, modeling the agents' behavior as switching non-linear dynamical systems. A soft MDP policy describes the nonlinear motion dynamics, and the latent goal variable governs the switches. The method uses hand-crafted costs for each surface type (e.g. sidewalk, crosswalk, road, grass), and handles time-dependent information such as traffic signals. Instead of using an MDP formulation, Vasquez [324] proposes the Fast Marching Method (FMM) to compute the cost-to-go function

for a set of goals. The predictor uses a velocity-dependent probabilistic motion model, describes the temporal evolution along the predicted path, and offers a gradient-based goal prediction that allows quick recognition of the intended destination changes.

Multi-agent forward planning

Most planning-based methods discussed so far do not consider interactions between agents in the scene. To account for presence of other agents, several authors propose to modify individual optimal policies locally with physics-based methods [267, 268, 319, 335] or imitation learning Muench and Gavrila [217]. A crowd simulation approach that combines global planning and local collision avoidance is presented by van Den Berg et al. [319]. A global path for each agent is computed using a Probabilistic Road Map (PRM), considering only static obstacles. Local collision avoidance along the global path is done jointly for all agents using the Reciprocal Velocity Obstacles (RVO) [318] method. Another method [267], presented in Chapter 3 of this thesis, extends the MDP-based approaches [142, 370] with a fast random-walk based method to generate joint predictions for all observed people using social forces. In the follow-up publication this approach is extended considering group-based social motion constraints [268]. Wu et al. [335] extend the gridmap transition-based and reachability-based framework [66, 254] with automatic inference of local goal points, and calculate the stochastic policy in each cell, augmenting the physics-based dynamics with optimal motion direction. The motion of pedestrians is predicted jointly with other traffic participants by risk checking of future states based on gap acceptance model [42]. Instead of using a physics-based approach (e.g. social forces) for augmenting the MDP-based predictor, Muench and Gavrila [217] propose to learn an additional interaction-aware Q-function with imitation learning.

A number of approaches consider cooperative planning in joint state-space that includes all agents [18, 43, 56, 263]. Broadhurst et al. [43] use Monte Carlo sampling to generate probability distributions over future trajectories of the vehicles and pedestrians jointly. The approach considers several available actions for each agent in the scene: each vehicle executes one of the hand-crafted behaviors, and humans are assumed to move freely in all directions. Also Rösmann et al. [264] considers planning for cooperating agents. A set of topologically distinct candidate trajectories for each person is computed using trajectory optimization techniques [263]. Among those trajectories the best candidate is chosen according to a metric that includes group integrity, right versus left motion bias and curvature constraints. Finally, the encounter is resolved jointly in an iterative fashion. The interaction point of minimal spatial separation is computed between each two people, who adjust their trajectories accordingly, possibly switching to a different topological candidate. Mavro-

giannis and Knepper [208] represent multi-agent interaction through the use of braid groups (topological patterns) which formalize trajectories sets. At inference time, the problem of predicting joint trajectories is posed as a graph search in a permutation graph.

Joint planning for the robot and the human is addressed by several works [22, 56, 95]. Assuming availability of a fixed set of goals, Bandyopadhyay et al. [22] solve an optimal motion problem for each of it, and generate appropriate motion policies. The latter are used to estimate the future evolution of the joint state-space of the robot and the human. Galceran et al. [95] introduce a multi-policy decision-making systems to generate robot motions based on predicted movements of other agents in the scene, estimated with a changepoint-based technique [87]. Likelihood of future actions are sampled from the policies. The final prediction is generated by an exhaustive search of closed-loop forward simulations of these samples. The approach is well suited for predicting future macro-actions (i.e. turn left or right, slow down or speed up). Bahram et al. [18] generates joint robot and agents' motions using a sequential game theory technique. The approach presents an interactive prediction and planning loop where a sequence of predictions (i.e. motion primitives) is generated for the ego-vehicle by considering the sequential evolution of the entire scene. Chen et al. [56] develop a de-centralized multi-agent collision avoidance algorithm, which resolves local interactions with a learned joint value function that implicitly encodes cooperative behaviors.

2.6.2 Inverse Planning Approaches

Forward planning approaches, discussed so far, make an explicit assumption about the optimality criteria (reward or cost function) of an agent's motion. In this section we discuss algorithms that estimate the reward function of agents (or directly a policy) from observations, using statistical and imitation learning techniques (for a survey on imitation learning techniques applied to robotic systems we refer the reader to [227]). Inverse planning methods assume that the reward or cost function, which depends on contextual and social features and defines the rational behavior, can be learned from observations (see Fig. 2.8 (b)).

Single-agent inverse planning

In their influential work, Ziebart et al. [370] propose to learn a reward function yielding goal-directed behavior of pedestrians using maximum entropy inverse optimal control (MaxEnt IOC). Humans are assumed to be near-optimal decision makers with stochastic policies, learned from observations, which are used to predict motion as a probability distribution over trajectories. Building upon [370], Kitani et al. [153] expand it to include the labeled semantic map

of the environment. An IOC method takes the semantic map as an input, and learns the feature-based cost function that captures agents' preferences for e.g. walking on the sidewalk, or keeping some distance from parked cars. Previtali et al. [247] propose an approach that adopts linear programming formulation of IRL. Using a discrete and non-uniform representation of the 2D workspace, it scales linearly with respect to the size of the environment. Chung and Huang [62] present an MDP-based model that describes spatial effects between agents and the environment. The authors use IRL to estimate cost of each state as a linear combination of trajectory length, static and dynamic obstacle avoidance and steering smoothness. Special context-based spatial effects (SSE) are identified by comparing the costs of the states, learned with IRL, and the actual observed trajectories. A follow-up work [63] introduces a feature-based representation of SSEs, which can be modeled before being naturally observed, as in [62].

Instead of IRL, other works use different techniques to learn the reward function [119, 255]. Rehder et al. [255] solve the problem of intention recognition and trajectory prediction in one single Artificial Neural Network (ANN). The destinations and costly areas are predicted from stereo images using a recurrent Mixture Density Network (RMDN). Planning towards these destinations is performed using fully Convolutional Neural Networks (CNN). Two different architectures for planning are proposed: an MDP network and a forward-backward network, both using contextual features of the environment. Huang et al. [119] propose an approach that exploits two CNNs to learn a reward function considering spatial and temporal contextual information from a video sequences. A Spatial Matching Network (SMN) learns the spatial context of human motion. An Orientation Network (ON) is used to model the position variation of the object. The Dijkstra algorithm is used to find the minimum cost solution over a graph whose edges' weights are set by considering the reward function and the facing orientation computed by the two networks (SMN and ON).

All the detailed methods show that IRL or similar methods are providing powerful tools to learn human behaviors. Furthermore, Shen et al. [292] show that under some particular requirements (i.e. when the feature vector, model parameter and output representation are invariant under a rigid body transformation of the world fixed coordinate frame), IRL is suitable for learning location-independent transferable motion models.

Imitation learning

Instead of first learning a reward function and then applying planning techniques to generate motion predictions, imitation learning approaches directly extract a policy from the data. Generative Adversarial Imitation Learning (GAIL) approach, proposed by Ho and Ermon [114], aims for matching long-term dis-

tributions over states and actions. It uses a GAN-based [101] optimization procedure, in which a discriminator tries to distinguish between observations from experts and generated ones by making model rollouts. Afterwards, a model is trained to make predictions that yield similar long-term distributions over states and actions. This method has been successfully applied to learning human highway driving behavior [168] and training joint pedestrian motion models [105]. Li et al. [188] extend GAIL by introducing a component to the loss function, which maximizes the mutual information between the latent structure and observed trajectories. They test their approach in a simulated highway driving scenario, predicting the driver's actions given an input image and auxiliary information (e.g. velocity, last actions, damage), and show that it is able to imitate human driving, while automatically distinguishing between different types of behaviors.

Differently from GAIL, the deep generative technique by Rhinehart et al. [257] adopts a fully differentiable model, which is easy to train without the need of an expensive policy gradient search. By minimizing a symmetrized cross-entropy between the distributions of the policy and of the demonstration data, the method allows to learn a policy that generates predictions which balance precision (i.e. avoid obstacle areas) and diversity (i.e. being multi-modal).

Multi-agent inverse planning

In the following we review several inverse planning approaches that predict multi-agent motions [90, 162, 167, 179, 202, 241]. Kuderer et al. [167] and Kretzschmar et al. [162] propose a continuous formulation of the MaxEnt IOC [370] by considering a continuous spline-based trajectory representation. Their method relies on several features (e.g. travel time, collision avoidance) to capture physical and topological aspects of the pedestrians trajectories. Pfeiffer et al. [241] extend the latter works by introducing the variable end-position of the each trajectory, thus reasoning over the agents' goals. Walker et al. [331] present an unsupervised learning approach for visual scene prediction. The approach exploits mid-level elements (i.e. image patches) as building blocks for jointly predicting positions of agents in the scene and changes in their visual appearance. The learned reward function defines the probability of a patch moving to a different location in the image. To generate predictions, the method performs a Dijkstra search on the learned reward function considering several goals. Ma et al. [202] combine the Fictitious Play [45] game theory method with the deep learning-based visual scene analysis. Future paths hypothesis are generated jointly and iteratively: each pedestrian adapts her motion based on the predictions of the other pedestrians' actions. IRL's reward function features encode social compliance, neighborhood occupancy, distance to the goal and body orientation. Gender and age attributes, extracted with a deep network from video, define the possible average velocity of pedestrians.

Lee et al. [179] formulate the prediction problem as an optimization task. The method reasons on multi-modal future trajectories accounting for agent interactions, scene semantics and expected reward function, learned using a sampling-based IRL scheme. The model is wrapped into the single end-to-end trainable RNN encoder-decoder network, called DESIRE. The RNN architecture allows incorporation of past trajectory into the inference process, which improves prediction accuracy compared to the standard IRL-based techniques.

The previously discussed approaches for joint prediction assume multi-agent settings with rational and cooperative behavior of all agents. Differently, several approaches [111, 178] address the problem by modeling one target person as a rational agent, acting in a dynamic environment. The influence of other agents then becomes part of the stochastic transition model of the environment. For example, Henry et al. [111] propose an IRL-based method for imitating human navigation in crowded environments. They conjecture that humans take into account the density and velocity of nearby people and learn a reward function that weights between these and additional features. Another approach by Lee and Kitani [178] learns a reward function that explains behavior of a wide receiver in American football, whose strategy takes into account the behavior of the defenders. Models of the dynamic environment (e.g. linear or Gaussian Processes) are used as transitions in the IRL framework.

Rhinehart et al. [259] has developed a multi-agent forecasting model called Estimating Social-forecast Probabilities (ESP) that uses exact likelihood inference (unlike VAEs or GANs) derived from a deep neural network for forecast trajectories. In contrast to most standard trajectory forecasting methods, the approach is able to reason conditionally based on additional information that it was not trained to use by accepting agent goals at test time. The approach uses a generative multi-agent model in order to perform PREdiction Conditioned On Goals (PRECOG).

2.7 Contextual Cues

In this section we discuss the categorization of the contextual cues, in those dealing with the target agent (Sec. 2.7.1), the other dynamic agents (Sec. 2.7.2) and the static environment (Sec. 2.7.3).

2.7.1 Cues of the Target Agent

Most essential cues, used to predict future states of an agent, are related to the agent itself. To this end most of the algorithms use current position and velocity of the target agent [18, 23, 30, 31, 80, 91, 107, 142, 153, 166, 167, 201, 236, 257, 268, 313, 335, 370], often considering also the history of recent states/velocities. Position and velocity are also the main attributes of the target agent in vehicle motion prediction tasks [43, 112, 136]. Considering the

head orientation or full articulated pose of the person [41, 109, 156, 157, 212, 250, 265, 287, 317] may bring valuable insights on the target agent's immediate intentions or their awareness of the environment. Considering additional semantic attributes of the target agent may further refine the quality of predictions: gender and age in [202], personality type [32], class of the dynamic agent (e.g. a person or a cyclist in pedestrian areas, motorcycle, car or a truck on a highway) [6, 20, 66], person's attention and awareness of the robot's presence in [41, 157, 226], raised arm as a bending intention indicator for cyclists [157, 245].

2.7.2 Cues of Other Dynamic Agents

Most of the time all agents navigate in a shared environment, adapting their actions, timing and route based on the others' presence and behavior. Therefore for predicting motion it is beneficial to consider interaction between moving agents. We classify the existing approaches in three categories: *unaware predictors*, *individual-aware predictors* and *group-aware predictors*.

The class of unaware predictors includes all methods that generate motion prediction for a single agent, considering only the static contextual cues of the environment. Having no need to explicitly define or learn the interaction model, these methods are simpler to set up, require less training data to generalize, typically have less parameters to estimate. Simpler physics-based methods, such as linear velocity projection or constant acceleration models, are unaware predictors [19, 66, 82, 83, 94, 159, 322, 323, 341, 368]. Many pattern-based [29, 30, 51, 54, 57, 99, 107, 108, 112, 123, 147, 149, 165, 166, 206, 207, 214, 222, 243, 261, 279, 301, 302, 305, 311, 317, 334, 339, 344, 346] and planning-based methods [100, 142, 153, 257, 266, 324, 352, 370] are unaware predictors, due to the increase of complexity for conditioning the learned transition patterns or optimal actions on the presence and positions of other agents. Methods for predicting pedestrians crossing behavior [104, 143, 156, 212, 250, 265, 287] and cyclist motion [245, 246, 278, 358] typically treat each agent individually.

Individual-aware predictors methods consider the interaction between agents by modeling or learning their influence on each other. Physics-based methods that use social forces [41, 80, 91, 141, 198, 226, 356] or similar local interaction models [23, 139, 149, 201, 233, 236, 237, 240, 262, 347] are classical examples of individual-aware prediction models. A pattern-based approach by Ikeda et al. [124] models deviations from the desired path using social forces. In general, however, learning joint motion patterns is a considerably harder task. For example, Trautman and Krause [313], Trautman et al. [314] learn unaware motion patterns, and then evaluate the predicted probability distribution over the joint paths using an explicit interaction potential. Luber et al. [200] learn pairwise joint motion patterns of two humans approaching the spatial point of

interaction. The approach by Yoo et al. [354] learns which motion patterns are likely to occur at the same time and uses this information for predicting the future motion of several dynamic objects. Some approaches propose to learn a motion policy or reward function that accounts for dynamic objects in the surrounding [62, 63, 111, 178, 327]. Chapter 3 describeS an MDP planning-based method, where optimal policies of people are locally modified to account for other dynamic entities [267]. Wu et al. [335] and Zechel et al. [357] discount predicted transition probabilities to states in collision with other agents. Muench and Gavrila [217] decompose the interactive planning problem into two policies with the corresponding Q-functions: one for prediction in static environment, and another for interaction prediction in an obstacle-free environment. Many deep learning methods consider interactions between participants: explicitly modeling interacting entities [4, 11, 25, 60, 79, 89, 90, 105, 109, 120, 127, 158, 235, 242, 251, 259, 277, 280, 293, 298, 320, 321, 328, 343, 345, 362], implicitly as a result of pixel-wise prediction [331], or by learning a joint motion policy [179, 202, 290, 359]. Many vehicle prediction methods consider interaction between traffic participants, e.g. [3, 6, 18, 43, 53, 67–69, 71, 72, 117, 131, 136, 147, 170, 184, 234, 252, 297]. Kooij et al. [157] consider whether the ego-vehicle is on a potential collision course when predicting the road user path in their SLDS-based approach.

Group-aware predictors also recognize affiliations and relations of individual agents and respect the probability of them traveling together, as well as model an appropriate reaction of other agents to the moving group formation. For example, several physics-based methods model group relations by introducing additional attractive forces between group members [61, 140, 216, 237, 248, 262, 288, 295, 347]. Several learning-based approaches that use LSTMs [4, 25, 242, 293, 321, 360] may be capable of implicitly learning intra- and inter-group coherence behavior, however only the work by Bisagno et al. [39] states this capability explicitly. A planning-based approach which implicitly respects group integrity by increasing the costs of passing between group members is presented by Rösmann et al. [264] and an approach that explicitly models group motion constraints by Rudenko et al. [268], which is presented here in Chapter 3.

Algorithms using high-level context information about dynamic agents produce more precise predictions in a variety of cases. Learning advanced social features of human motion improves interactive predictors performance, for instance different parameters for interactions of heterogeneous agents [91], advanced motion criteria such as *social comfort* of navigation [167, 200, 241] or "desire to move with the flow" or "avoid dense areas" [111]. Some approaches model prior knowledge in terms of the dynamics of moving agents [179, 264], human attributes and personal traits [202]. Chung and Huang [63] present a general framework for learning context-related spatial effects, which affect the

human motion, such as avoiding going through a waiting line, or in front of a person, who observes the work of art in a museum.

Modeling also the influence of the robot's presence on the agents' paths is another interesting line of research: Trautman and Krause [313] and Oli et al. [226] tackle this problem by placing the robot as a peer-interacting agent among moving humans. Several authors [162, 167, 241, 264] optimize joint trajectories for all humans and the robot. A relevant case of modeling the effect of robotic herd actions on the location and shape of the flock of animals is studied by Sumpter and Bulpitt [299]. Similarly, Schmerling et al. [281] condition human response on the candidate robot actions for modeling pairwise human-robot interaction. Eiffert and Sukkarieh [79] include the robot as an interacting agent in the LSTM-based predictor. Tang and Salakhutdinov [307] compute a conditional probability density over the trajectories of other agents given the hypothetical rollout for the robot.

2.7.3 Cues of the Static Environment

Humans adapt their behaviors according not only to the movements of the other agents but also to the environment's shape and structure, making extensive use of its topology to reason on the possible paths to reach the long-term goal. Many existing prediction algorithms make use of such geometric information of the environment.

Some approaches produce *unaware predictions*, assuming an obstacle-free environment. This category includes several physics-based approaches [19, 41, 82, 83, 94, 240, 283, 368]. Pattern-based methods usually model obstacles implicitly, by learning collision-free patterns [29, 51, 54, 57, 81, 88, 108, 109, 123, 129, 134, 149, 163, 165, 166, 206, 207, 214, 243, 278, 279, 300, 301, 305, 309, 311, 325, 333, 334, 344, 346, 354]. When facing a change in the obstacles' configuration, such patterns become obstacle-unaware. Location-independent motion patterns are usually obstacle-unaware [99, 112, 200, 222, 317, 339]. Pedestrian crossing prediction methods typically assume obstacle-free environment [104, 143, 156, 157, 212, 250, 265, 287], as well as most of the vehicle prediction methods [6, 69, 71, 147, 234, 252, 302], which assume the road-surface to be free of static obstacles. Finally, many methods consider only dynamic entities, but no static obstacles in the environment [4, 9–11, 18, 25, 31, 39, 43, 68, 79, 89, 90, 105, 120, 127, 136, 150, 167, 184, 235, 242, 251, 293, 298, 313, 314, 321, 327, 328, 343, 356, 360].

In several approaches the exact pose of the objects is known and utilized to compute more informed predictions (we refer to such methods as to *obstacle-aware* methods). Mainly the social force-based and similar techniques model the interaction between the moving agents and individual static obstacles [80, 91, 139, 141, 142, 162, 198, 201, 226, 233, 236, 237, 259, 262, 318, 347,

357]. Several location-independent pattern-based methods [13, 14] can handle static objects avoidance.

Still, obstacle-aware methods may fail in very cluttered environments, due to the complexity of representing an environment with a set of individual obstacles. To overcome this difficulty many prediction approaches use maps which are a more complete representation of the environment (we call them *map-aware* methods). Occupancy grid maps are the most common representation for these approaches, e.g. in the physics-based approach by Rehder and Klöden [254] reachability-based transitions are calculated on a binary grid-map. Particularly the planning-based approaches use this kind of representation: thanks to the map they can infer global, intentional behaviors of the agents [35, 46, 56, 62, 63, 100, 111, 124, 190, 241, 247, 264, 266–268, 324, 340, 352, 353, 370]. Fig. 2.5 shows the difference between the *pure motion based predictions*, the *obstacle-aware* and the *map-aware* approaches. The latter perform better in terms of global obstacle avoidance behavior during prediction.

Semantic map based approaches extend the map-aware approaches by considering various semantic attributes of the static environment. A semantic map [20, 66, 142, 153, 217, 255, 257, 258, 261, 280, 292, 305, 320, 322, 323, 362] or extracted features from a top-down image [158, 277, 307, 345] can be used to capture people preferences in walking on a particular type of surfaces. Furthermore, planning-based methods often use prior knowledge on potential goals in the environment [35, 142, 247, 266, 324]. Location- and time-specific information in the particular environment may help to improve prediction quality [214, 300].

Due to the high level of structure in the environment, methods in autonomous driving scenarios extensively use available semantic information, such as street layout and traffic rules [3, 53, 61, 67, 72, 104, 117, 131, 143, 156, 170, 179, 239, 245, 246, 297, 341] or current state of the traffic lights [104, 131, 142], also for predicting pedestrian and cyclist motion [107, 157, 159].

2.8 Motion Prediction Evaluation

An important challenge for motion prediction methods is the design of experiments to evaluate their performance with respect to other methods and the requirements from the targeted application. In this section we review and discuss common metrics and datasets to this end.

2.8.1 Performance Metrics

Due to the stochastic nature of human decision making and behavior, exact prediction of trajectories is rarely possible, and we require measures to quantify the similarity between predicted and actual motion. Different prediction types – see Fig. 1.1 – require different measures: for single trajectories we need geomet-

ric measures of trajectory similarity or final displacement, for parametric and non-parametric distributions over trajectories we can use geometric measures as well as difference measures for probability distributions. Metrics, commonly used in the literature, are summarized in Table 2.1.

Geometric accuracy metrics

Geometric measures are the most commonly used across all application domains. Several surveys have considered the topic of trajectory analysis and comparison [215, 231, 249, 361, 364] where, based on the previous ones, only the recent survey by Quehl et al. [249] specifically considers geometric similarity measures for trajectory prediction evaluation. In addition to that, we review the probabilistic metrics and the assessment of distributions with geometric methods, and the experiments to evaluate robustness in the following sub-sections.

Summarizing [215, 249], we consider eight metrics:

Mean Euclidean Distance (MED), also called *Average Displacement Error* (ADE), averages Euclidean distances between points of the predicted trajectory and the ground truth that have the same temporal distance from their respective start points. An alternate form computes MED in a subspace between coefficients of the trajectories' principal components (PCA-Euclid). A third variant (MEDP) is a path measure able to compare paths of different length. For each (x, y)-point of the predicted path, the nearest ground truth point is searched. Being a path measure, MEDP is invariant to velocity differences and temporal misalignment but does not account for temporal ordering. A fourth variant (n-ADE) measures MED only on non-linear segments of trajectories. MED measures are widely used by many authors across all domains, see Table 2.1. Many authors evaluate probabilistic predictions by computing expected MED under the predictive distribution, referring to it as *mean ADE*, *weighted mean ADE*, or, abusing notation, simply MED or ADE. This type of evaluation, however, does not measure how good the predictive distribution matches the ground truth distribution, falling short of being a true probabilistic measure. For example, it favors point predictions and avoids larger variances, as they often increase the expected ADE.

Dynamic Time Warping (DTW) [34] computes a similarity metric between trajectories of different length as the minimum total cost of warping one trajectory into another under some distance metric for point pairs. As DTW operates on full trajectories, it is susceptible to outliers.

Modified Hausdorff Distance (MHD) [78] is related to the Hausdorff distance as the maximal minimal distance between the points of predicted and actual trajectory. MHD was designed to be more robust against outliers by allowing slack during matching and to compare trajectories of different length. A further variant is the *trajectory Hausdorff* measure (THAU) [177], a path

metric that computes a weighted sum over three distance terms each focusing on differences in perpendicular direction, length, and orientation between the paths. The weights can be chosen to be application-dependent.

Longest Common Subsequence (LCS) [49] aligns two trajectories of different length so as to maximize the length of the common subsequence, i.e. the number of matching points between both trajectories. A good match is determined by thresholding a pair-wise distance and time difference where not all points need to be matched. LCS is more robust to noise and outliers than DTW but finding suitable values for the two thresholds is not always easy.

CLEAR multiple object tracking accuracy (CLEAR-MOTA) was initially introduced as a performance metric for target tracking [33]. In the context of prediction evaluation, it is similar to LCS in that it sums up good matches between points on the predicted trajectory and the ground truth. The difference is that the concept of pair-wise matches/mismatches is more complex including false negatives, false positives and non-unique correspondences.

In addition to the metrics considered in [215, 249], relevant metrics used in the reviewed literature include the *Quaternion-based Rotationally Invariant LCS* (QRLCS), which is the rotationally invariant counterpart of LCS [112], and several measures that quantify different geometric aspects in addition to trajectory or path similarity:

Final Displacement Error (FDE) measures the distance between final predicted position and the ground truth position at the corresponding time point. If the prediction is represented by a distribution, many authors compute expected FDE. FDE however, is not appropriate when there are multiple possible future positions.

Prediction Accuracy (PA) uses a binary function to classify a prediction as correct if the predicted position fulfills some criteria, e.g. is within a threshold distance away from the ground truth. Percentage of correctly predicted trajectories is then reported. PA allows to incorporate suitable invariances into the distance function such as allowing certain types of errors.

As also pointed out by Quehl et al. [249], the challenge in choosing a suitable measure is that each of these measures usually produce quite different results. For the sake of an unbiased and fair evaluation of different prediction algorithms, measures should be chosen not to suit a particular method but based on the requirements from the targeted application. An application which includes a lot of different velocities, for example, should not solely rely on path measures.

Probabilistic accuracy metrics

One of the drawbacks of geometric metrics is their inability to measure uncertainty and also multimodal nature of predictions, e.g. when the target agent may take different paths to reach the goal, or when an observed partial trajec-

	Metric	Used by
Geometric	Average Displacement Error (ADE)	[4, 6, 11, 25, 39, 41, 53, 60, 67–69, 72, 79, 89, 90, 105, 109, 112, 117, 120, 123, 127, 131, 142, 147, 150, 157, 158, 184, 201, 212, 222, 234–236, 242, 246, 250–252, 257, 264, 277–280, 287, 293, 297, 298, 300, 325, 327, 328, 335, 341, 343–347, 353, 354, 358, 360, 362]
	Final Displacement Error (FDE)	[4, 11, 39, 41, 60, 62, 79, 89, 105, 109, 120, 123, 127, 158, 201, 222, 235, 251, 277, 293, 298, 321, 327, 328, 343–346, 360, 362]
	Modified Hausdorff Distance (MHD)	[66, 90, 107, 129, 153, 266–268, 280, 292, 324, 354]
	Prediction Accuracy (PA)	[31, 35, 71, 91, 117, 124]
Probabilistic	Negative Log Likelihood (NLL)	[53, 66, 127, 131, 206, 245, 259, 266, 302]
	Negative Log Loss (NLL)	[153, 202, 247, 307, 324]
	Predicted Probability (PP)	[156, 157, 254, 267, 268]
	Min. Avg. or Final Displacement Error (mADE, mFDE)	[11, 53, 117, 127, 179, 184, 234, 257, 259, 261, 307, 320]
	Cumulative Probability (CP)	[302]

Table 2.1: Metrics to evaluate motion prediction

tory matches several previously learned motion patterns. Moreover due to the stochasticity of the human behaviors, motion prediction algorithms need to be evaluated on their accuracy to match the underlying probability distribution of human movements. Several probabilistic accuracy metrics can be used for this purpose.

Many variational inference and machine learning algorithms [40, 203] use the Kullback-Leibler (KL) divergence [171] to measure dissimilarity of two distributions, e.g. the unknown probability distribution of human behavior $p(\mathbf{s}_{1:T})$ and the predicted probability distribution $q(\mathbf{s}_{1:T}|\theta)$, with θ being a set of parameters of the chosen prediction model. The KL divergence is computed as $d_{KL}(p||q) \simeq \sum_{\mathbf{s}_{1:T}\in\mathbb{S}}\{-p(\mathbf{s}_{1:T})\log q(\mathbf{s}_{1:T}|\theta) + p(\mathbf{s}_{1:T})\log p(\mathbf{s}_{1:T})\}$ with the space of all trajectories $\mathbb{S}$. Minimizing the $d_{KL}(p||q)$ corresponds to maximizing the log-likelihood function for θ under the predicted distribution $q(\mathbf{s}_{1:T}|\theta)$. Different surveyed papers have adopted variants of the KL divergence as accuracy metric for their stochastic predictions.

For example, the **average Negative Log Likelihood** or **average Negative Log Loss** evaluates the negative log likelihood term $\left(\simeq \sum_{\mathbf{s}_{1:T}\in\mathbb{D}} \log q(\mathbf{s}_{1:T}|\theta)\right)$ of d_{KL} from a set of ground truth demonstrations $\mathbb{D} = \left\{\mathbf{s}^i_{1:T}\right\}_{i=1}^{N}$ with the total number of demonstrations N. Furthermore, several approaches use the **Predicted Probability** (PP) metric, $\left(\simeq \sum_{t=1}^{T} q(\mathbf{s}_t|\theta)\right)$ or its negative logarithm, to calculate the probability of the ground truth path (*i.e* $\mathbf{s}_{1:T}$) on the predicted states distri-

bution. For the above metrics, the computation of the log likelihood depends on the chosen model, its induced graph and the corresponding factorization. Finally, the **Cumulative Probability** (CP) metric computes the fraction of the predictive distribution that lies within a radius r from the correct position for various values of r.

Several recently introduced metrics follow a sampling approach to evaluate a probability distribution. **Minimum Average Displacement Error** (mADE) metric [259, 284, 307, 310, 332], as well as *variety loss*, *oracle*, *Minimum over N*, *Best-of-N*, *top n%*, or *minimum Mean Squared Distance* (minMSD), computes Euclidean distance between the ground truth position of the agent $\mathbf{s}_t^*$ at time t and the closest (or the n% closest) of the K samples from the predicted probability distribution: $\min_k \|\mathbf{s}_t^* - \mathbf{s}_t^k\|$. Similarly, **minimum Final Displacement Error** (mFDE) evaluates only the distribution at the prediction horizon T. Such metrics encourage the predicted distribution to cover multiple modes of the ground truth distribution, while placing probability mass according to the mode likelihood. An evaluation of the robustness of top 1 vs. top n% metrics by Bhattacharyya et al. [38] has shown that the *top n%* metric produces more stable results.

Other performance metrics

Prediction accuracy is by far the primary performance indicator in the reviewed literature across approaches and application domains. In particular for long-term prediction methods, authors evaluate accuracy against the prediction horizon [18, 41, 60, 62, 69, 95, 99, 112, 124, 129, 142, 143, 178, 241, 242, 250–252, 254, 267, 268, 300, 302, 311, 323, 335, 343]. Much fewer authors address other aspects of robustness and investigate the range of conditions under which prediction results remain stable and how they are impacted by different types of perturbations.

Experiments to explore robustness evaluate prediction accuracy as a function of various influences: the length or duration of the observed partial trajectory until prediction (addresses the question of how long the target agent needs to be observed for a good prediction) [153, 179, 251], the size of the training dataset [123, 302, 323, 326], number of agents in the scene [259], input data sampling frequency and the amount of sensor noise [31] or amount of anomalies in the training trajectories [108]. Several authors report a separate accuracy measurement for the more challenging (e.g. non-linear or anomalous) part of the test set [89, 123, 157], or evaluate the model's performance on different classes of behavior, e.g. walking or stopping [279]. Analysis of generalization, overfitting and input utilization by a neural network, presented by Schöller et al. [284], makes a good case for robustness evaluation.

Furthermore, to quantify efficiency of a prediction method, some authors relate inference time to the number of agents in the scene [267, 268, 311], and

only a few papers provide an analysis of their algorithms' complexity [35, 54, 143, 268, 362].

2.8.2 Datasets

In order to evaluate the quality of predictions, predicted states or distributions are usually compared to the ground truth states using standard datasets of recorded motion. Availability of annotated trajectories, represented with the sequence of states or bounding boxes in the top-down view, sets prediction benchmarking datasets aside from the other popular computer vision datasets, where the ground truth state of the agent is not available and is difficult to estimate.

Common recording setup includes a video-camera with static top-down view of the scene, or ground-based lasers and/or depth sensors, mounted on a static or moving platform. Detected agents in each frame are labeled with unique IDs, and their positions with respect to the global world frame are given as (x, y) coordinates together with the frame time-stamp t, i.e. (id, t, x, y). Often the coordinate vector is augmented with orientation and velocity information. Furthermore, social grouping information, gaze directions, motion mode or maneuver labels and other contextual cues can be provided. Apart from this specific form of labeling, further requirements to prediction benchmarking datasets include interaction between agents, varying density of agents, presence of non-convex obstacles in the environment, availability of the semantic map and long continuous observations of the agents.

In Table 2.2 and 2.3 we review the most popular datasets, used for evaluation in the surveyed literature. Out of many datasets, used for benchmarking by different authors, we picked those used by at least two independent teams, excluding the creators of the dataset. We believe that this is a good indication of the dataset's relevance, which also supports the primary purpose of benchchmarking – comparing performance of different methods on the same dataset. Additionally, in Table 2.4 we include three recent datasets, which do not meet the selection criterion, but cover valuable aspects, missing from the earlier datasets. This includes the first dataset of cyclists trajectories [246], the first dedicated benchmark for human trajectory prediction [160, 276] and our first dataset of human motion trajectories with accurate motion capture data [270], presented in Chapter 5 of this thesis.

Dataset	Location	Agents	Sensors	Scene description	Duration and tracks	Annotations and sampling rate
ETH [236]	Outdoor	People	Camera	2 pedestrian scenes, top-down view, moderately crowded	25 min, 650 tracks	Positions, velocities, groups, maps @2.5 Hz
Used by: [4, 11, 31, 39, 41, 62, 105, 120, 123, 127, 150, 158, 201, 222, 235, 242, 251, 277, 313, 321, 327, 328, 343, 345, 347, 360, 362]						
UCY [181]	Outdoor	People	Camera	2 pedestrian scenes (sparsely populated Zara and crowded Students), top-down view	16.5 min, over 700 tracks	Positions, gaze directions –
Used by: [4, 11, 25, 35, 39, 41, 105, 109, 120, 123, 127, 158, 201, 202, 222, 235, 237, 251, 277, 320, 321, 328, 343, 345, 347, 360, 362]						
Stanford Drone Dataset [262]	Outdoor	People, cyclists, vehicles	Camera	8 urban scenes, ~900 m^2 each, top-down view, moderately crowded	5 hours, 20k tracks	Bounding boxes @30 Hz
Used by: [53, 66, 79, 90, 129, 206, 261, 277, 280, 320, 321, 362]						
NGSIM [64, 65]	Outdoor	Vehicles	Camera network	Recording of the US Highway 101 and Interstate 80, road segment length 640 and 500 m	90 min	Local and global positions, velocities, lanes, vehicle type and parameters, @10 Hz
Used by: [6, 68–71, 137, 145, 168, 184, 218, 307, 362]						
highD Dataset [161]	Outdoor	Vehicles	Camera	6 different highway locations near Cologne, top-down view, varying densities with light and heavy traffic	Over 110k vehicles, 447 driven hours	Positions and additional features, e.g. THW, TTC @25 Hz
Used by: [70, 128, 145, 148, 210, 218, 336]						
Edinburgh [205]	Outdoor	People	Camera	1 pedestrian scene, top-down view, 12 x 16 m^2, varying density of people	Several months, 92k tracks	Positions @9 Hz
Used by: [51, 80, 89, 247, 266, 344]						
Grand Central Station Dataset [366]	Indoor	People	Camera	Recording in the crowded New York Grand Central train station	33 minutes	Tracklets @25 Hz
Used by: [89, 298, 343, 344, 346, 353]						

Table 2.2: Overview of the motion trajectories datasets (part 1)

Dataset	Location	Agents	Sensors	Scene description	Duration and tracks	Annotations and sampling rate
VIRAT [224]	Outdoor	People, cars, other vehicles	Camera	16 urban scenes, 20–50° camera view angle towards the ground plane, homographies included	25 hours	Bounding boxes, events (e.g. entering a vehicle or using a facility) @10, 5 and 2 Hz
Used by: [153, 247, 324, 331, 340]						
KITTI [97]	Outdoor	People, cyclists, vehicles	Velodyne, 4 cameras	Recorded around the mid-size city of Karlsruhe (Germany), in rural areas and on highways	21 training sequences and 29 test sequences	3D Positions @10 Hz
Used by: [142, 179, 257, 297, 335]						
Town Center Dataset [28]	Outdoor	People	Camera	Pedestrians moving along a moderately crowded street	5 minutes, 230 hand labelled tracks	Bounding boxes @15 Hz
Used by: [109, 202, 345, 346]						
ATC [48]	Indoor	People	3D range sensors	Recording in a shopping center, 900 m^2 coverage, varying density of people	92 days, long tracks	Positions, orientations, velocities, gaze directions, @10-30 Hz
Used by: [214, 267, 268]						
Daimler Pedestrian Path Prediction Dataset [283]	Outdoor	People	Stereo camera	Recording from a moving or standing vehicle, pedestrians are crossing the street, stopping at the curb, starting to move or bending in	68 tracks of pedestrians, 4 sec each	Positions, bounding boxes, stereo images, calibration data @17 Hz
Used by: [279, 280, 287]						
L-CAS [349]	Indoor	People	Velodyne	Recording in a university building from a moving or stationary robot	49 minutes	Positions, groups, Velodyne scans @10 Hz
Used by: [251, 300]						

Table 2.3: Overview of the motion trajectories datasets (part 2)

Dataset	Location	Agents	Sensors	Scene description	Duration and tracks	Annotations and sampling rate
Tsinghua-Daimler Cyclist [246] Used by: [278]	Outdoor	Cyclists	Stereo camera	Recording from a moving vehicle	134 tracks	Positions, road topology @5 Hz
TrajNet [276] **TrajNet++** [160] Used by: [346]	Outdoor	People	Cameras	Superset of datasets, collecting also relevant metrics and visualization tools	Superset of image-plane and world-plane datasets	Bounding boxes and tracklets, datasets recording at different frequencies
THÖR [270] Used by: [365]	Indoor	People	Motion capture	Human-robot navigation study in a university lab	Over 600 person and group trajectories in 60 minutes	Positions, head orientations, gaze directions, groups, map, Velodyne scans @100 Hz

Table 2.4: Additional motion trajectories datasets

2.9 Discussion

There has been great progress in developing advanced prediction techniques over the last years in terms of method diversity, performance and relevance to an increasing number of application scenarios. In this section, we summarize and discuss the state of the art and pose the three questions initially raised in the introduction:

Q1: are the evaluation techniques to measure prediction performance good enough and follow best practices?

This is discussed in Sec. 2.9.1 by reviewing the existing benchmarking practices including metrics, experiments and datasets.

Q2: have all prediction methods arrived on the same performance level and the choice of the modeling approach does not matter anymore?

This is discussed in Sec. 2.9.2 where we consider the theoretical and demonstrated ability of the different modeling approaches to solve the motion prediction problem by accounting for contextual cues from the environment and the target agent. And:

Q3: is motion prediction solved?

This is discussed in Sec. 2.9.3 by revisiting the requirements from the different application scenarios. Finally, in the conclusions to this thesis (Chapter 7) we summarize the state of motion prediction, outline open challenges and suggest future research directions.

2.9.1 Benchmarking

Evaluating the performance of a motion prediction algorithm requires choosing appropriate testing scenarios and accuracy metrics, as well as studying the methods's robustness against various variables, such as the number of interacting agents or amount of maneuvering in the data.

Depending on the application area, the testing scenario may be an intersection, a highway, a pedestrian crossing, shared urban street with heterogeneous agents, a home environment or a crowded public space. Existing datasets, summarized in Sec. 2.8.2, cover a wide range of scenarios, e.g. indoor [48, 270, 366] and outdoor environments [181, 224, 236], pedestrian areas [28, 205], urban zones [262, 283] and highways [64, 65, 161], and include trajectories of various agents, such as people, cyclists and vehicles. However, these datasets are usually semi-automatically annotated and therefore only provide incomplete and noisy estimation of the ground truth positions (due to annotation artifacts). Furthermore, length of the trajectories is often not sufficient for evaluation in

some application domains, where long-term predictions are required. Moreover, the amount of interactions between recorded agents is often limited or disbalanced (very few agents are interacting, ergo misinterpreting such cases is not reflected in the lower benchmark scores). Finally, relevant semantic information about static (i.e. grass, crosswalks, sidewalks, streets) and dynamic (i.e. human attributes such as age, gender or group affiliation) entities is usually not recorded.

Accuracy metrics, described in Sec. 2.8.1, offer a rich choice for benchmarking, ranging from computing geometric distances between points (ADE, FDE) also accounting for temporal misalignments (DTW, MHD), to probabilistic policy likelihood measures (NLL) and sampling-based distribution evaluation (mADE). For long-term forecasts made in topologically non-trivial scenarios, results are usually multi-modal and associated with uncertainty. Performance evaluation of such methods should make use of metrics that account for this, such as negative log-likelihood or log-loss derived from the KLD. Not all authors are currently using such metrics. Even for short-term prediction horizons, for which a large majority of authors use geometric metrics only (AED, FDE), probabilistic metrics are preferable as they better reflect the stochastic nature of human motion and the uncertainties involved from imperfect sensing.

Another issue of benchmarking is related to variations in exact metric formulation and different names used for the same metric, e.g. for the ADE- and likelihood-based metrics, as indicated in Sec. 2.8.1. Additionally, precision is often evaluated on a single arbitrary prediction horizon. These aspects obstruct comparison of the relative precision of various methods.

Furthermore, very few authors currently address robustness as a relevant issue/topic. This is surprising as prediction needs to be robust against a variety of perturbations when deployed in real systems. Examples includes sensing and detection errors, tracking deficiencies, self-localization uncertainties or map changes.

On question 1:

We conclude that *Q1* is not confirmed. Despite the numerous metrics, datasets and experiment designs, used in individual works, benchmarking prediction algorithms lacks a systematic approach with common evaluation practices.

For evaluating prediction quality, researchers should opt for more complex testing scenarios (which include non-convex obstacles, long trajectories, collision avoidance maneuvers and non-trivial interactions) and the complete set of metrics (both geometric and probabilistic). It is a good practice to condition the forecast precision on various prediction horizons, observation periods and the complexity of the scene, e.g. defined by how many interacting agents are tracked simultaneously. Furthermore, perfect sensing, perception and tracking is not always achieved in real-life operation, and therefore algorithms' per-

formance ideally should be investigated in realistic conditions and supported by robustness experiments, e.g. see Sec. 2.8.1. Performing proper performance analysis would clarify application potential and effective prediction horizon of many methods.

Similar benchmarking practices should be applied to runtime evaluation. Considering efficiency on embedded CPUs of autonomous systems is important for the algorithm's design and evaluation. To prove applicability in real-life scenarios (e.g. in the pipeline with time-sensitive local and global motion planners), discussion should include formal complexity and runtime analysis, conditioned on the scene complexity and prediction horizon.

For a fair objective comparison of the prediction algorithms, developing a standard benchmark with testing scenarios and metrics is becoming a task of critical importance, e.g. given the rapid growth in published literature (see Fig. 2.2). The first attempt to build such a benchmark, TrajNet, is taken by Sadeghian et al. [276], with the follow up, TrajNet++ [160], actively being promoted in thematic workshops. TrajNet and TrajNet++ are based on selected trajectories from the ETH, UCY and Stanford Drone Dataset and uses the ADE and FDE evaluation metrics. We encourage more researchers to follow this example and contribute to the unification of benchmarking practices.

2.9.2 Modeling Approaches

With such a wide variety of motion modeling approaches, a natural question arises: which one should be preferred? In this section we discuss the inherent strengths and limitations of different approaches' classes and the efforts to incorporate various contextual cues. This discussion continues in Sec. 2.9.3 with highlighting the specifics of several key tasks in the application domains.

Physics-based approaches are suitable in those situations where the effect of other agents or the static environment, and the agent's motion dynamics can be modeled by an explicit transition function. Many of the physics-based approaches naturally handle joint predictions and group coherence. With the choice of an appropriate transition function, physics-based approaches can be readily applied across multiple environments, without the need for training datasets (some data for parameter estimation is useful, though). The downside of using explicitly designed motion models is that they might not capture well the complexity of the real world. The transition functions tend to lack information regarding the "greater picture", both on the spatial and the temporal scale, leading to solutions that represent local minima ("dead ends"). In practice, this limits the usability of physics-based methods to short prediction horizons and relatively obstacle-free environments. All in all, the existence of fast approximate inference, the applicability across multiple domains under mild conditions, and the interpretability make physics-based approaches a pop-

ular option for the collision avoidance of the mobile platforms (e.g. self-driving vehicles, service robots) and the people tracking applications.

Pattern-based approaches are suitable for environments with complex unknown dynamics (e.g. public areas with rich semantics), and can cope with comparatively large prediction horizons. However, this requires ample data that must be collected for training purposes in a particular type of location or scenario. One further issue is the generalization capability of such learned model, whether it can be transferred to a different site, especially if the map topology changes (cf. service robot in an office where the furniture has been moved). Pattern-based approaches tend to be used in non-safety critical applications, where explainability is less of an issue and where the environment is spatially constrained.

Planning-based approaches work well if goals, that the agents try to accomplish, can be explicitly defined and a map of the environment is available. In these cases, the planning-based approaches tend to generate better long-term predictions than the physics-based techniques and generalize to new environments better than the pattern-based approaches. In general, the runtime of planning-based approaches, based on classical planning algorithms (i.e. Dijkstra [285], Fast Marching Method [289], optimal sampling-based motion planners [132, 138], value iteration [195]) scales exponentially with the number of agents, the size of the environment and the prediction horizon [275].

On question 2:

In our view, *Q2* is not confirmed. As we have seen, the different modeling approaches have various strengths and weaknesses. Although in principle it could be possible to incorporate the same contextual cues, there have been so far insufficient studies to compare prediction performance across modeling approaches. Moreover, different modeling approaches exhibit varying degree of complexity and efficiency in including contextual cues from different categories. Physics-based methods are by their very nature aware of the target agent cues and may be easily extended with other ones (e.g. social-force-based [110] and circular distribution-based [66]). Pattern-based methods can potentially handle all kind of contextual information which is encoded in the collected datasets. Some of them are intrinsically map-aware [30, 165, 265]. Several others can be extended to include further types of contextual information (e.g. [4, 25, 242, 313, 328]) but such extension may lead to involved learning, data efficiency and generalization issues (e.g. for the clustering methods [30, 57]). Planning-based approaches are intrinsically map- and obstacle-aware, natural to extend with semantic cues [153, 257, 268, 370]. Usually they encode the contextual complexity into an objective/reward function, which may fail to properly incorporate dynamic cues (e.g. changing traffic lights). Therefore, authors have to design specific modifications to include dynamic cues into the predic-

tion algorithm (such as Jump Markov Processes in [142], local adaptations of the predicted trajectory in [267, 268], game-theoretic methods in [202]. Unlike for the pattern-based approaches, target agents cues are natural to incorporate, e.g. as in [167, 202, 267], as both forward and inverse planning approaches rely on a dynamical model of the agents. Contextual cues-dependent parameters of the planning-based methods (e.g. reward functions for inverse planning and models for forward planning) are trivial and typically easier to learn but inference-wise less efficient for high-dimensional (target) agent states compared to the simple physics-based models.

2.9.3 Application Domains

In Sec. 2.9.2 we have shown that all modeling approaches theoretically can handle various contextual cues. However, the question of preferring one approach over the others also depends on the task at hand.

Service robots

Predictors for mobile robots usually estimate the most likely future trajectory of each person in the vicinity of the robot. The usual setup includes cameras, range and depth sensors mounted on the robot, operating on a limited-performance mobile CPU.

Physics-based or pattern-based human interaction models, capable of providing short-term high-confidence predictions (i.e. for 1-2 seconds), are best suited for local motion planning and collision avoidance in the crowd. Methods used to this end should have fast and efficient inference for predicting short-term dynamics of several people around the robot. In the simplest case, even linear velocity projection is sufficient for smoothing the robot's local planning [19, 56]. More advanced methods should handle human-human interaction [4, 91, 105, 216, 236], the influence of robot's presence and actions on human motion [79, 226, 259, 281] and high-level body cues of human motion for disambiguating the immediate intention [109, 157, 250, 317]. In safety-critical applications, reachability-based methods provide a guarantee on local collision avoidance [23]. Furthermore, understanding local motion patterns is useful for compliant and unobstructive navigation [229, 329].

For global path and task planning, on the other hand, long-term multi-hypothesis predictions (i.e. for 15-20 seconds ahead) are desired, posing a considerably more challenging task for the prediction system. Reactivity requirement is relaxed, however understanding dynamic [32, 202] and static contextual cues [62, 66, 153, 300], which influence motion in the long-term perspective, reasoning on the map of the environment [142, 267] and inferring intentions of observed agents [35, 255, 324] becomes more important. For both local and global path planning, location-independent methods are best suited for predicting motion in a large variety of environments [23, 90, 293].

In terms of accuracy of the current state-of-the-art methods, experimental evaluations on simpler datasets, such as the ETH and UCY, show an average displacement error of 0.19 – 0.4 m for 4.8 s prediction horizon [4, 251, 328, 347]. Linear velocity projection in these scenarios is estimated at 0.53 m ADE. In more challenging scenarios of the ATC dataset with obstacles and longer trajectories an average error of 1.4 – 2 m for 9 s prediction has been reported [4, 268, 300].

Self-driving vehicles

The early recognition of maneuvers of road users in canonical traffic scenarios is the subject of much interest in the self-driving vehicles application. Several approaches stop short of motion trajectory prediction (i.e. regression) and consider the problem as action classification, while operating on short image sequences. Sensors are typically on-board the vehicle, although some work involves infrastructure-based sensing (e.g. stationary cameras or laser scanners) which can potentially avoid occlusions and provide more precise object localization.

Most works consider the scenario of the laterally crossing pedestrian, dealing with the question what the latter will do at the curbside: start walking, continue walking, or stop walking [143, 156, 157, 283]. Some works enlarge the pedestrian crossing scenario, by allowing some initial pedestrian movement along the boardwalk before crossing (Schneider and Gavrila [283] perform trajectory prediction, while other approaches are limited to crossing intention recognition, e.g. [85, 154, 282]). This scenario is safety-critical and crucial for autonomous vehicles to solve with high confidence. Pose and high-level contextual cues of the target agent [157], and the scene context modeling (e.g. location and type of the obstacles [217, 330], state of the traffic lights [142]) are helpful to improve the crossing trajectory prediction.

As to cyclists, Kooij et al. [157] consider the scenario of a cyclist moving in the same direction as the ego-vehicle, and possibly bending left into the path of the approaching vehicle. Pool et al. [246] consider the scenario of a cyclist nearing an intersection with up to five different subsequent road directions. Both involve trajectory prediction.

For predicting motion of both cyclists and vehicles is it important to consider multi-modality and uncertainty of the future motion. Recently many authors have proposed solutions to this end [53, 67, 117, 362]. Furthermore, it is important to consider coordination of actions between the vehicles [259, 281].

It is difficult to compare the experimental results, as the datasets are varying (different timings of same scenario, different sensors, different metrics). Several works report improvements vs. their baselines. For example, Fig. 2 in [156] shows that during pedestrian stopping, 0.9 and 1.1 m improvements in lateral position prediction can be reached with a context-based SLDS, compared to

a simpler context-free SLDS and basic LDS (Kalman Filter), respectively, for prediction horizons up to 1 s. A live vehicle demo of this system at the ECCV'14 conference in Zurich, showed that the superior prediction of the context-based SLDS could lead to evasive vehicle action being triggered up to 1 s earlier, than with the basic LDS.

Surveillance

The classification of goals and behaviors as well as the accurate prediction of human motion is of great importance for surveillance applications such as retail analytics or crowd control. Common setups for these applications use stationary sensors to monitor the environment. While single-frame based systems allow to partially solve some tasks such as perimeter protection, incorporating a sequence of observations and making use of behavior prediction models often improve accuracy in cases of occlusions or measurements with low quality (e.g. noise, bad lighting conditions).

Traffic monitoring and management applications can benefit from from long-term prediction models, as they allow to associate new observations with existing tracks (e.g. Luber et al. [198], Pellegrini et al. [236, 237], Yamaguchi et al. [347]) and to model long-term distributions over possible future positions of each person [63, 352]. Furthermore, it enables the analysis and control of customer flow in populated areas such as malls and airports, by gathering extensive information on human motion patterns [81, 149, 309, 354], understanding crowd movement in light and dense scenarios, tracking individuals within them, and making future predictions of individuals or crowds (e.g. crowd density prediction). Often these methods benefit from employing sociological methods, such as understanding of social interaction, behavior analysis, group and crowd mobility modeling [13, 31, 202, 367].

Furthermore identifying deviation from usual patterns often makes the foundation for anomaly detection methods that go beyond perimeter protection, as they analyze trajectories instead of the pure existence of a pedestrian in a specific region.

Also in this application area it is difficult to compare results obtained by different approaches, due to the diversity of the used datasets and the way the evaluation has been performed (e.g. different prediction horizons). In terms of prediction accuracy, we report the most interesting results obtained in densely crowded environments using mainly image data. In these settings, recent state-of-the-art approaches achieve an average displacement error of 0.08 – 1.2 m on the ETH, UC, NY Grand Central, Town Center and TrajNet datasets, and a final displacement error of 0.081 – 2.44 m, with a prediction horizon that generally goes from 0.8 s up to 4.8 s (Shi et al. [293], Xue et al. [344, 345, 346], Zhou et al. [367], the latter using a proprietary dataset and going up to a prediction horizon of 10 s).

On question 3:

As we show in Sec. 2.9, requirements to the motion prediction framework strongly depend on the application domain and particular use-case scenarios therein (e.g. vehicle merging vs. pedestrian crossing within the Intelligent Vehicles domain). Therefore, it is not possible to conclude achievement of absolute requirements of any sort. When considering concrete use-cases, industry-driven domains, such as intelligent vehicles (IV), appear to be the most mature in terms of formulated requirements and proposed solutions. For instance, requirements to the prediction horizon and metric accuracy for emergency braking of IV in urban driving scenarios are described in the ISO 15622:2018 [126] standard, which defines norms for comfortable acceleration/deceleration rates for vehicles, conditioned on the maximum speed and traffic rules, as well as the distribution of pedestrian speed and acceleration. Therefore we conclude, that for specific use-cases, in particular for basic emergency braking for IV, solutions have achieved a level of performance that allows for industrialization into consumer products. Those use-cases can be considered solved. For other use-cases we expect more standardization and explicit formulation of requirements to take place in the near future. For instance, the standard for safety requirements for personal care robots ISO 13482:2014 [125] suggests using sensors for detecting a human in the vicinity of the robot to issue a protective stop, and controlling the speed and force when the robot is in close proximity to humans to reduce the risk of collision. This standard, however, does not propose motion anticipation to improve the risk assessment.

Furthermore, several aspects of performance, robustness and generalization to new environments, discussed in the following sections, need to be explored before reaching further conclusions on maturity of the solutions. Finally, in order to reliably assess the quality of existing solutions across all application domains, is it critical to address the issues of benchmarking.

2.10 Conclusions and Outlook

In this chapter we present a thorough analysis of the human motion trajectory prediction problem. We survey the literature across multiple domains and propose a taxonomy of motion prediction techniques. Our taxonomy builds on the two fundamental aspects of the motion prediction problem: the model of motion and the input contextual cues. We review the relevant trajectory prediction tasks in several application areas, such as service robotics, self-driving vehicles and advance surveillance systems. Finally, we summarize and discuss the state of the art along the lines of three major questions and outlined several prospective directions of future research.

Prediction is very difficult, especially about the future.

This quote (whose origin has been attributed to multiple people) certainly remains applicable to motion trajectory prediction, despite three decades of research and the 200+ prediction methods listed in this chapter. Hopefully this thesis will increase visibility in this rapidly expanding field and the will stimulate further research along the directions discussed.

Following the findings and insights of this chapter, the thesis proceeds to develop on several topics in the upcoming chapters:

- Having discussed the benefits and drawbacks of the modeling approaches, a novel combination of the physics-based and planning-based methods into one powerful framework, which is both interaction- and obstacle-aware, is proposed in Chapter 3. This MDP-based method naturally supports available semantic maps, and the social force-based interaction module accounts for social grouping, thus incorporating several key contextual cues in one method. In its high level of context awareness, this method bridges a gap between the short-term motion prediction problem, where the dynamic environment cues are dominant, and long-term prediction, where semantics, obstacles and goals strongly influence the motion of people.

- Chapter 4 describes the contribution to the usage of semantic maps in autonomous systems. In this chapter we propose a data-driven method to learn occupancy priors of walking people and infer them for such environments, where no data is available, using only semantic map as input. This method allows assessing the semantically-rich environment regardless of the specific observed and tracked pedestrians, enabling an autonomous system to anticipate the dynamics and better plan the motion route in a distant area.

- Having shown that the available datasets of motion trajectories often lack important contextual cues, in Chapter 5 we propose a flexible, weakly-scripted data collection procedure to generate motion trajectories in interactive settings. Our new dataset THÖR contains one hour of diverse and accurate motion data in presence of obstacles and a moving robot, with groups of up to 6 people, and includes eye-gaze data and LiDAR recordings.

- Finally, to progress the topic of benchmarking, we propose a benchmark design in Chapter 6 which includes automated generation of testing scenarios, systematic variation of parameters and relevant experiments, rarely present in the literature, as we discussed in Sec. 2.8.1.

Chapter 3
Interaction-aware Planning-based Trajectory Prediction

If Croesus made war on the Persians, he would destroy a mighty empire.

ORACLE OF APOLLO AT DELPHI
gives Croesus a multimodal
prediction, 560 BC

Human motion is influenced by a variety of factors, with elements of the static and dynamic environment being the key ones. Considering the complex non-convex obstacles and heterogeneous social interactions between the observed agents is a challenging but necessary step for operation in cluttered and crowded environments. This chapter presents a planning-based approach for long-term human motion prediction that accounts for local interactions and can accurately predict joint motion of multiple agents. Long-term predictions are handled using an MDP formulation that computes a set of stochastic goal-directed motion policies. To obtain distributions over future motion trajectories, the policies are sampled with a weighted random walk algorithm in which each person is locally influenced by social forces from other nearby agents. The interaction model accounts for social grouping information, reflecting the soft formation constraints that groups typically impose on their members' motion. The presented algorithm produces multi-modal predictions with flexible non-parametric uncertainty representation, can account for individual agent velocities and requires no training phases. Qualitative and quantitative experiments demonstrate that the proposed method obtains more accurate predictions in

comparison to several state-of-the-art methods in terms of accuracy measured with probabilistic and geometrical metrics.

3.1 Introduction

A mobile robot is expected to operate in a variety of distinct environments alongside people. As such, the system should not fail when encountering crowds, obstacles or abnormal behavior. Predicting human motion is a key component for safe and reliable operation in such environments. This task is challenging due to the many factors that influence human motion: other agents with their intentions, actions, attributes or social rules, and the environment with its geometry, semantics or affordances. The prior art has addressed this task using different approaches based on physical dynamics modeling, learning and planning methods, considering both the single-agent case and the multi-agent case, in which predictions are made jointly.

As we reviewed in Sec. 2.7, it is largely established, that prediction should account for interactions between the observed people and their influence on each other's motion. A large selection of recent methods from all categories shows this capability to model interactions between people. Also obstacle-awareness is found more often in the presented methods. In fact, the idea to use MDP for motion prediction is not new. Ziebart et al. [370], Kitani et al. [153], Karasev et al. [142] and several others have formulated MDP-based predictors for individually observed people both indoors and outdoors. However, the combination of these two properties was not fundamentally explored in the prior art.

Considering social grouping information is a key feature for social-awareness of a prediction method, which is still relatively rarely found in the literature. Our attention to groups is motivated by the insight that social relations among people are an important factor for predicting future motion, as individuals in groups typically form and maintain certain spatial patterns, which e.g. in [216] is described by a model based on social communication between group members. An example of the group motion in real-world data is given in Fig. 3.2.

Research in computational social science and human crowd dynamics has found that up to 70% of people move in groups of two and more members and that they maintain rather stable formations depending on crowd density [58, 216]. These findings motive our hypothesis that social grouping is an important cue for the long-term prediction of human motion.

Related work in modeling group structure, crowd simulation or behavior analysis of pedestrian groups include [140, 248, 295] with applications e.g. for building design or mass event planning [21, 288]. Detecting groups has also applications in video surveillance [296] and tracking [176, 197, 237, 347] where group-informed motion modeling was shown to improve data association. The detection task has been addressed using clustering of geometrically similar tra-

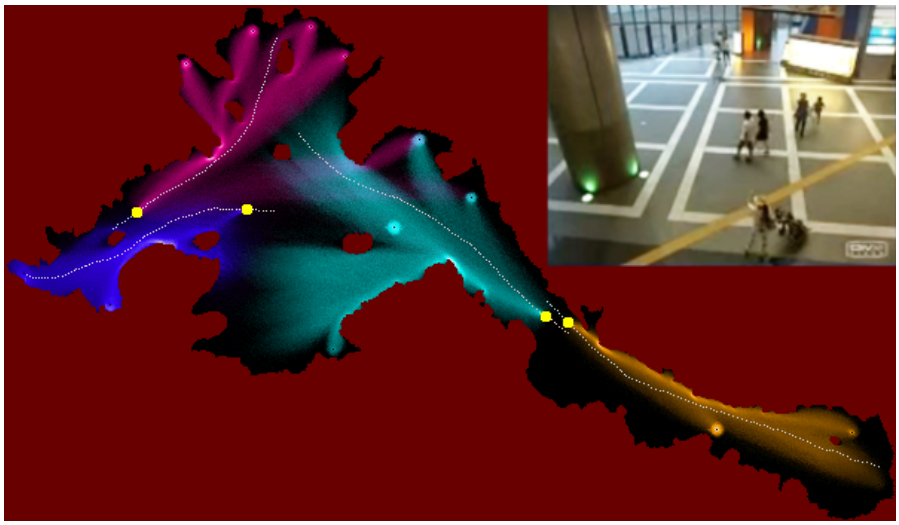


Figure 3.1: Prediction results for four persons in the ATC shopping center dataset, obtained with our algorithm and shown in individual colors. The current position of each person is indicated by a **yellow** circle, ground truth trajectories are shown in **white**.

jectories, estimating inter- and intra-group forces among individuals in crowds or so-called coherent motion indicators [216]. Common techniques for modeling group motion include imposing attraction forces to other group members [21, 237, 248], to the geometrical center of the group [216] or to the group's leader [288], or imposing a certain relative formation in which the group is assumed to be moving [140, 295]. An extension of the social force model that uses group information was proposed by Moussaïd et al. [216].

Social grouping was also incorporated in a prediction algorithm by [193, 197, 237, 347]. However, differently from the presented approach, those methods do not reason globally about goals and the environment topology, i.e. their predictions might lead into local minima. Furthermore, they are only used for short-term prediction within a tracking framework.

3.1.1 Contribution

In this chapter we present a novel planning-based approach that accounts for local social interactions to accurately predict the motion of multiple agents jointly and in real-time. Developing a planning-based method, we build on the assumption that humans essentially behave like planners by finding a (near-) optimal path through the environment. Our method extends the state-of-the-art by a novel MDP formulation of the joint motion prediction problem using

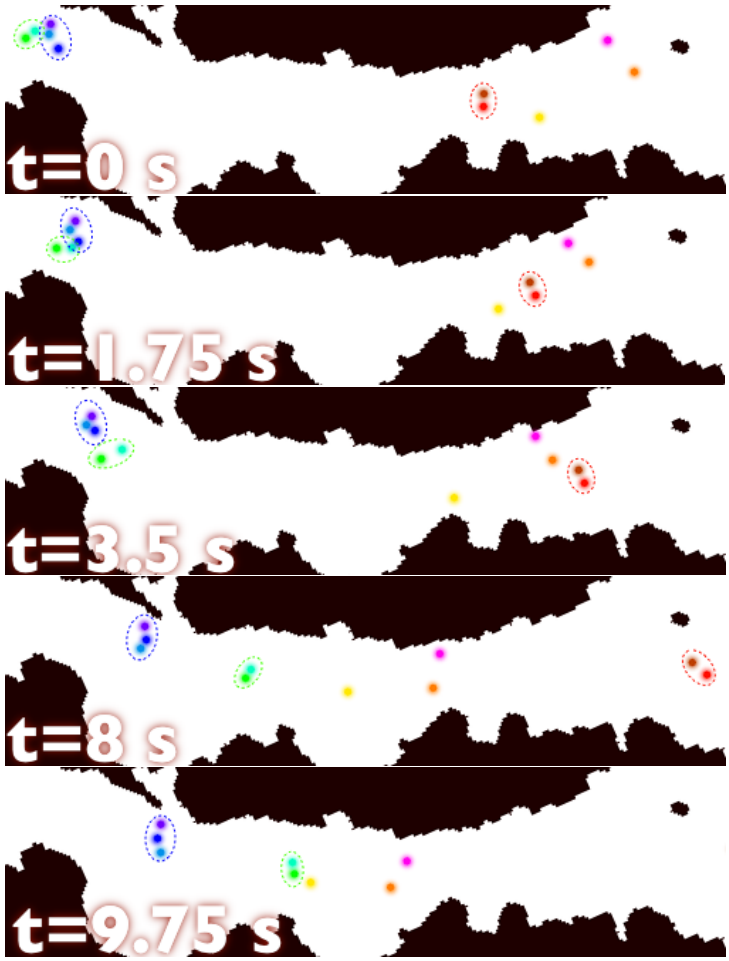


Figure 3.2: Illustrating example of group motion in a shopping mall. Colored points show ground truth positions of people from the ATC dataset at different points in time, groups are encircled. It can be seen, that people in groups stay together and move in relatively stable formations. Social grouping is thus an important cue for long-term motion prediction.

joint stochastic policy sampling to produce goal-directed global motion policies. For online prediction, those policies are locally modified based on social interactions between agents and soft formation constraints for agents in groups modeled by the group social force by Moussaïd et al. [216]. Experiments on real-world datasets show that our method can accurately predict long-term tra-

jectories of people involved in socially interactive tasks in real-time, outperforming four relevant state-of-the-art methods.

We also propose a method for performing predictive robot motion planning under the obtained policies.

3.1.2 Outline

The chapter is structured as follows: in Sec. 3.2 we describe our approach. Experiments and results are presented in Sec. 3.3, and Sec. 3.4 concludes the chapter with an outlook on possible extensions and applications of our method, both in ongoing and future work.

3.2 Joint sampling MDP for Motion Prediction

3.2.1 Problem Formulation

We frame the task of predicting a person's future location as estimating the probability $p(s|t)$ that the person will be in state s at time t, $\forall s \in \mathcal{S}$, $t_0 < t < T$, where t_0 is the current time and T is the prediction horizon. We use 2D grid maps of the discretized environment M to represent occupied and free space, as well as predicted positions of agents. Thus, $p(s|t = t_i)$ is a probability distribution over the prediction domain, estimated at time instance t_i. This is a powerful representation, which natively supports arbitrary multi-modal distributions, not limited to some parametric form or mixtures thereof.

Our method consists of two main components:

1. Global motion policy in the static environment is modeled using Markov Decision Processes (MDPs)

2. Local interaction between observed people is modeled using group social forces

For modeling the global motion policy we make the assumption, common to the planning-based prediction methods, that goal states are known a-priori or can be learned off- or online, and that the observed people intend to move towards those goals in an optimal or near-optimal fashion. Furthermore, we assume that the *static costmap* $C(s)$, which carries the unitary cost of each state, is known. The costmap is set to 1 for occupied states and to a small value $\epsilon > 0$ for free states. Furthermore, values in $(\epsilon, 1)$ can be used to represent unitary costs of walkable states, e.g. defined by a *semantic map*.

In the following we review the notation for Markov Decision Processes in Sec. 3.2.2, and describe the MDP formulation for global motion prediction in Sec. 3.2.3. Details in these two sections are sufficient for predicting the trajectories in isolation. We go on to present an interaction module in Sec. 3.2.4 and

combine it with the global motion policies in Sec. 3.2.5. Afterwards, Sec. 3.2.6 and 3.2.7 present complexity analysis and implementation details respectively.

3.2.2 Markov Decision Process Notation

Markov Decision Processes provide a mathematical framework for modeling decision making problems for a discrete-time stochastic control process. Formally, a MDP is described by a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$ where $\mathcal{S}$ and $\mathcal{A}$ are finite sets of agent *states* and *actions*, respectively. The *transition function* $\mathcal{P}(s, s', a)$ defines the probability of getting to state s' from state s when executing action a. The *reward function* $\mathcal{R}(s, a)$ specifies the immediate reward gained for taking action a in state s. The discount factor γ controls the importance of future rewards relative to immediate rewards. The agent's *policy* $\pi : \mathcal{S} \to \mathcal{A}$ defines the action the agent should take in each state. The *optimal policy* π^*, which maximizes the cumulative expected future rewards (Eq. 3.2), is obtained alongside with the state and action values, $V^*(s)$ and $Q^*(s, a)$, by solving the recursive Bellman equations (Eq. 3.1), using e.g. value iteration [303].

$$\begin{cases} Q^*(s, a) = \mathcal{R}(s, a) + \gamma \sum_{s'} \mathcal{P}(s, s', a) V^*(s') \\ V^*(s) = \max_a Q^*(s, a) \end{cases} \tag{3.1}$$

$$\pi^*(s) = \arg\max_a Q^*(s, a) \tag{3.2}$$

3.2.3 MDP for Global Motion Prediction

In this section we describe the model of global agent motion towards a goal. We use the MDP-based formulation of the optimal path planning problem in a known environment. Given a 2D static map M of the environment representing occupied and free space, and a set of goal states $\mathcal{G}$, we formulate a separate MDP path finding problem for each goal $g \in \mathcal{G}$ to obtain the cost-to-go state values $V^*_g(s)$, action values $Q^*_g(s, a)$ as well as the optimal policy $\pi^*_g(s)$ in each state $s = (s_x, s_y) \in M$. We assign negative rewards to all states and actions except for the goal state g which is represented as an *absorbing zero* state, i.e. has only one self-transitioning action with zero reward. Therefore the obtained policies are goal-directed: the agent will collect negative rewards until the goal state is reached.

For modeling the action space, we assume that human motion is unconstrained in orientation and acceleration. We describe actions as orientation-velocity pairs: $a = \langle \theta, v \rangle$, where $\theta \in [0, 2\pi)$ is the orientation and $v \in [0, v_{max}]$ is velocity. An action $a = \langle \theta, v \rangle$ defines the deterministic transition between states $s \xrightarrow{a} s'$, calculated as $s'_x = s_x + v\cos(\theta)$, $s'_y = s_y + v\sin(\theta)$, and reads as "making a move in direction θ with velocity v". This action space allows

transitions up to a large maximum velocity v_{max}. This action space, and the following definition of the reward function, is universal for all agents moving with their own preferred paces. Later we present a simple modification to reflect the velocity of a particular observed person.

The reward function $\mathcal{R}_g(s, a)$ is constructed as a weighted sum of Euclidean distance covered by a, and the unitary cost of the target state $C(s')$, provided by the optional input semantic map $C(s)$. Higher cost corresponds to states the agent is assumed to avoid, such as the road or unpaved regions.

$$\mathcal{R}_g(s, a) = \begin{cases} -w_1 C(s') - w_2 \|s - s'\|, \text{ if } s \neq g \\ 0, \text{ otherwise}, \end{cases} \tag{3.3}$$

where $s' = \mathcal{P}(s, a)$ and $w_1, w_2 > 0$ control the relative importance of each component: the unitary cost of s' and the Euclidean distance $\|\cdot\|$ covered with the action a. Since the reward function is negative everywhere except the goal state, we solve the MDP with $\gamma = 1$ in Eq. 3.1. Thus, the $V^*_g(s)$ value of a state is actually the *cost-to-go* from s to g.

To predict also alternative paths to the goal and allow deviations from the optimal policy, we relax the obtained π^*_g with the *stochastic Boltzmann policy* that assigns to each action a probability to be executed in state s proportional to its value $\hat{Q}^*_g(s, a)$. Temperature parameter α controls the level of stochasticity, i.e. the probability that sub-optimal actions are chosen by the agent. We denote the stochastic policy as π_g and compute it as in Eq. 3.5, where $\hat{Q}^*_g(s, a)$ is the value of action a, and $V^*_g(s)$ is the value of the optimal action.

$$\hat{Q}^*_g(s, a) = w_a \mathcal{R}_g(s, a) + V^*_g(s') \tag{3.4}$$

$$a \sim \pi_g(s) \text{ with prob. } \propto \exp(\alpha(\hat{Q}^*_g(s, a) - V^*_g(s))) \tag{3.5}$$

Here an additional weight $w_a \in (0, 1)$ in Eq. 3.4 is introduced to encourage the agent to perform faster actions with larger v. This modification is necessary for adapting the policies to the actual observed speed, described in the following.

The obtained policy π_g allows actions up to a pre-defined very large velocity v_{max}. For handling individual observed velocities $v_{obs} < v_{max}$, we use a simple *policy cutting* technique that incorporates information about v_{obs} into the obtained policy. For each person i, the action space is redefined with $v \in [0, 2\, v^i_{obs}]$. The individual stochastic policy $\hat{\pi}^i_g$ is then computed as in Eq. 3.6. In $\hat{\pi}^i_g$ the probability of faster actions $a = \langle \theta, v \rangle$ with $v > v^i_{obs}$ is set the same as for the symmetrically slower actions with $v < v^i_{obs}$.

$$p(a) \quad \text{in} \quad \hat{\pi}^i_g \qquad \propto \qquad \begin{cases} p(\langle \theta, v \rangle) \text{ in } \pi_g, \text{ if } v \leqslant v^i_{obs}, \\ p(\langle \theta, 2\, v^i_{obs} - v \rangle) \text{ in } \pi_g, \text{ if } v > v^i_{obs} \end{cases} \tag{3.6}$$

Basically, we assign the same probability to faster actions with $v > v^i_{obs}$ as to the symmetrically slower actions with $v < v^i_{obs}$. The original policy π_g is "cut" at the point of v^i_{obs} and "mirrored" backwards, hence the name *policy cutting*.

3.2.4 Joint Human Motion Prediction with Group Social Forces

In this subsection we present our method for jointly predicting trajectories of all agents in the scene. We assume that a person tracking system delivers short sequences of observed agent positions, called tracklets, and that this system also provides group detection as partitionings of individual agents into groups. These are both realistic assumptions as many tracking systems, for example [194], are able to robustly track people also across misdetection and occlusions using e.g. advanced data association techniques. Such systems have also been extended with the ability to detect and reason about social grouping hypotheses as discussed in Sec. 3.1.

Given N people in the scene, the observed track of length $l(i)$, associated with person i, is denoted as $\mathcal{T}^i = \langle s^i_1, s^i_2, ..., s^i_{l(i)} \rangle$, where $s^i_t = (s^i_{x,t}, s^i_{y,t})$ is the state where the person was observed at time t, and $i \in [1, \dots, N]$. The tracklet's end $s^i_{l(i)} = s^i(t_0)$ is the position of person i at the current time t_0 and $\mathcal{T}$ is the set of all observed tracks. Membership in one and only one of the groups $Gr_h \in \mathbf{Gr}$ is assigned to each person: $i \in Gr_h,\ Gr_h \cap Gr_{h'} = \emptyset\ \forall h' \neq h,\ \cup_h Gr_h = \{1, \dots, N\}$.

From each tracklet we derive the observed speed v^i_{obs}, orientation θ^i_{obs} and the discrete probability distribution $p^i(\mathcal{G})$ over destinations $\mathcal{G}$. We predict the final destination of person i based on the observed tracklet. Similarly to [370] and [324], for each goal $g \in \mathcal{G}$ we estimate the gradient of the cost-to-go $V^*_g(s)$ along $\mathcal{T}^i$ as the difference between the costs at s^i_1 and $s^i_{l(i)}$ using a softmax function:

$$p(g) \ \propto\ \exp\left(\beta\left(V^*_g(s^i_{l(i)}) - V^*_g(s^i_1)\right)\right). \tag{3.7}$$

Temperature parameter β defines to what extent alternative goals are considered. Members of the same group Gr_h share the goal probability vector, computed as the average of individual vectors: $p^h_{Gr}(\mathcal{G}) = |Gr_h|^{-1} \sum_i p^i(\mathcal{G}),\ i \in Gr_h$.

Local interaction modeling with social forces

The concept of social forces [110] describes how the intended motion of a person changes according to the influence of other people and the environment by superimposing repulsive forces from obstacles and other people with attractive forces to the goal. The approach, initially developed for crowd behavior analysis and egress research, performs well in modeling short-term local influences but performs poorly in making accurate long-term predictions, as we have seen

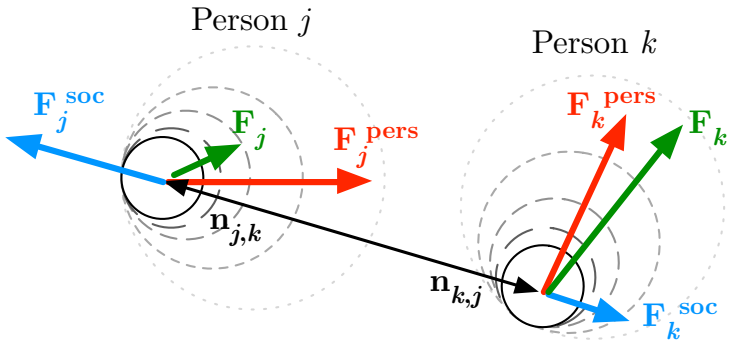


Figure 3.3: Example of the anisotropic social force model with $\lambda = 0$. Person k is crossing in front of person j. Intended directions $\mathbf{F}^{\text{pers}}$ are shown with **red** arrows. Person j is influenced by a strong social force $\mathbf{F}^{\text{soc}}$, depicted in **blue**, while the effect on person k is limited due to the anisotropic factor. Resulting directions $\mathbf{F}$ are shown in **green**.

in Chapter 2. Therefore, in our method the long-term aspects – attraction to the goal and repulsion from obstacles – are handled by the MDP formulation. We only utilize the local influence aspects of the social force model.

Formally, social force $\mathbf{f}_{i,k}^{\text{soc}}$, emitted by person k in the direction of person i is

$$\mathbf{f}_{i,k}^{\text{soc}} = a_k e^{\left(\frac{r_{i,k} - d_{i,k}}{b_k}\right)} \mathbf{n}_{i,k} \left(\lambda + (1-\lambda)\frac{1+\cos(\varphi_{i,k})}{2}\right), \tag{3.8}$$

where $a_k \geqslant 0$ specifies the magnitude and $b_k > 0$ the range of the force, $d_{i,k}$ is the distance between people and $r_{i,k}$ is the sum of their radii. The term $\mathbf{n}_{i,k}$ is the normalized vector pointing from k to i, which describes the direction of the force. An anisotropic factor, controlled by $\lambda \in [0,1]$, scales the force in the person's direction of motion: the force reaches its full magnitude when the angle $\varphi_{i,k}$ between the intended motion direction of person i and $\mathbf{n}_{k,i}$ is zero, and has no effect when $\varphi_{i,k} = \pi$. The factor postulates that influences in the front of a person are stronger than those to the sides and weak in the back (see also Fig. 3.3). Social forces on person i are added for all k and used to change the motion direction $\mathbf{F}_i^{\text{pers}}$ which in our case is the action $a = \langle \theta, v \rangle$ sampled from the stochastic policy:

$$\mathbf{F}_i = \mathbf{F}_i^{\text{pers}} + \mathbf{F}_i^{\text{soc}} = \mathbf{F}_i^{\text{pers}} + \sum_{k \neq i} \mathbf{f}_{i,k}^{\text{soc}}. \tag{3.9}$$

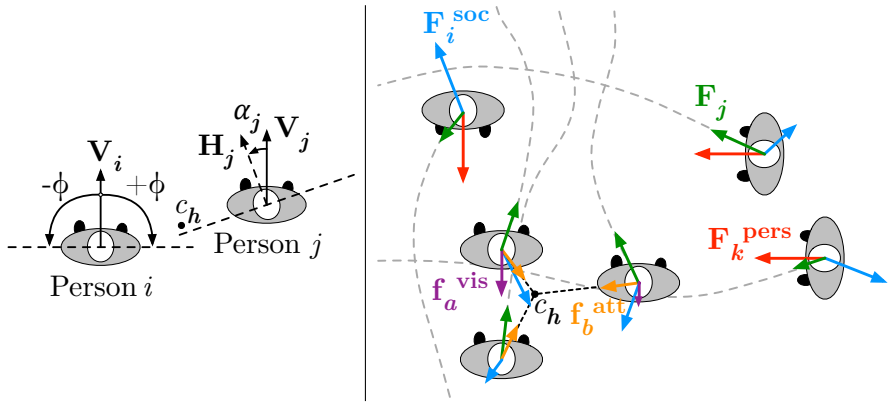


Figure 3.4: Left: illustration of the group social force parameters. **Right:** in this example, three people a, b and c in the bottom are walking upwards as a group $\mathrm{Gr_h}$. Three individual pedestrians i, j and k are opposing them from different directions. Intended directions $\mathbf{F}^{\text{pers}}$ are shown with **red** arrows, omitted for the group members for the sake of clarity. Person i is influenced by a strong social force $\mathbf{F}^{\text{soc}}$, depicted in **blue**, and has to halt and adjust the motion trajectory, shown as a **gray dotted line**. Person k stops and lets the group pass, while j attempts to cross in front of the group. Resulting motion directions **F** are shown in **green**. Intra-group social forces $\mathbf{F}^{\text{vis}}$ and $\mathbf{F}^{\text{att}}$ are shown in **blue** and **orange** respectively.

An extension to group social forces

An extension of the social force model to include group interaction was proposed by Moussaïd et al. [216]. Several new forces are defining attraction of people walking in groups to other members of the group (attraction term) and imposing soft constraints on the walking formation that resembles typical patterns of humans in groups (visibility term). For each member i of the group $\mathrm{Gr_h}$, the visibility term $\mathbf{f}_\mathrm{i}^{\text{vis}}$ is defined as

$$\mathbf{f}_\mathrm{i}^{\text{vis}} = -\beta_1 \alpha_\mathrm{i} \mathbf{V}_\mathrm{i}, \tag{3.10}$$

where β_1 is a model parameter describing the strength of the social interaction between group members, and $\mathbf{V}_\mathrm{i}$ is the current velocity vector of person i. This deceleration component $\mathbf{f}_\mathrm{i}^{\text{vis}}$ is oriented in the opposite direction of current movement $\mathbf{V}_\mathrm{i}$, and it is proportional to the angle α_i between the gazing direction $\mathbf{H}_\mathrm{i}$ of person i and the group center of mass $\mathrm{c_h}$, given the person's field of view ϕ. An illustration of the parameters is given in Fig. 3.4, left.

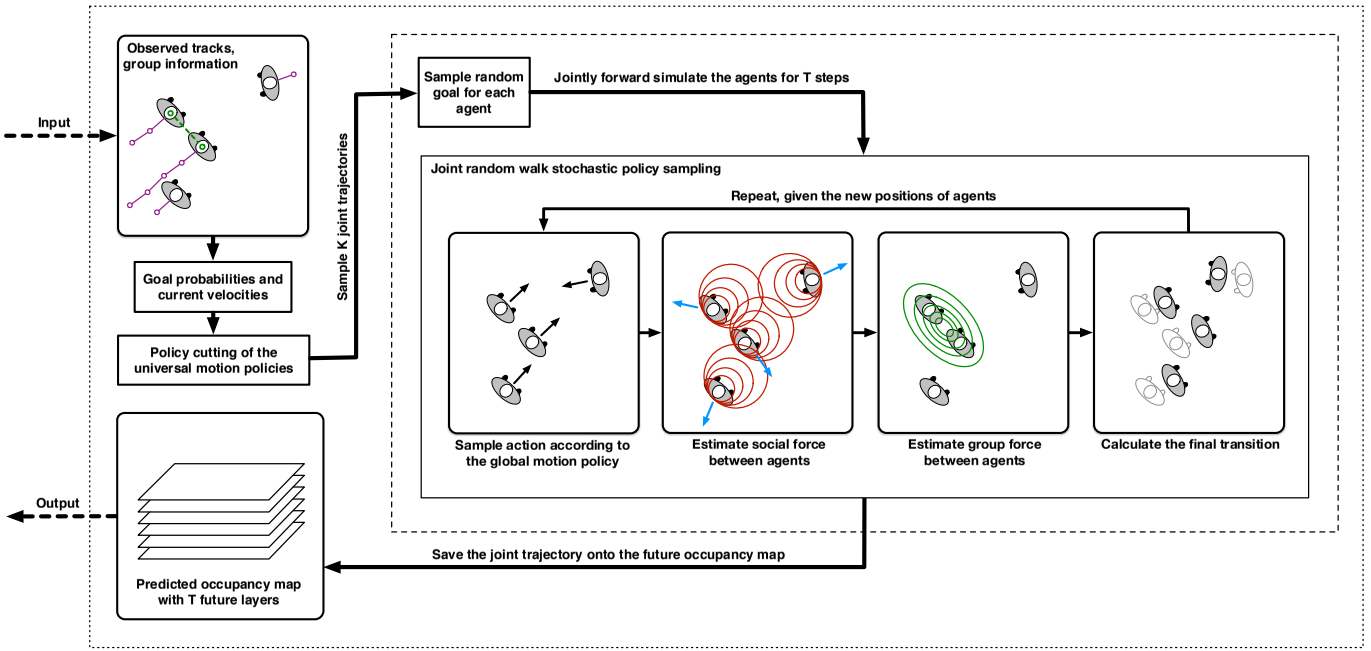


Figure 3.5: Summary and flow diagram of the approach. Based on the observed tracklets and group detections, our prediction method samples K joint trajectories, each time drawing a random goal for the people and groups. The random walk stochastic policy sampling procedure estimates the social interactions and forward-simulates the agent positions T steps ahead. Joint trajectories are then saved onto the future occupancy map.

Formulation of $\mathbf{f}_i^{\text{vis}}$ imposes a *line formation*, perpendicular to the direction of motion, as the preferred walking pattern of a group. However, in order to facilitate intra-group social interactions, members of larger groups of 4 or more people often switch to the more compact *V-formation*. The same happens in cluttered spaces, as well as in crowded environments, where the members have to balance between comfortable interaction and efficient movement. To model this behavior, the attraction term $\mathbf{f}_i^{\text{att}}$ to the geometrical center of the group is introduced as

$$\mathbf{f}_i^{\text{att}} = \beta_2 q_A \mathbf{U}_i, \tag{3.11}$$

where β_2 is the strength of the group attraction effect, and $\mathbf{U}_i$ is the unit vector pointing from pedestrian i to the center of masses c_h of Gr_h. This force is only activated if the distance between person i and c_h exceeds a certain threshold q_A, otherwise the attraction force is zero.

The added intra-group forces $\mathbf{f}_i^{\text{vis}}$ and $\mathbf{f}_i^{\text{att}}$ yield a decelerating effect on pedestrians, whose stochastic motions often lead them in front of the group. In reality this effect is not present as humans by nature are able to better coordinate their motion within the group. To counterbalance the deceleration effect and get more precise predictions on average, we simply scale the observed speed v_{obs}^i of each human i by a factor $q_S > 1$.

The final direction of motion for person i is computed as

$$\mathbf{F}_i = \mathbf{F}_i^{\text{pers}} + \mathbf{F}_i^{\text{soc}} + \mathbf{F}_i^{\text{group}} = \mathbf{F}_i^{\text{pers}} + \sum_{k \neq i}^{N} \mathbf{f}_{i,k}^{\text{soc}} + \mathbf{f}_i^{\text{vis}} + \mathbf{f}_i^{\text{att}}. \tag{3.12}$$

An example of the social forces affecting the motion of people in a social scenario is given in Fig. 3.4, right.

3.2.5 Stochastic Policy Sampling Using Random Walks

To generate predictions using the stochastic policy π_g, we propose a random walk algorithm (Alg. 1) that samples K joint paths for all people in the scene, each path starting in the corresponding current state $s_{l(i)}^i$ of person i at time t_0. Each joint path is representing a possible future interaction given the observed tracklets and available group information. In each of the K samples we randomly draw a goal $g(i)$ for person i from the distribution $p^i(\mathcal{G})$ and randomly generate actions $a^i = (\theta^i, v^i)$ from the policy corresponding to $g(i)$. This is done by sampling the normalized discrete distribution $\hat{\pi}_{g(i)}^i(s_n^i)$ obtained from Eq. 3.5. Group members share the same goal, sampled from $p_{Gr}^h(\mathcal{G})$. During this random walk, we evaluate social interactions among the agents that affect each agent's instantaneous stochastic policy according to the group social force model (see Fig. 3.6 for illustration). The position of each person at time t is then saved in the corresponding layer L_t^i of the probabilistic occupancy map L, that

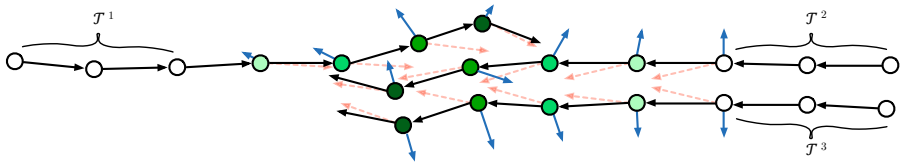


Figure 3.6: Social interaction example with our random walk algorithm using social forces. There are two persons side-by-side and a third person moving in the opposite direction. Observed tracklets $\mathcal{T}^{1,2,3}$ are shown by **white** circles, jointly predicted future positions are shown by the same **shade of green**. **Red** arrows indicate the sampled action of each person, **blue** arrows show the social force and **black** arrows show the resulting action the person executes.

is shared among the K samples. After K iterations L is normalized $\sum_s L_t^i(s) = 1$ to properly represent the probability distribution $p(s|t = t_i)$ of the person's possible location at time t_i (see also Fig. 3.7).

To achieve smoother path predictions, we introduce a model-free motion inertia term, parametrized by $I = (I_\theta, I_\nu)$, that prevents sudden changes in speed and direction between $t-1$ and t. When sampling the stochastic policy π_g, we obtain action $a = (\theta_t, \nu_t)$ which is then shifted towards the current speed and direction as in Eq. 3.13. By varying I_θ and I_ν, together with the temperature α, we get a flexible control of angular and translational variability in a person's goal-directed motion behavior.

$$(\theta_t, \nu_t) := (1 - I) \cdot (\theta_t, \nu_t)^T + I \cdot (\theta_{t-1}, \nu_{t-1})^T \tag{3.13}$$

Finally, we have Alg. 2 as our method for prediction, with an illustrative summary in Fig. 3.5. Its inputs are the obstacles map M of the environment, the set of goals $\mathcal{G}$, the set of tracklets $\mathcal{T}$, the social grouping information **Gr**, and the prediction horizon T. Its parameters are the cost of free space ϵ, motion stochasticity α and goal uncertainty β, motion inertia coefficients I_ν and I_θ, social force parameters $SF_p = (a_k, b_k, \lambda)$, group social force parameters $GSF_p = (\beta_1, \beta_2, q_A, \phi, q_S)$ and the value of K samples from the stochastic policy. We keep the inertia and social force parameters constant for all people, however, online estimation of their values for individually observed persons is possible in future work. Lines 3-6 prepare and solve the MDP. This part of the algorithm can be precomputed offline or updated online at lower frequency since the stochastic policy remains valid as long as the map stays relatively static. Line 7 calls the joint stochastic policy sampling method that computes predictions for all people and returns the occupancy map L. See Fig. 3.1 for example predictions obtained with our method.

Algorithm 1 Joint Random Walk Stochastic Policy Sampling

function JointStochPolicySampling($\mathfrak{T}$, **Gr**, $V^*_g(s)$, $\pi_g(s)$, K, T)
Compute θ_{obs}, v_{obs}, $p(\mathcal{G})$ for each person using $\mathfrak{T}$ and $V^*_g(s)$
Initialize $T \times N$ empty layers of the L occupancy map:
for $t = 1, \ldots, T$, $i = 1, \ldots, N$ **do**
    $L^i_t \leftarrow \texttt{zeros}(|\mathcal{S}|)$
end for
Sample K joint paths for all people:
for $k = 1, \ldots, K$ **do**
    For person i set initial state s^i_n, orientation θ^i_n and velocity v^i_n,
    and sample the goal $g(i)$:
    for $i = 1, \ldots, N$ **do**
        $(s^i_n,\ \theta^i_n,\ v^i_n) \leftarrow (s^i_{l(i)},\ \theta^i_{obs},\ v^i_{obs})$
        $g(i) \leftarrow \texttt{sample}(p^i(\mathcal{G}))$
    end for
    Jointly predict for T steps ahead:
    for $t = 1, \ldots, T$ **do**
        for $i = 1, \ldots, N$ **do**
            repeat
                Sample random action a of person i according to $\hat{\pi}^i_{g(i)}$:
                $(\theta^i_a,\ v^i_a) \leftarrow \texttt{sample}(\hat{\pi}^i_{g(i)}(s^i_n))$
                Apply inertia given current orientation θ^i_n and velocity v^i_n:
                $(\theta^i_a,\ v^i_a) \leftarrow (1 - I) \cdot (\theta^i_a,\ v^i_a) + I \cdot (\theta^i_n,\ v^i_n)$
                Calculate state transition of person i executing action a:
                $s^i_{n+1} \leftarrow \mathcal{P}(s^i_n,\ a)$
                Social force on person i given current agents' positions s_n:
                $F_s \leftarrow \texttt{socialForce}(s_n, i)$
                Modify transition of person i given the current social force:
                $s^i_{n+1} \leftarrow s^i_{n+1} + F_s$
            until $\texttt{lineOfSight}(s^i_n,\ s^i_{n+1})$
            Add the next position s^i_{n+1} of person i to the occupancy map:
            $L^i_t(s^i_{n+1}) \leftarrow L^i_t(s^i_{n+1}) + 1$
        end for
        Update current positions, orientations and velocities of all people:
        for $i = 1, \ldots, N$ **do**
            $(s^i_n,\ \theta^i_n,\ v^i_n) \leftarrow (s^i_{n+1},\ \theta^i_a,\ v^i_a)$
        end for
    end for
end for
for $t = 1, \ldots, T$, $i = 1, \ldots, N$ **do**
    $L^i_t \leftarrow \texttt{normalize}(L^i_t)$
end for
return L

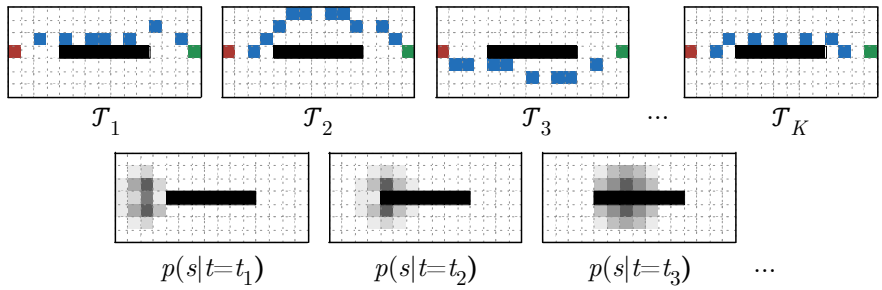


Figure 3.7: Illustration of the random walk stochastic policy sampling in a scenario with one person. **Top row:** For each of the K sampled paths $\mathcal{T}_1 ... \mathcal{T}_K$, locations at time t_j are added to the corresponding time layer $p(s|t = t_j)$, which is then properly normalized. The path is depicted in **blue**, with **red** being the start s_0 and **green** the goal. **Bottom row:** Probability distributions $p(s|t = t_j)$ are shown by **shades of gray**, darker areas mean higher probabilities.

Algorithm 2 Group Social Force MDP Motion Prediction

1: Parameters: $\alpha, \beta, K, \epsilon, I, SF_p, GSF_p$
2: Inputs: $M, \mathcal{T}, T, \mathcal{G}$, **Gr**
3: **for all** $g \in \mathcal{G}$ **do**
4: compute $\mathcal{R}_g(s, a)$ as in Eq. 3.3
5: $V_g^*,\ Q_g^*,\ \pi_g^* \leftarrow$ `ValueIteration`$(\mathcal{R}_g(s, a))$ as in Eq. 3.1, 3.2
6: compute π_g as in Eq. 3.5
7: **end for**
8: $L \leftarrow$ `JointStochPolicySampling`$(\mathcal{T}$, **Gr**$, V_g^*(s), \pi_g(s), K, T)$
9: **return** L

3.2.6 Complexity Analysis

Alg. 3 summarizes the operations required to obtain predictions with our algorithm. We assume that K joint random paths are requested, N people are in the scene and T prediction steps are made. The complexity of the goal sampling operation for every human (line 2) depends on the number of goals $|\mathcal{G}|$. Group center calculation is done only once for each time step (line 4). The random action sampling procedure (line 6) depends on the action space discretization (A angles and V velocities) and has the worst-case complexity of $\mathcal{O}(AV)$. This happens when the agent is moving with velocity close to v_{max}. The social force in the direction of agent i (line 7) is computed for each surrounding agent within a certain radius. In the worst-case, when all agents are densely located, the complexity is $\mathcal{O}(N)$. The group social force computation (line 8) is a constant time operation.

Algorithm 3 Joint Stochastic Policy Sampling Operation Analysis

```
for k = 1, ..., K do
    Sample a goal for each person: O(N|G|)
    for t = 1, ..., T do
        Calculate group center for each group: O(N)
        for i = 1, ..., N do
            Sample a random action: O(AV)
            Calculate social force: O(N)
            Calculate group social force: O(1)
        end for
    end for
end for
```

The overall complexity of our prediction algorithm is then $\mathcal{O}(K(N|\mathcal{G}| + T(N(AV + N))))$. Runtime measurements with comparison to the considered baselines are given in Sections 3.3.1 and 3.3.2.

3.2.7 Implementation Details

When solving the MDP problem for a goal g, we need to make sure that g is reachable from every free state, i.e. there are no isolated states which may prevent convergence of the value iteration algorithm in the absorbing zero setting. We use a wavefront algorithm starting from the goal state to determine the subset of approachable states, and solve the MDP problem only for those states. To speed-up convergence, we process the states in the order of increasing Manhattan distance to the goal. Moreover, since in our MDP $\forall s, a, a' : \mathcal{P}(s, a) = \mathcal{P}(s, a') = s' \Rightarrow Q^*(s, a) = Q^*(s, a')$, i.e. actions a and a' have the same effect, we iterate directly over target states s' instead of every pair $\langle \theta, v \rangle \in \mathcal{A}$. Starting with undefined state values, we run value iteration until all states are assigned with some positive value, thus obtaining approximate costs-to-go. In our experiments, value iteration converges to the approximate cost-to-go after only one iteration.

For a fine discretization of action space $\mathcal{A} = \langle \theta, v \rangle$, storing the stochastic policy $\pi_g(s)$ for every state implies significant storage burden. We store the policy in a sparse form, saving only actions with probability higher than the factor of $\frac{1}{|\mathcal{A}|}$. This yields no visible effect on the random walk predictions, but saves up to 10x storage space, depending on the level of stochasticity in the original policy. To reduce the number of samples K, we smooth the layers of L with three iterations of a separable box filter. We found that this leads to very similar distributions compared to those obtained with 10 to 20 times more samples.

3.3 Experiments

In this section we evaluate our method both qualitatively and quantitatively in a series of experiments and compare it to several baselines. In the first round of experiments, presented in Sec. 3.3.1, we use scenarios where no group motion is present. Next, in Sec. 3.3.2 we study how considering social grouping context improves the quality of prediction. By doing this, we show that already considering social interactions without groups improves the prediction results over the typical MDP-based methods, which are not interaction-aware. On the other hand, using global context, provided by the MDP component, gives our solution an edge over the pure social force-based methods.

In both rounds of experiments we use two types of environments. Firstly, we use several simulated scenarios to demonstrate the predicted trajectories. Then, we qualitatively evaluate the methods using the ATC dataset[1] of real-world trajectories recorded in a shopping center. The map of the environment, covering an area of 900 m^2, is shown in Fig. 3.1. Using the large selection of trajectories, we identify 15 common goal points in the area with trajectory endpoint clustering. From the dataset we select 25 scenarios without groups and 21 scenarios with groups, each having several interacting pedestrians (i.e. between 2 and 10). In each scenario, people are following various paths to their intended destinations with different velocities, adjusting paths to comply with other agents nearby. The presence of high-level motion stochasticity, observation noise and close proximity to other people makes this dataset a challenging one, especially for longer prediction horizons.

Since our method combines a planning-based and a social force-based prediction approach, we choose as baselines the planning-based approach by Karasev et al. [142] and the social force-based approach by Elfring et al. [80]. For the sake of a fair comparison, we apply our own goal estimation technique (that requires no learning data, see Eq. 3.7) to the baselines.

To stress the performance of the various components of our method, in the experiments we refer to it as:

- IS-MDP – our first motion prediction method, published in [266], which stands for *Independent* or *Individual Sampling MDP*. This method does not consider any interactions between people, essentially predicting trajectories as if people were moving in isolation.
- JS-MDP, published in [267], stands for *Joint Sampling MDP*. This method uses the social forces, but does not consider social grouping.
- GSF-MDP has the full functionality of the *Group Social Force MDP*, presented in [268] and described in this chapter.

[1] http://www.irc.atr.jp/crest2010_HRI/ATC_dataset/

We evaluate the predictive performance of all algorithms using the following metrics: *Negative Log-Probability* (NLP) is a direct measure of ground truth path $\mathcal{T}$ probability, measured at each point of path $\mathcal{T}_i$ according to predictions for that time instance t_i: $\text{NLP}(\mathcal{T}) = -\frac{1}{T}\sum_{i=1}^{T} \log p(\mathcal{T}_i|t_i)$. *Modified Hausdorff Distance* (MHD) [153] is a geometric measure of distance between the ground truth path and the most probable path in the predicted probability distribution. For both metrics, lower values correspond to better prediction accuracy or smaller geometric deviation, respectively.

All algorithms are implemented in C++, running on a laptop with a 2.8 GHz Intel Xeon processor and 32 GB RAM. The following sections provide further technical details of the experiments, such as parameter values and the numbers of random trials.

3.3.1 Environments with no Groups

In this section we evaluate our *Joint Sampling MDP* (JS-MDP) approach both qualitatively and quantitatively and compare it to the baselines.

Experiment 1: Predicting social interactions

The first experiment aims to qualitatively evaluate the local predictive ability of JS-MDP in predicting future trajectories of humans involved in cooperate collision avoidance. We simulated four scenarios: two people walking together side-by-side, one person overtaking another person, people walking in opposite flows, and a situation in which a person blocks a narrow passage (see Fig. 3.8 top row, Fig. 3.9 top row).

Experiment 2: Prediction accuracy evaluation

In this experiment we quantitatively evaluate the predictive performance of JS-MDP using the 25 scenarios without groups from the ATC dataset. MHD and NLP metric values are calculated for each trajectory in these scenarios and averaged across 50 experiments for each scenario. We use 2 seconds as observation period, and predictions are obtained for $T = [2.5, \ldots, 15]$ seconds ahead. We also give the average times to compute predictions using our algorithm.

For each algorithm we perform hyperparameter optimization using the SMAC3 toolbox [192] with results, summarized in Table 3.1. Values of w_1, w_2 and w_a in Eq. 3.4 are estimated to match the expected behavior of the pedestrian to the best of our knowledge: $w_1 = 1,\ w_2 = 1,\ w_a = 0.5$ and the cost of the free space is $\epsilon = 10^{-10}$. Action space parameters are set as follows: angular discretization of θ is $\pi/20$; translational discretization of v is 0.1 m/s, $v \in [0, 3]$ m/s. Cell sizes of our grid maps are 0.1 m in Experiment 1 and 0.15 m in Experiment 2. The frequency of predictions is 4 Hz, the number of random walk samples $K = 100$.

Method	Parameters
IS-MDP	$\alpha = 15.95,\ \beta = 5.44$
JS-MDP	$\alpha = 5.03,\ \beta = 13,\ I = (0.687, 0.725),\ (a_k, b_k, \lambda) = (0.271, 0.221, 0)$
Karasev et al. [142]	$(w_{g,t}, w_{s,t}) = (0.031,\ 0.140),\ \alpha = 21.31,\ \beta = 18.68$
Elfring et al. [80]	$(q_w, f_w, c_w) = (1.436,\ 0.23,\ 3.097),\ \zeta_\rho = 83.74$

Table 3.1: Estimated hyperparameters in experiments with no groups

Results

Fig. 3.8 and Fig. 3.9 show the qualitative results of the first experiment. Our method correctly predicts the development of each scenario, handling typical cooperative actions that people carry out in social spaces: the approach is able to predict overtaking and avoidance maneuvers (see Fig. 3.9), and to infer usual social interactions such as walking side-by-side or offering the way to a pedestrian moving in opposite direction (Fig. 3.8), without discarding the goal intentionality of the pedestrians.

Fig. 3.10 presents the results of the quantitative evaluation, conducted in Exp. 2, showing the mean of the NLP and MHD metrics over the prediction horizon of 2.5–15 sec. The NLP results show that our method outperforms the other three approaches, assigning a higher probability to the future ground truth location of people, which is essential e.g. for predictive motion planning as in Sec. 3.4.3. The two planning-based methods [142, 266] accumulate errors with growing prediction horizon from non-predicted social interactions, while the social force-based method of Elfring et al. [80] gives worse results as its predictions do not account for the global environment structure. The results for the probabilistic MHD metric show that our approach is on par with the others methods for short prediction horizons but outperforms them for the more relevant longer horizons.

We also give the runtime results for JS-MDP in Fig. 3.11. Our approach takes on average 0.4 seconds to predict for $T = 7.5$ seconds ahead in a scenario with 5 people. Short-term predictions for $T = 2.5$ seconds can be quickly obtained in less than 0.2 seconds. Note that these measurements exclude time for Value Iteration (Alg. 2 line 5) and the stochastic policy computation (Alg. 2 line 6), which take 0.4 s and 2.35 s for each goal respectively, but can be computed offline for a known environment or updated in a low-frequency cycle.

3.3.2 Experiments with Groups

In this section we present several experiments conducted to evaluate our *Group Social Force MDP* (GSF-MDP) approach and compare its predictive capabilities with the baselines.

Method	Parameters
GSF-MDP	$\alpha = 4.64$, $\beta = 18.65$, $I = (0.09, 0.02)$, $(a_k, b_k, \lambda) = (0.09, 0.32, 0)$, $(\beta_1, \beta_2, q_A, \phi, q_S) = (0.05, 1.18, 2.93, 0.38, 1.49)$
JS-MDP	$\alpha = 13.26$, $\beta = 9.12$, $I = (0.01,\ 0.19)$, $(a_k, b_k, \lambda) = (1.46,\ 0.11,\ 0)$
Karasev et al. [142]	$(w_{g,t}, w_{s,t}) = (0.03,\ 0.14)$, $\alpha = 21.31$, $\beta = 18.68$
Elfring et al. [80]	$(q_w, f_w, c_w) = (1.44,\ 0.23,\ 3.1)$, $\zeta_\rho = 83.74$

Table 3.2: Estimated hyperparameters in experiments with groups

Experiment 1: Predicting social Interactions

This experiment includes several qualitative demonstrations of the predicted group collision avoidance behavior of people. To this end we use maps of two environments and simulate observed trajectories in those maps to see the predicted development of interactive scenarios. The first scenario (Fig. 3.12) stages an experiment with 5 people in a narrow corridor. The second scenario (Fig. 3.13) sets up a challenging crowded environment with multiple non-convex obstacles and 21 people walking in 7 groups.

Experiment 2: Prediction accuracy evaluation

Quantitative evaluation of GSF-MDP is conducted using the 21 social scenarios, extracted from the ATC dataset. These scenarios feature trajectories of 172 people, including 90 pedestrians walking in groups, observed for long periods of time (see Fig. 3.2 for an example scenario). MHD and NLP metric values are calculated for each trajectory in the 21 interactive scenarios and averaged across 20 experiments for each scenario. We use 1.5 seconds as observation period, and predictions are obtained for $T = [2.5, \ldots, 12.5]$ seconds ahead. We also measure the average time to compute predictions using our algorithm and the baselines.

Prior to the main experiment, we perform hyperparameter optimization using the SMAC3 optimization toolbox [192] for each algorithm. Optimization criteria are to minimize the sum of NLP and MHD values. The optimal parameters are found to be as summarized in Table 3.2. The reward function and the action space parameters are used the same as in Sec. 3.3.1. Cell sizes of the grid maps are 0.05 m in Experiment 1 and 0.15 m in Experiment 2. The frequency of prediction is 4 Hz, the number of random walk samples $K = 200$.

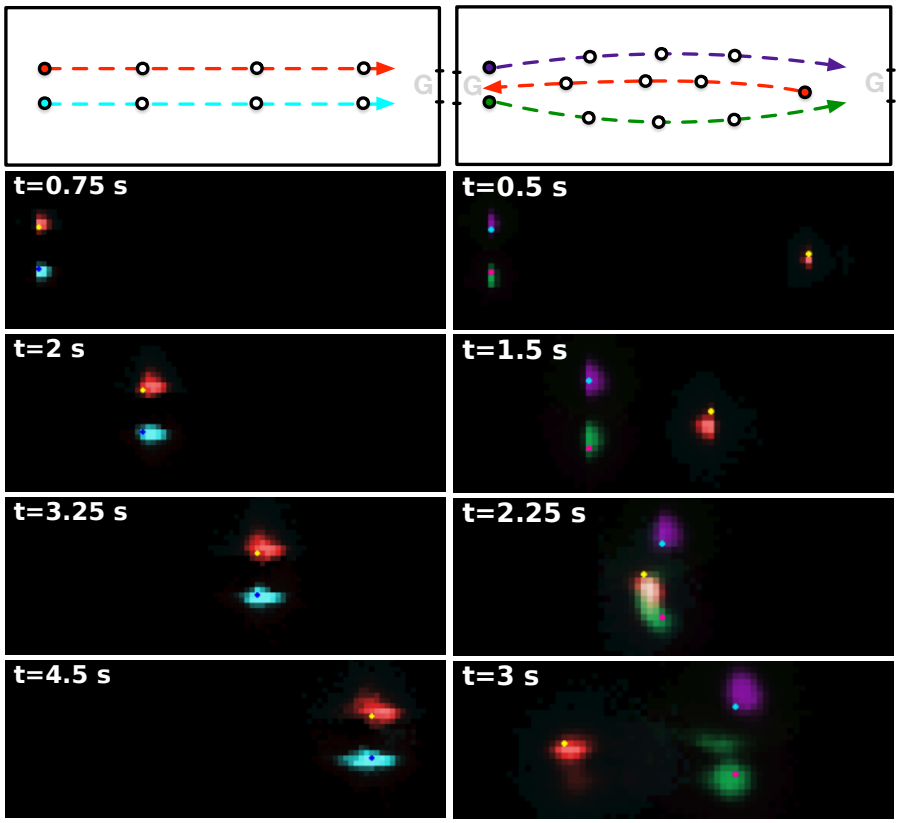


Figure 3.8: Prediction results in simulated interactive scenarios with no groups (part 1). Predicted distributions are color-coded, augmented with the ground truth position shown as a dot in contrasting color. **Top row:** schematic depiction of the situation, dashed lines show the path of each person. **Left:** two people walking together side-by-side, sharing a common goal ahead of them. **Right:** people walking in opposite flows, two of them make room for the third person walking in between.

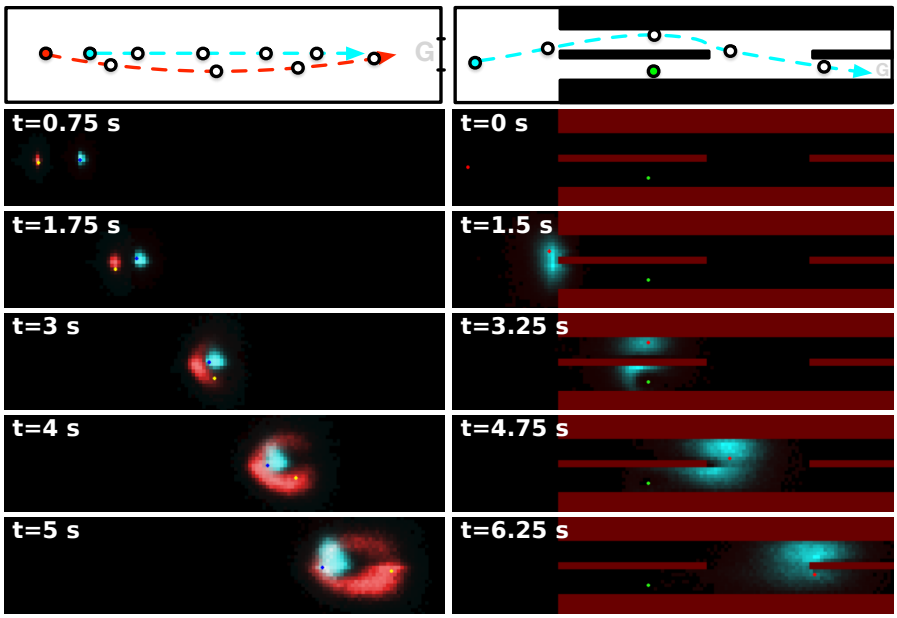


Figure 3.9: Prediction results in simulated interactive scenarios with no groups (part 2). **Left:** a fast walking person is overtaking a slow walker, both of them are heading towards the same goal. **Right:** a person causes a hindrance by blocking a narrow passage. In all cases, the algorithm makes goal-oriented predictions that correctly represent the local ambiguity caused by the other agent or the environment.

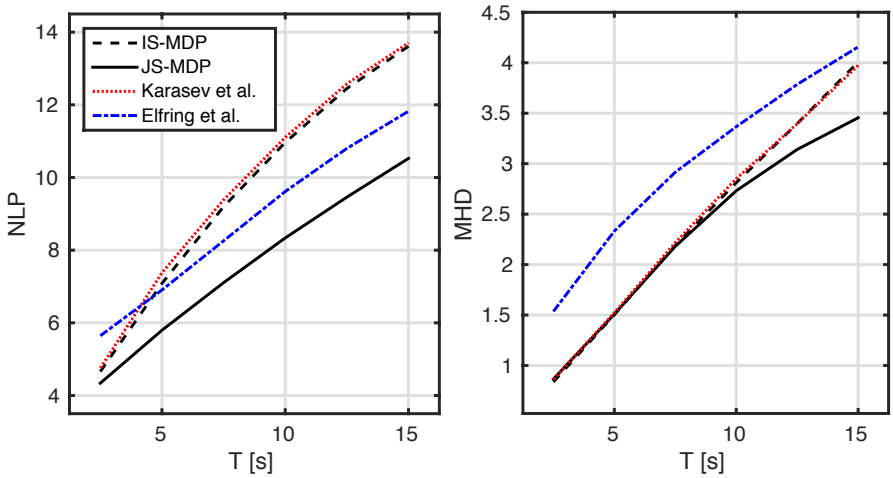


Figure 3.10: NLP and MHD evaluation results in scenarios with no groups. **Left:** Mean of the Negative Log-Probability (NLP) metric in the ATC dataset. Our approach outperforms the baselines along the entire prediction horizon of up to 15 seconds. **Right:** Mean of the Modified Hausdorff Distance (MHD) metric. Our approach is on par with the baselines for short-term predictions and outperforms them for longer horizons.

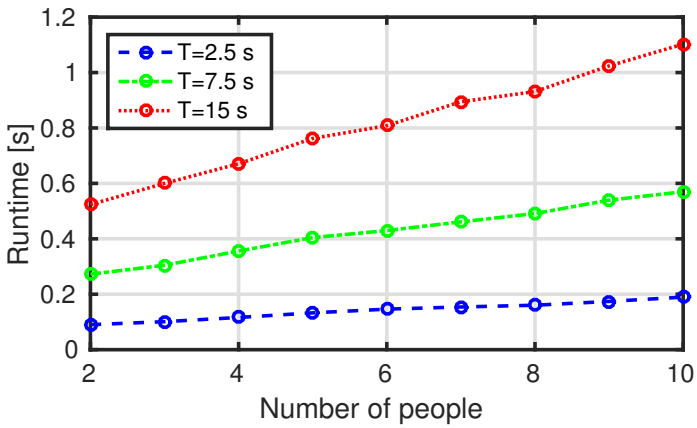


Figure 3.11: Average runtime of our algorithm in the ATC scenarios with no groups. The runtime is measured for prediction horizons T = 2.5, 7.5, 15 seconds ahead and conditioned on the number of people in the scenario.

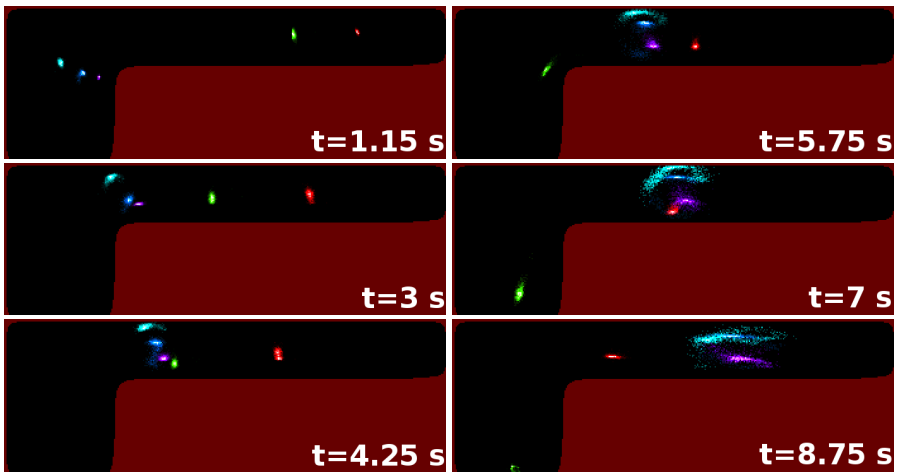


Figure 3.12: Prediction results in a simulated scenario with groups. Predicted distributions are color-coded. At $t = 1.15$ seconds a group of three people, depicted in **blue, cyan** and **purple** is walking upwards and then turns into the corridor to the right without losing its formation. At $t = 3$, $t = 4.25$ seconds the group is handling a hindrance caused by the **green** pedestrian, at $t = 5.75$, $t = 7$ seconds the group is handling another hindrance with the **red** pedestrian.

Figure 3.13: Prediction results in a simulated scenario with obstacles and 21 people walking in 7 groups. Goals are placed in the four corners of the map. **Left:** initial positions of people are shown in colored circles, each color corresponding to one group. **Right, top row:** predicted positions with GSF-MDP for several points in time. Consider e.g. the green group that waits until the passage is cleared by the red and blue groups without losing its formation. Then it gives way for the faster orange group. People in the red group are correctly predicted to maintain a side-by-side walking formation. **Right, bottom row:** predicted positions with the JS-MDP baseline, where group motion is not modeled. The green group performs unnecessary maneuvers, then gets separated. The same happens with the red and orange groups, who lose their members in the crowd.

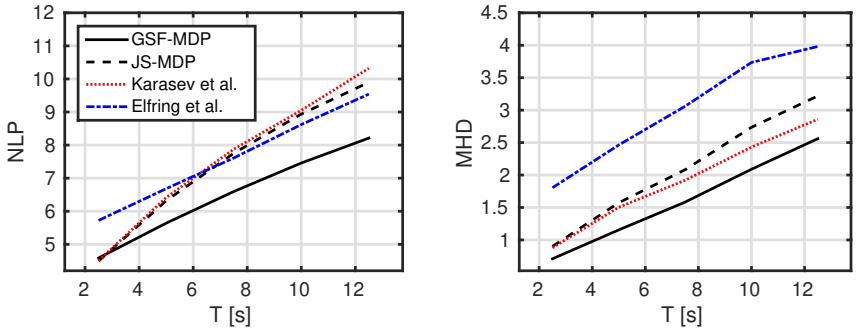


Figure 3.14: NLP and MHD evaluation results in scenarios with groups. **Left:** Mean of the Negative Log-Probability (NLP) metric in the ATC dataset. Our approach outperforms the baselines along the entire prediction horizon of up to 12.5 seconds. **Right:** Mean of the Modified Hausdorff Distance (MHD) metric. Our approach delivers more precise results on both short and long prediction horizons.

Results

Fig. 3.12 and Fig. 3.13 show the results of Experiment 1. The first simulated scenario (Fig. 3.12) demonstrates a collision avoidance maneuver, performed by a group of three pedestrians in a narrow corridor. The group is able to keep its "social" linear walking formation that facilitates intra-group interaction. In the end, however, the spreading of samples indicates the predicted possibility of re-grouping into a more compact V-formation – a behavioral pattern observed in real crowds [216]. In the second scenario (Fig. 3.13) our method predicts a realistic behavior of group members. In particular, they are able to wait for the passage to clear before continuing their motion as a group, keeping the broad V-shape walking pattern when the available space allows it, and not lose their members behind in the dense crowd. Predicted results are visually compared with a baseline, where the group motion is not modeled.

Fig. 3.14 presents the quantitative results of Experiment 2, displaying the mean of the NLP and MHD metrics over the prediction horizon of 2.5–12.5 seconds. The NLP results suggest that our algorithm assigns higher probabilities to the ground truth states of the person's future location, outperforming all the baselines. The planning-based method of Karasev et al. [142] accumulates errors from non-predicted social interactions over the growing prediction horizon, while JS-MDP [267] suffers from the lack of group awareness. The social force-based method of Elfring et al. [80] generates worse results due to the lack of global knowledge of the environment's structure. The MHD evaluation results further confirm the improvement of our method over the state-of-the-art on both short and long-term prediction horizons.

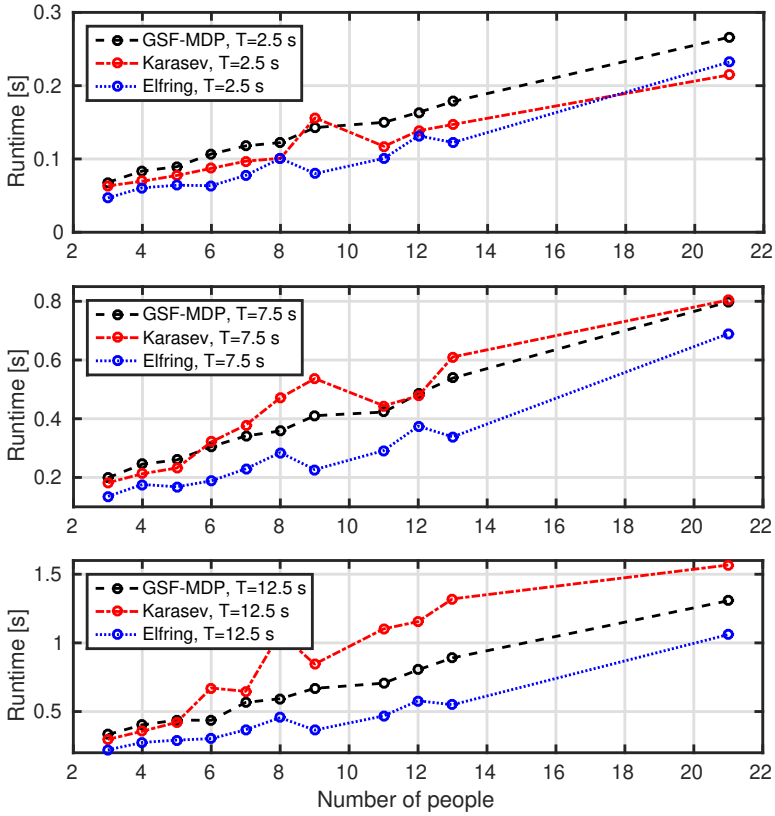


Figure 3.15: Average runtime of our algorithm for prediction horizons T = 2.5, 7.5 and 12.5 seconds ahead on the ATC dataset with various numbers of people. On average, our method performs on par with the baselines.

In Fig. 3.15 we give the prediction runtime of GSF-MDP compared to the baselines. For example, our method is capable of computing 2.5 seconds of predictions for 5 people in less than 0.1 seconds, or predict 7.5 seconds of 10 people motion in 0.4 seconds. On average, our method performs on par with the state-of-the-art. Given that the range of the social forces is not large, and people are typically not agglomerated in a single region, the method most often scales linearly with the number of people, and not quadratically as in the worst-case, described in Section 3.2.6.

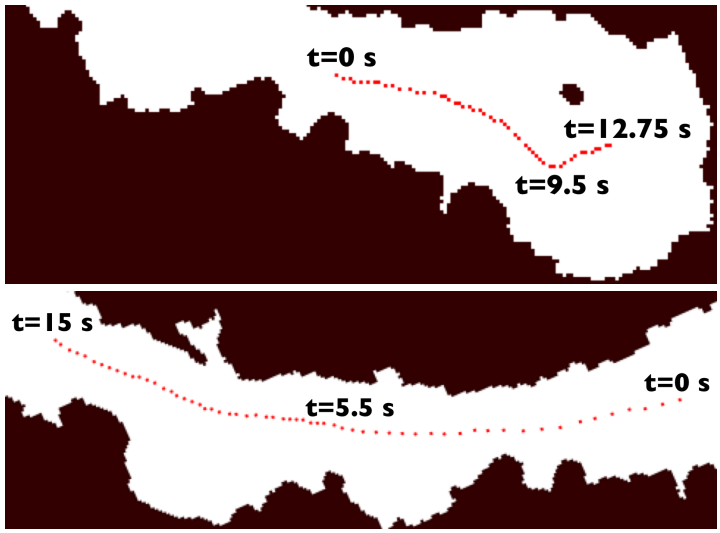


Figure 3.16: Challenging changes of motion dynamics in the ATC dataset. Pedestrian positions are measured at 4 Hz and plotted in **red**. We observe a change in motion intent (**top**, at $t = 9.5$ seconds) and motion velocity (**bottom**, at $t = 5.5$ seconds) not explainable by nearby people, group membership, environment geometry or other observable factors in the data.

3.4 Conclusion and Outlook

In this chapter we have discussed an elaborate solution for long-term motion prediction of pedestrians. Our method accumulates many key concepts from the motion prediction methodology in one efficient modular solution. In particular, such concepts include: explicit goal inference, usage of occupancy maps, accounting for varied velocities, modeling social aspects of human motion, utilization of social grouping information, predicting multi-modal probability distributions over possible future locations. The modular structure allows for further development and improvement of specific components, while the structure of the layered grid-map output is natively compatible with robot motion planners. Accordingly, in this section we outline the possibilities for future research along these two directions.

3.4.1 Semantic Context-awareness

The evaluation results presented above are encouraging. Performing at similar runtime with the state-of-the-art, our method is capable of delivering more accurate predictions across the entire prediction horizon. Still, during our ex-

periments we have encountered situations, which are generally challenging for long-term predictors, see for instance Fig. 3.16. Our stochastic policy accounts for variations in paths and homotopy classes, but does not handle sudden velocity or motion intent changes – this limitation in a long-term setting is a common unexplored aspect in the literature. Predicting paths accurately in situations shown in Fig. 3.16 could be done with a dynamic α value, which increases uncertainty for more distant points in time. Learning relevant stimuli for motion behavior in the environment and spatially incorporating them into the local behavior model could be another possibility to better foresee the uncertainty from sudden intention or velocity changes.

One could imagine the utility of context in semantically-rich indoor environments: for instance, a person approaching a narrow passage or a sharp corner, would probably slow down, and such behavior can be predicted in advance. A vector field-based velocity distribution, either learned [166] or modeled [75] in a given environment, could provide an insight into such behavior changes.

3.4.2 Combination with a Pattern-based Interaction Model

Our method shows an efficient way to make obstacle-aware predictions – an aspect in which purely pattern-based methods traditionally struggle. Clearly, to learn behaviors in every conceivable obstacle layout, such methods would require a tremendous amount of training data. On the contrary, our method achieves this with only a few trajectories for hyper-parameter estimation, and then can be applied in arbitrarily complex environments.

On the other hand, pattern-based methods have shown remarkable progress in modeling spatio-temporal interactions between people in obstacle-free spaces, outperforming the classical social force-based models in many situations. In the future work we intend to combine the global motion policies with a learned interaction model.

3.4.3 Robot Motion Planning Using Predictions

One key direction for future research on our prediction method is integration with the motion planner of a robot. The general idea for incorporating predictions into motion planning is to penalize robot locations that will probably be occupied by other agents at the same time. To this end, we overlay the predicted regions of occupancy L with the gridmap of static obstacles M, typically used by the robot for path planning and collision avoidance.

In the preliminary study [266] we compare three different predictive planning approaches: spatio-temporal discrete search in a time-augmented state space, used e.g. in [30], *costmap inflation* suggested by Bai et al. [19] and the *inferring collision points* (ICP) technique by Ziebart et al. [370]. We found the

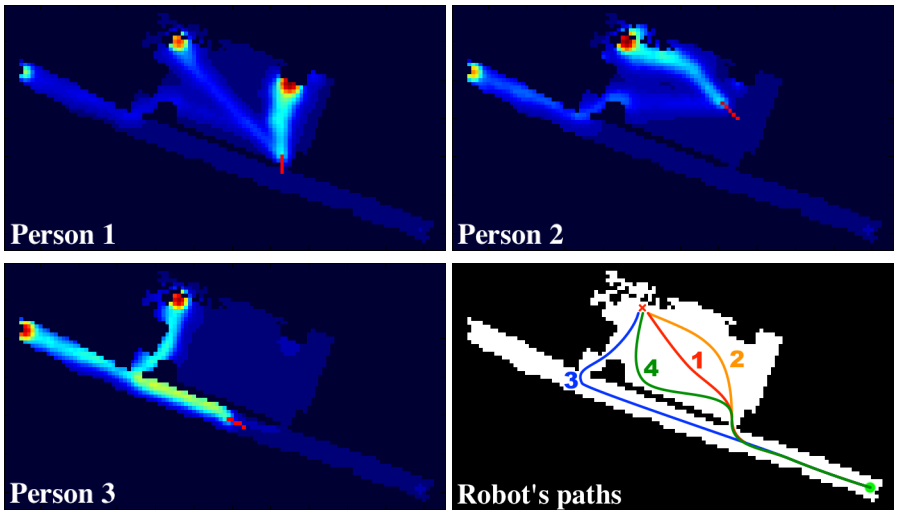


Figure 3.17: Predictive planning results. Predictions for three people are shown in **top row** and **bottom left**. Observed tracklets are depicted in **red**, predicted trajectories are represented with **heatmaps** (warmer colors correspond to higher occupancy probability). **Bottom right:** the robot is located in the top left corner of the room, its goal is in the bottom right corridor. Using the *inferring collision points* algorithm, the robot iteratively plans three paths (in **red**, **orange** and **blue**), before it finds the **green** path with no predicted collisions.

ICP method to be the best compromise between performance and efficiency. The method iteratively shapes a time-independent navigational cost function to remove known hindrance points. Initialized with the costmap $C = M$, at each iteration, ICP finds the A* solution in C, simulates it forward in time and compares the position of the robot to the corresponding time layer $L_t = \sum_i L_t^i$, inflating the cost of collision regions. See Fig. 3.17 for examples of predictive planning with the ICP algorithm.

In the future work we intend to study and validate the robot motion planning methods with explicit predictions. More details to this end are presented in Sec. 7.3.2.

Chapter 4
Occupancy Priors of Human Motion in Urban Environments

Understanding and anticipating human activity is an important capability for intelligent systems in mobile robotics, autonomous driving, and video surveillance. While learning from demonstrations with on-site collected trajectory data is a powerful approach to discover recurrent motion patterns, generalization to new environments, where sufficient motion data are not readily available, remains a challenge. In many cases, however, semantic information about the environment is a highly informative cue for the prediction of pedestrian motion or the estimation of collision risks. In this chapter, we infer occupancy priors of human motion using only semantic environment information as input. To this end we apply and discuss a traditional Inverse Optimal Control approach, and propose a novel one based on Convolutional Neural Networks (CNN) to predict future occupancy maps. Our CNN method produces flexible context-aware occupancy estimations for semantically uniform map regions and generalizes well already with small amounts of training data. Evaluated on synthetic and real-world data, it shows superior results compared to several baselines, marking a qualitative step-up in semantic environment assessment.

4.1 Introduction

Throughout this thesis it is shown that understanding and predicting human motion is an increasingly popular subject of research with the goal of improving the safety and efficiency of autonomous systems in spaces shared with people. Application areas include mobile service robots, intelligent vehicles, collaborative production assistants, video surveillance, or urban city planning. Human motion in these scenarios is influenced by many factors including other agents in the scene and the environment itself, which can be represented by a topometric map and semantic information. Indoor human navigation is often driven by

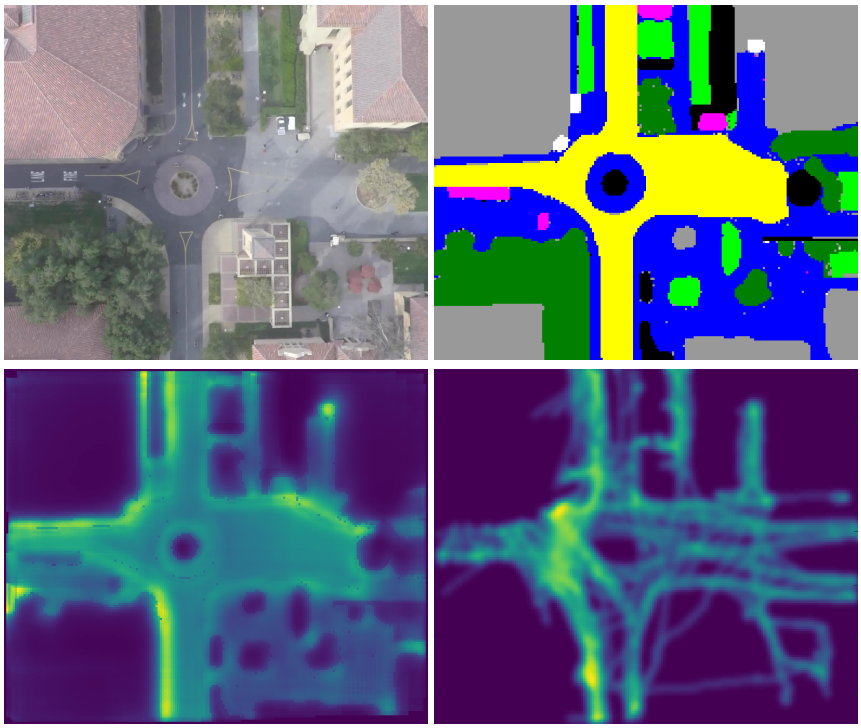


Figure 4.1: Predicting occupancy priors in semantically-rich urban environments. **Top left:** an urban scene from the Stanford Drone dataset. **Top right:** semantic map of the environment. **Bottom left:** CNN-predicted occupancy distribution priors of walking people in the environment, encoded with a heatmap: warmer colors correspond to states with higher probability of observing pedestrians **Bottom right:** ground truth occupancy distribution.

avoiding collisions with static and dynamic obstacles, motivating our prediction method design in Chapter 3. On the other hand, surface semantics have a strong impact in outdoor (e.g. urban) environments. For instance, pedestrians walk most of the time on sidewalks, sometimes on streets, unpaved areas and greenspaces, and very rarely over obstacles. Modeling the influence of semantics is a challenging task, typically approached with data-driven methods using human trajectory data in a given environment [81, 166, 333] without knowing the goal of the target agent. Powerful in scenes known beforehand, such approaches may suffer from poor generalization to never-seen or changing environments where no data is available.

In this chapter we research the possibility of inferring occupancy priors of walking people in previously unseen places with limited input, namely using

only the semantic map of the area. A prior occupancy distribution is intuitively interpretable and beneficial for a large variety of applications, such as improved goal estimation [254, 335] (for instance in Fig. 4.1 not all walking directions are likely to be the goal of a person) and crossing intention recognition in autonomous driving tasks [118, 245], where possible "illegal crosswalks" could be easily detected. The usage of semantic maps may further improve the accuracy of map-based motion prediction approaches [142], which often assume constant priors for each semantic class. Such occupancy estimation can guide a cleaning robot towards more heavily used areas, or a service robot in search of people to assist.

Traditionally, Inverse Reinforcement Learning (IRL) has been used to learn semantic preferences of walking people in urban and semantically-rich environments [153, 256, 320]. It is indeed possible to use the learned preferences to simulate trajectories and infer the prior occupancy distribution in a new environment. In this chapter we review the IRL methodology, applied to the occupancy prior distribution inference, and discuss its limitations. In order to address them, we introduce a novel extension to the recent method by Doellinger et al. [74], which uses a Convolutional Neural Network (CNN) to predict average occupancy maps indoors, with semantic map input for the urban scenes. We train our method on scenes from the Stanford Drone Dataset [262], as well as on simulated environments. In comparison to several baselines, our CNN method predicts much more accurate prior occupancy priors in terms of KL-divergence to the ground truth distributions, and makes a qualitative improvement of estimating flexible occupancy priors for semantically-uniform areas by considering local context and interconnections between different semantic regions.

4.1.1 Contribution

In summary, in this chapter we make the following contributions:

- We analyze and discuss state-of-the-art methodology for inferring occupancy prior distribution in semantically-rich urban environments.
- Addressing the limitations of the prior art, we propose a novel method based on Convolutional Neural Networks (displayed in Fig. 4.1 and 4.2).
- We execute a thorough comparison of the discussed methods with several baselines, and show qualitative and quantitative improvement of the KL-divergence scores when using our CNN method.

4.1.2 Outline

The chapter is structured as follows: in Sec. 4.2 we briefly review the related work on modeling semantics-awareness for autonomous systems, in Sec. 4.3 we

detail the proposed solutions and in Sec. 4.4 we describe the training and evaluation. Results are presented in Sec. 4.5, and a discussion in Sec. 4.6 concludes the chapter.

4.2 Related Work

The ability to understand a human environment and its affordances is useful in a number of tasks where intelligent autonomous systems need to reason on observed events, anticipate future events, evaluate risks and act in a dynamic world. Examples include person and group tracking [175, 194, 236], in particular over a camera network with non-overlapping fields of view, human-aware motion planning [19, 94, 229], motion behavior learning [225], human motion prediction [63, 268], human-robot interaction [174], video surveillance [4] or collision risk assessment [196]. Apart from basic geometric properties of the workspace, its semantics have a large impact on human motion in these tasks. Modeling this impact is challenging, therefore a popular approach is learning the motion patterns directly from data without explicitly specifying semantic features [81, 166, 313, 333]. However, many of those methods either need additional training input in new environments or experience transfer issues. To the best of our knowledge only a few methods explicitly highlight the performance in new environments outside the training scenario [20, 153, 292, 297]. As we will later describe, our approach explicitly uses only semantic maps as input and there will be no need to adapt the learned models to new environments, described with the same semantic classes.

Modeling the effect of surface classes on human motion was mainly used in reactive approaches such as [66] and planning-based approaches [142, 153]. In particular, these methods leverage semantic segmentation tools for understanding and detecting the semantics in the environment. Several approaches exist to segment available semantics [17, 219, 337] and to build semantic maps of the environment [96, 103] – a prerequisite to the methods presented in this chapter. We build on those methods to predict areas frequently used by pedestrians based on the semantic class of the surface.

Several Inverse Reinforcement Learning (IRL, or Inverse Optimal Control, IOC) approaches make use of semantic maps for predicting future human motion [142, 153, 292]. In particular, they use the semantic maps for encoding the features of the reward function. However, these IRL approaches are limited to one weight per feature and thus do not generalize well to new environments or heteregenous datasets with different geometries [342, 355]. In this chapter, we review an adaption of the Maximum Entropy IRL algorithm [369] to the task of occupancy priors estimation, and compare our CNN-based approach to it.

4.3 IOC and CNN Approaches for Occupancy Priors Estimation

In this chapter, we study the problem of estimating occupancy priors of walking humans in semantically-rich urban environments. The problem is formulated as follows: given a grid-map of the environment $\mathcal{M}$ with associated feature responses $\mathbf{f}(s) = [f_1(s), \ldots, f_K(s)]$, $\sum_{k=1}^{K} f_k(s) = 1$ for each state $s \in \mathcal{M}$ over the set of K semantic classes, we seek to estimate the probability $p(s)$ of a walking human being observed in this state.

If we assume having access to a large set of trajectories $\mathcal{T}_{\mathcal{M}}$ in $\mathcal{M}$, the problem of estimating $p(s)$ can be solved by counting visitation frequencies in each state:

$$p(s) = \frac{D(s)}{\sum_{s' \in \mathcal{M}} D(s')}\Big|_{\mathcal{T}_{\mathcal{M}}}, \tag{4.1}$$

where $D(s)$ is visitation count of state s over all trajectories $\mathcal{T}_{\mathcal{M}}$. In this paper, we estimate this distribution in environments where no trajectory data are available. One natural way to overcome the lack of trajectories is to simulate them, in particular using learned human walking preferences. To this end, in Sec. 4.3.1 we first review an Inverse Optimal Control (IOC) method [153] for predicting motion trajectories in semantic environments, and discuss its applicability and limitations. Then, in Sec. 4.3.2, we propose a novel approach based on Convolutional Neural Networks, which is an extension of the occupancy prior estimation method by Doellinger et al. [74]. For this task, we assume a semantic map of the environment $\mathbf{f}(\mathcal{M})$, or a method to extract it, to be available. Without loss of generality and for the sake of visual clarity, the states in this chapter are represented with one-hot vectors, i.e. $\forall s\ \exists\ k$ s.t. $f_k(s) = 1$ and $\forall j \neq k : f_j(s) = 0$.

4.3.1 Inverse Optimal Control on Multiple Maps (IOCMM)

Both Reinforcement Learning (RL) and Inverse RL or Inverse Optimal Control (IOC) frameworks deal with modeling optimal behavior of an agent, operating in a stochastic world $\mathcal{S}$ and collecting rewards $\mathcal{R}$ on the way to their goal state $s_g \in \mathcal{S}$. An agent's behavior is encoded in a policy $\pi(a|s)$, which maps the state $s \in \mathcal{S}$ to a distribution over actions $a \in \mathcal{A}$. When the reward function is not known beforehand, which is the case in many real-world applications, one possiblity is to learn it from a set of observations $\mathcal{T}$ with an IOC method. In this case, the reward function is parametrized by a set of parameters θ.

Modeling the behavior of an agent navigating in the environment, which is described with a set of features $\mathbf{f}(s)$ for each state, suits the problem of recovering occupancy priors from semantic map inputs well. Prior art, however,

Algorithm 4 Inverse Optimal Control: Backward pass

function BackwardPass($\mathcal{T}^i_{\mathcal{M}}, \theta$)
$V(s) \leftarrow -\infty$
for $n = N, \dots, 1$ **do**
  $V^{(n)}(s_g) \leftarrow 0$
  $Q^{(n)}(s, a) \leftarrow \mathcal{R}(s, \theta) + \mathrm{E}_{P^{s'}_{s,a}}[V^{(n)}(s')]$
  $V^{(n-1)}(s) \leftarrow \mathrm{softmax}_a\, Q^{(n)}(s, a)$
end for
$\pi(a|s) \leftarrow e^{\alpha(Q(s,a)-V(s))}$
return π

has not dealt with abstract quantities, such as occupancy expectations, focusing rather on the policy of an individual agent [370] or multiple agents jointly [167]. In this chapter we adapt the IOC framework to this task. As our IOC implementation is based on [153], we give a short summary of their approach in this section.

MDP-based Maximum Entropy Inverse Reinforcement Learning (MaxEnt IRL) [369] assumes that the observed motion of agents is generated by a stochastic motion policy, and seeks to estimate this policy with maximum likelihood to the available demonstrations. The reward an agent gets in state s is linear with respect to the feature responses in that state: $\mathcal{R}(s, \theta) = r_0 + \theta^\top \mathbf{f}(s)$, where $r_0 > 0$ is the base reward of a transition and θ is a set of weights or *costs* of the semantic classes: $\sum_{k=1}^{K} \theta_k = 1, \theta_k \in [0, 1]$. Given $\mathcal{R}$, the distribution over the sequence of states $\mathbf{s}$ is defined as

$$p(\mathbf{s}, \theta) = \frac{\prod_t e^{\mathcal{R}(s_t, \theta)}}{Z(\theta)} = \frac{e^{\sum_t r_0 + \theta^\top \mathbf{f}(s_t)}}{Z(\theta)} \tag{4.2}$$

Finding the optimal θ^* vector is equivalent to maximizing the entropy of $p(\mathbf{s}, \theta)$ in Eq. 4.2 while matching the semantic class feature counts of the training trajectories. An iterative procedure based on the exponentiated gradient descent of the log-likelihood $\mathcal{L} \triangleq \log p(\mathbf{s}|\theta)$ is described by Kitani et al. in [153]. The gradient $\nabla \mathcal{L}_\theta$ is computed as the difference between the *empirical* mean feature count $\bar{\mathbf{f}} = \frac{1}{|\mathcal{T}|} \sum_i^{|\mathcal{T}|} \mathbf{f}(\mathcal{T}^i)$, i.e. the average features accumulated over the $\mathcal{T}$ training trajectories in the map $\mathcal{M}$, and the *expected* mean feature count $\hat{\mathbf{f}}_\theta$, the average features accumulated by trajectories generated by the current parameters θ: $\nabla \mathcal{L}_\theta = \bar{\mathbf{f}} - \hat{\mathbf{f}}_\theta$. The weight vector is then updated as

$$\theta \leftarrow \theta e^{\lambda \nabla \mathcal{L}_\theta}, \tag{4.3}$$

where λ is the learning rate, and the expected mean feature count $\hat{\mathbf{f}}_\theta$ is computed using an iterative algorithm described below.

Algorithm 5 Inverse Optimal Control: Forward pass

```
function ForwardPass(T^i_M, π)
D ← 0
for n = 1, ..., N do
    s ← s_0
    while s ≠ s_g do
        D(s) ← D(s) + 1
        s′ ← π(a|s)
        s ← s′
    end while
end for
f̂_θ ← Σ_s f(s)D(s)
return f̂_θ
```

The algorithm iterates backward and forward passes, detailed in Alg. 4 and 5 respectively. The *backward pass* uses the current θ vector to compute the value function $V(s)$ for each state s in $\mathcal{M}$ given the goal state s_g – the final state of the trajectory $\mathcal{T}^i_{\mathcal{M}} \in \mathcal{T}_{\mathcal{M}}$. A stochastic motion policy $\pi_\theta(a|s)$ to reach s_g in $\mathcal{M}$ under $\mathcal{R}(s, \theta)$ is then computed and used in the *forward pass* to simulate several trajectories from s_0 to s_g, where s_0 is the initial state of $\mathcal{T}^i_{\mathcal{M}}$. The expected mean feature count is computed as a weighted sum of feature counts $\hat{\mathbf{f}}_\theta = \sum_s \mathbf{f}(s)D(s)$ in the simulated trajectories, and the backward-forward iteration is repeated for a batch of trajectories in $\mathcal{T}_{\mathcal{M}}$. The θ vector is updated as in Eq. 4.3 using the cumulative $\hat{\mathbf{f}}_\theta$ for the trajectories in the batch, and the algorithm is iterated until the gradient $\nabla\mathcal{L}_\theta$ reaches zero. In order to learn from multiple maps, including those where only a subset of K features is present, we run the backward and forward passes for a batch of trajectories *in a batch of maps*, accumulating the visitation counts $D(s)$ across several maps. The resulting Inverse Optimal Control on Multiple Maps (IOCMM) method is detailed in Alg. 6.

Having obtained the optimal θ^* weights, it is possible to compute the reward $\mathcal{R}(s, \theta^*)$ and simulate trajectories in any environment which is described by a subset of K semantic features. By simulating semantic-aware trajectories, an average visitation count for each state, normalized across all states in $\mathcal{M}$, yields the occupancy probability $p(s)$, as in Eq. 4.1. Apart from the θ^* vector, this simulation depends on the distributions from which the initial and goal states s_0 and s_g (hereinafter denoted $s_{0,g}$) are drawn: since the algorithm is inherently unaware of the semantics behind classes, omitting this step may result in $s_{0,g}$ generation inside of obstacles or other high-cost areas. To counteract this issue, we consider two strategies: (1) directly learn probabilities to sample the start or goal position in a state s, conditioned on the semantic class $\mathbf{f}(s)$ of

Algorithm 6 Inverse Optimal Control on Multiple Maps (IOCMM)

$\theta \leftarrow 1/K$
repeat
  $\hat{\mathbf{f}}_\theta \leftarrow 0,\ \bar{\mathbf{f}} \leftarrow 0$
  Batch B_m maps
  for $m = 1, \ldots, B_m$ **do**
    $\mathcal{T} \leftarrow$ Batch B_t trajectories from $\mathcal{M}_m$
    $\bar{\mathbf{f}} \leftarrow \bar{\mathbf{f}} + \frac{1}{|\mathcal{T}|} \sum_i^{|\mathcal{T}|} \mathbf{f}(\mathcal{T}^i)$
    $\mathcal{R}(s, \theta) \leftarrow r_0 + \theta^\top \mathbf{f}(s)$
    for $i = 1, \ldots, B_t$ **do**
      $\pi \leftarrow$ `BackwardPass`(s_g)
      $\hat{\mathbf{f}}_\theta \leftarrow \hat{\mathbf{f}}_\theta +$ `ForwardPass`(s_0, π)
    end for
  end for
  $\bar{\mathbf{f}} \leftarrow$ `normalize`$(\bar{\mathbf{f}})$
  $\hat{\mathbf{f}}_\theta \leftarrow$ `normalize`$(\hat{\mathbf{f}}_\theta)$
  $\nabla \mathcal{L}_\theta \leftarrow \bar{\mathbf{f}} - \hat{\mathbf{f}}_\theta$
  $\theta \leftarrow \theta e^{\lambda \nabla \mathcal{L}_\theta}$
until $\|\nabla \mathcal{L}_\theta\| < \epsilon$

the state: $p(s_{0,g}|\mathbf{f}(s))$, and (2) generate the $s_{0,g}$ only from low-cost regions with the softmax function over the estimated cost of the state: $p(s_{0,g}) \sim e^{-\mathcal{R}(s,\theta^*)/\tau}$. Furthermore, to generate long trajectories spanning across the map, both distributions (1) and (2) are scaled linearly proportional to the distance between the $s_{0,g}$ and the center of the map.

Analysis and discussion

While delivering adequate results in our experiments, as we show in Sec. 4.4, the IOCMM approach to occupancy priors estimation has an inherent drawback. With rigid costs of a semantic class k, defined by the corresponding θ_k weight, the IOCMM method cannot produce flexible estimations for a spatial region given its position in a wider topological structure. For instance, if we assume that the grass surface is walkable, then it will have a low cost and predicted people would largely ignore paved paths in a park. However, this behavior is probably not confirmed in the training data, which will increase the costs of the grass regions, potentially making them not traversable in some cases where such behavior is expected. Controlled by one θ_k parameter, the cost of the semantic class stays constant over the entire map. Similarly, learning to step on the road surface in places where this behavior is unavoidable will inevitably lead to decreasing the costs of the road surface *everywhere* in

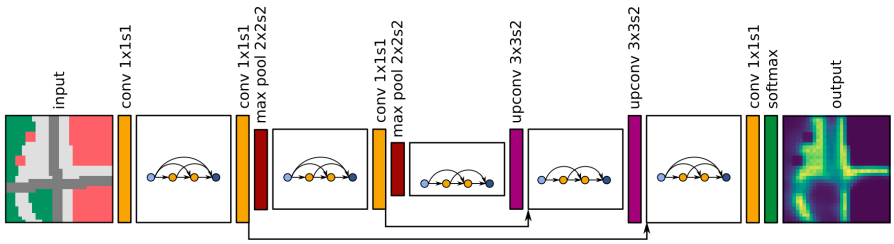


Figure 4.2: Structure of the *semapp* network. Input semantic tensor of size `<height>x<width>x<features>` is downsampled twice before passing through a bottleneck and getting upsampled to the original resolution again. Black arrows indicate skip connections. The kernel sizes and strides are denoted as `<kernel width>x<kernel height>s<stride>`.

the map. Learning such flexible behavior requires reasoning over the local context and interconnections between different semantic attributes and the surface. To this end, we propose our Convolutional Neural Network-based approach "Semantic Map-Aware Pedestrian Prediction" (*semapp*), described in the next section.

4.3.2 Semantic Map-Aware Pedestrian Prediction (semapp)

Convolutional Neural Networks (CNNs) have shown great successes for operations on map data, such as semantic segmentation [219] or value function estimation for deep reinforcement learning [306]. For our task of predicting occupancy distributions of walking humans in semantic environments, we need a method to map the feature responses $\mathbf{f}(\mathcal{M})$ to probabilities $\mathrm{p}(\mathrm{s})$. To this end, we extend the network to predict occupancy values in semantics-free geometric environments [74]. This network, based on the FC-DenseNet architecture [133], has reasonably few parameters which helps to avoid overfitting when training on limited amounts of data. Experiments with different architectures have been made in [74] but the authors have found their results to be very robust to such changes. We thus decided to perform no further optimization on the network architecture. The method in [74] is referred to as Map-Aware Pedestrian Prediction (*mapp*), therefore we call our extension "Semantic mapp", or *semapp*. Extending the architecture from [74] to semantic inputs by changing the input from one binary input channel to one channel for each semantic class allows the network to differentiate between pedestrians walking on grass, sidewalks and streets additional to avoiding obstacles. The architecture is outlined in Fig. 4.2.

The network directly outputs the map-sized tensor with the occupancy distribution, so, unlike IOCMM, semapp requires no trajectory simulation for inference. Consequently, for training we convert the trajectories $\mathfrak{T}_{\mathcal{M}}$ in each map $\mathcal{M}$ into the occupancy distribution using Eq. 4.1. This conversion itself

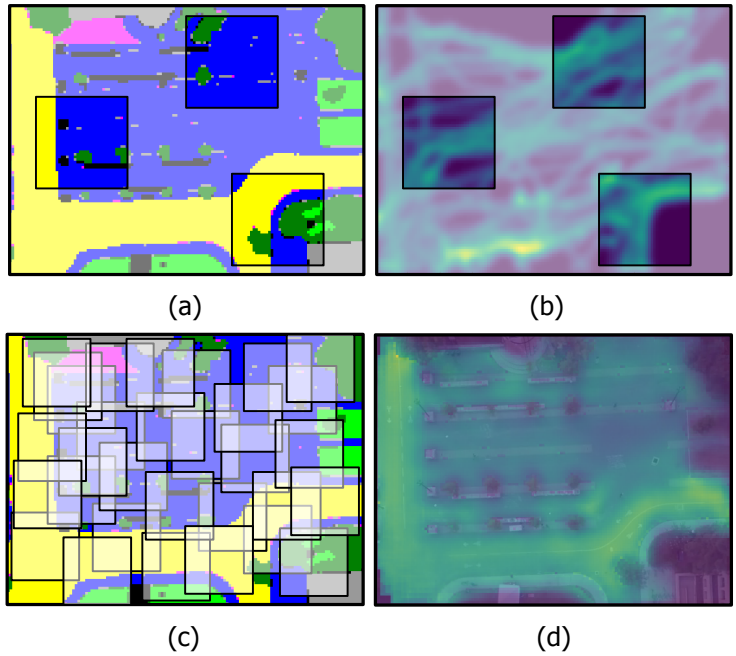


Figure 4.3: (a) CNN training on crops from larger maps in the Stanford Drone Dataset, (b) corresponding crops from the ground truth distribution. (c) Inference for a large map by averaging the inference for random crops in the test image, (d) full predicted occupancy distribution from 500 crops.

is not without meaning: trajectories, as compared to the processed occupancy distribution, contain additional temporal information. However, this information is not necessarily relevant for the task at hand - in fact, we are deliberately discarding the temporal aspect of human motion, inferring instead the generalized prior of observing a person in any state of the environment. Using directly the distribution emphasises the relationship between the topology of the environment and the desired occupancy priors. Furthermore, it relaxes the requirements to the data itself: detections are sufficient, and there is no need for continuous tracks.

Since the network operates on map crops of fixed size, we decompose a larger input image into a number of random crops of appropriate size and then rebuilt the final distribution $p(s)$ for state s as an average of predicted occupancy values of s in all crops which include that state (see Fig. 4.3b). In this case random crops, as compared to regular grids, remove the aliasing issues from combining adjacent crops.

Dataset	U4	Stanford Drone
Number of maps	80	25
Map size in pixels	32×32	$\sim 146 \times 152$
Resolution	simulation	0.4 m
Number of trajectories per map	~ 30	~ 132
Number of semantic classes in the dataset	4	9

Table 4.1: Datasets summary

4.4 Experiments

In this section, we give an overview of the training data (Sec. 4.4.1) and the experiments' design, as well as details on the training and baseline implementation (Sec. 4.4.2).

4.4.1 Datasets

We evaluate all methods on two datasets of human trajectories in semantically-rich environments: the Stanford Drone Dataset [262] and a set of simulated maps. Both datasets are summarized in Table 4.1.

To prove the concept of learning occupancy distributions from semantic maps, we created the "U4" dataset which includes 80 hand-crafted maps of Urban environments with four semantic classes (sidewalk, grass, road and obstacle) and manually marked trajectories in each map. In this dataset, we pay particular attention to "illegal crosswalk" detection, i.e. such scenes where global topology of the environment encourages people to step onto the driveway and cross it. Additionally, as people often tend to cut sharp corners by walking over grass, such behavior is also included in this dataset. Several scenes from U4 are shown in Fig. 4.4.

The *Stanford Drone Dataset* (SDD) [262] was recorded on the Stanford University grounds, which include a wide variety of environments and semantic classes, e.g. shared roads for cyclists and vehicles, pedestrian areas, college buildings, vegetation and parking lots. The dataset includes 51 top-down scenes with bounding boxes for various agents, from which we extracted trajectories of people, approximating the position by the center of the bounding box. We chose 25 scenes sufficiently covered by trajectories and scaled the maps to the constant physical resolution of 0.4 m per cell. We manually segmented each scene into nine semantic classes: pedestrian area, vehicle road, bicycle road, grass, tree foliage, bulging, entrance, obstacle and parking. Some example scenes from SDD are shown in Fig. 4.1, 4.3 and 4.5.

	U4 dataset			Stanford Drone dataset		
	Training		**Inference**	**Training**		**Inference**
IOCMM	Traj. batch B_t: 10 Map batch B_m: 7 Base reward r_0: 0.01	Stoch. policy α: 0.1 Learning rate λ: 1.0	$s_{0,g}$ sampl. τ: 0.01 Num. sim. traj.: 500	Traj. batch B_t: 50 Map batch B_m: 5 Base reward r_0: 0.01	Stoch. policy α: 0.1 Learning rate λ: 1.0	$s_{0,g}$ sampl. τ: 0.01 Num. sim. traj.: 1000
Semapp	Pooling layers: 1 Growth rate: 2 Layers per block: 5 Num. conv. layers: 19 Num. param.: 6.5 k Weight decay: 4.5×10^{-5}	Crop size: 32 Batch size: 32 Dropout prob.: 0.35 Learning rate: 0.01 Learn. rate decay: 0.9985	Num. crops: 1	Pooling layers: 1 Growth rate: 2 Layers per block: 5 Num. conv. layers: 19 Num. param.: 6.5 k Weight decay: 4.5×10^{-5}	Crop size: 64 Batch size: 32 Dropout prob.: 0.35 Learning rate: 0.01 Learn. rate decay: 0.9985	Num. crops: 500

Table 4.2: IOCMM and semapp parameters used for training and inference in the U4 and Stanford Drone datasets

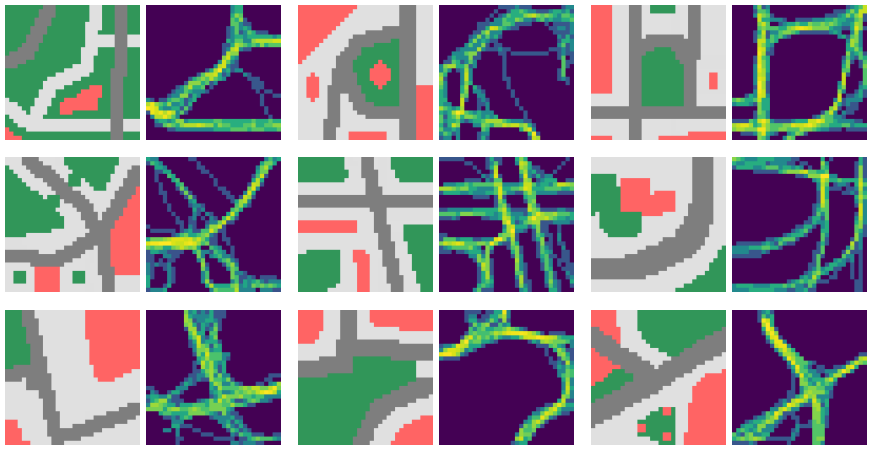


Figure 4.4: Training examples from the U4 dataset: each pair shows the semantic map on the left and the ground truth occupancy distribution on the right. Semantic classes include vehicle road in **dark gray**, pedestrian areas and sidewalks in **light gray**, unpaved areas and grass in **green** and obstacles in **red**.

4.4.2 Training and Evaluation

To our knowledge, there exist quite some works on human motion prediction, but none of the existing methods predicts prior occupancy distribution of walking people in urban environments only based on semantic information – a task considerably different from trajectory prediction [142, 153]. Therefore in our experiments we mainly compare the IOC and CNN solutions to the problem against the ground truth distributions and the following baseline methods:

1. uniform distribution over $\mathcal{M}$
2. uniform distribution over the walkable states in $\mathcal{M}$
3. semantics-unaware *mapp* network [74]

For quantitative evaluation we measure *Kullback-Leibler divergence* (KL-div) between the predicted and the ground truth distribution:

$$D_{KL}(P_{GT} \| Q_{Pred.}) = \sum_{x \in \mathcal{M}} P_{GT}(x) \log \frac{P_{GT}(x)}{Q_{Pred.}(x)}. \tag{4.4}$$

We train and evaluate IOCMM, semapp and the baselines separately on the U4 and Stanford Drone datasets. Training and inference parameters are summarized in Table 4.2.

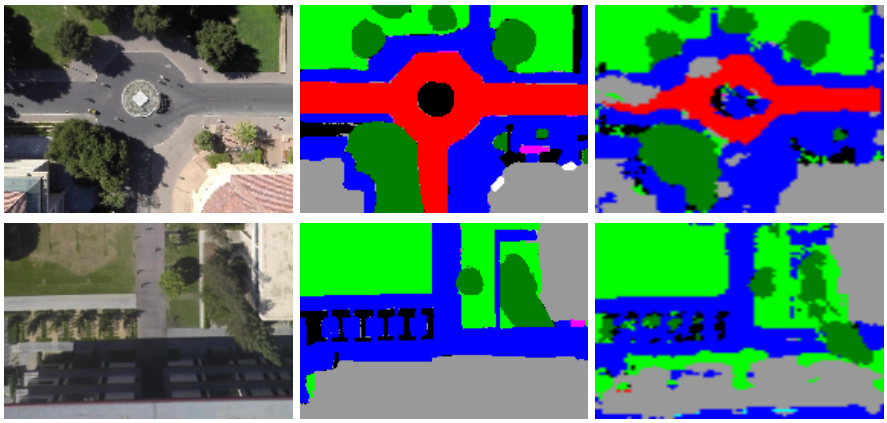


Figure 4.5: Semantic segmentation, obtained by our proof-of-concept pipeline in the Stanford Drone dataset. **Left:** input images. **Middle:** ground truth labels. **Right:** predicted semantic map.

We optimized the hyperparameters for IOCMM prior to the main experiments on a small portion of data from both datasets. For KL-div benchmarking we learn the θ weights from leave-one-out maps in the dataset, and validate the result in the remaining map, iterating over all maps in the respective dataset. Furthermore, we evaluate the impact of the number of simulated trajectories for inference in the new map (as in Eq. 4.1), measuring runtime and solution quality. Finally, we estimate the two $s_{0,g}$ sampling strategies, detailed in Sec. 4.3.1, separately.

Training the mapp and semapp networks on the U4 dataset is straightforward as the size of the maps (32 by 32 pixels) is equivalent to the network input. From the larger images in the SDD (on average 146 by 152 pixels) we take 500 random crops of size 64 by 64 pixels. Each crop in the training data is augmented 7 times by rotation and mirroring. In both cases, leave-one-out maps are used for training and validation (50/50), and the remaining map is used for evaluation. We followed the training procedure and hyperparameters from [74] which turned out robust enough for our application. In particular, we trained the networks for 100 epochs with binary cross-entropy loss using the Adam optimizer [151], and stopped the training when the performance on the validation set did not improve for 15 epochs. Further parameters are detailed in Table 4.2.

4.4.3 A Remark on Semantic Segmentation

As a proof-of-concept that the semantic maps, required for the methods presented in this chapter, can be obtained from images during runtime, we have set up a preliminary pipeline for semantic segmentation using UNet implementation in Keras and TensorFlow [106]. We trained the CNN with the 19 images from the Stanford Drone dataset, augmented 3 times with rotation, and tested on the remaining 6 images. Even with such negligible amount of data, our experiments (see Fig. 4.5) reached 0.64 frequency-weighted IoU in the training dataset (78% accuracy), and 0.53 IoU (69% accuracy) in the test dataset. We are confident, that using state-of-the-art semantic segmentation techniques [1] the performance, necessary for the application of our method, will be reached. Combining these two pipelines is of prime priority for our future work.

4.5 Results

We report the mean and standard deviations of the KL-divergences for both datasets in Table 4.3. In the U4 dataset, both IOCMM and semapp outperform the other baselines, furthermore both proposed sampling strategies for IOCMM show similar performance after appropriate hyperparameter optimization. Semapp, in addition to the quantitative improvement of minimum 14% over the closest baseline (IOCMM), offers a clear qualitative improvement in identifying crucial non-linearities in the predicted priors, as displayed in Fig. 4.6. This figure shows the extent to which semantics of the environment impact the distribution prediction – all semantics-unaware methods in our comparison, e.g. uniform $\mathrm{p}(\mathrm{s})$ over $\mathcal{M}$ and mapp, perform poorly. On the contrary, in both datasets semapp outperforms all baselines, due to its ability to reason over spatially-connected regions using convolutions, learning not only local contexts where motion probability is high, but also which locations are usually avoided by pedestrians. Interestingly, in the SDD dataset, due to incomplete ground truth coverage of the scenes (as seen in Fig. 4.1 and 4.3), removing unwalkable spaces from the uniform distribution over all states in $\mathcal{M}$ only decreases performance of this baseline. The reason here is that for many walkable states, where no motion is recorded in the ground truth, probabilities increase, resulting in worse KL-div scores. Despite using this imperfect training material, semapp consistently outperforms all baselines with the best KL-div score and smaller standard deviation between maps.

In Fig. 4.7, we visualize the generalization capabilities of the IOCMM method. To this end, we show the optimal θ^* weights in each individual map in both datasets, and compare them to the globally optimal set of weights, learned from training on all maps in the respective dataset. Here lies one benefit of the Inverse Optimal Control strategy to find occupancy priors: IOCMM does not only generalize well on a large amount of maps, but also retains high performance when learning from small amounts of data. In fact, when training on a

Method	U4 dataset	Stanford Drone
Uniform $\mathbf{p}(\mathbf{s})$ over $\mathcal{M}$	1.21 ± 0.26	1.40 ± 0.31
Uniform $\mathbf{p}(\mathbf{s})$ over walkable states in $\mathcal{M}$	0.97 ± 0.28	1.69 ± 0.42
Uniform $\mathbf{p}(\mathbf{s}\|\mathbf{f}(\mathbf{s}))$ learned from $\mathcal{T}_{\mathcal{M}}$	0.53 ± 0.09	1.09 ± 0.26
mapp CNN	0.93 ± 0.27	1.03 ± 0.19
IOCMM with learned $\mathbf{p}(\mathbf{s}_{0,g})$	0.42 ± 0.07	1.04 ± 0.24
IOCMM with modeled $\mathbf{p}(\mathbf{s}_{0,g})$	0.43 ± 0.07	1.21 ± 0.32
semapp CNN	$\mathbf{0.37 \pm 0.11}$	$\mathbf{0.71 \pm 0.13}$

Table 4.3: Average KL-div in the U4 and Stanford Drone datasets

fraction of maps from the dataset (e.g. as little as 10 random maps), IOCMM on average still converges to the globally-optimal θ^* costs for semantic classes, and thus the KL-div scores do not drop. This property is not shared by semapp, which needs a large selection of maps sufficiently covered by trajectories to generalize across various local contexts.

Both methods, IOCMM and semapp, depend on the number of random samples during inference. IOCMM samples trajectories between random start and goal positions, while semapp samples random crops from the larger semantic map. In Fig. 4.8 we show the relation between the number of samples, performance and inference time in the large maps of the SDD dataset. Runtimes were measured for Python implementations of both algorithms on an ordinary laptop with Intel Xeon 2.80GHz $\times$ 8 CPU and 32 GB of RAM. The CNN is implemented using Theano on the built-in GPU Quadro M2000M. It is worth mentioning that the inference time of semapp on one crop of size 64 by 64 pixels (equivalent to 25.6×25.6 m) is ~ 0.054 seconds, appropriate for real-time application.

4.6 Conclusions and Outlook

In this chapter, we looked into the problem of learning human occupancy priors in semantically-rich urban environments using only the semantic map as input. Considering two established classes of approaches to this end (Inverse Optimal Control and Convolutional Neural Network), we show that our CNN-based semapp approach is outperforming all baselines already with limited training data. The IOCMM approach, on the other hand, can be used to reasonably estimate the costs of semantic classes from several maps and few trajectories. However, it is limited to constant weights, which may not reflect behavior of people in all local contexts of semantically-complex environments. This approach lacks reasoning on spatial relevance of surfaces to infer cases where people may prefer to walk on one surface class over another, or not walk at all.

In future work we intend to further investigate the possibilities of applying advanced IOC techniques, for instance non-linear IRL with complex fea-

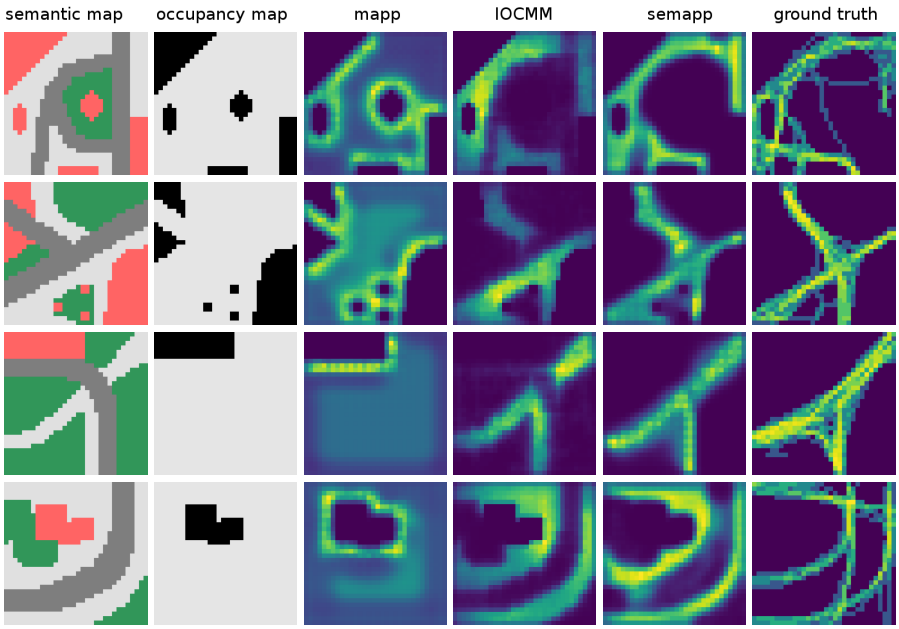


Figure 4.6: Qualitative comparison of results in the U4 dataset. A binary *occupancy map* of the environment highlights the amount of structure imposed by semantics in urban scenes. There is little surprise that the semantics-unaware mapp approach for learning occupancy priors [74] is not learning any meaningful behaviors apart from the fact that people (often) tend to be found close to obstacles. On the contrary, IOCMM correctly estimates the priors in different walkable areas. On top of that, CNN-based semapp is capable of detecting all "illegal crosswalks" in these scenes, as well as cutting over grass in such places where the topology of the environment encourages to do so, e.g. see the sharp corner in the third row.

tures [183], non-linear reward modeling [204], automated feature extraction to exploit local correlations in the environment [182, 211], IOC with multiple locally-consistent reward functions [221] and with CNN-based reward function approximator [338]. Furthermore, we plan to validate semapp with on-the-fly semantics estimation and extend it to first-person view for application in automated driving to infer potential pedestrians' entrance points to the road.

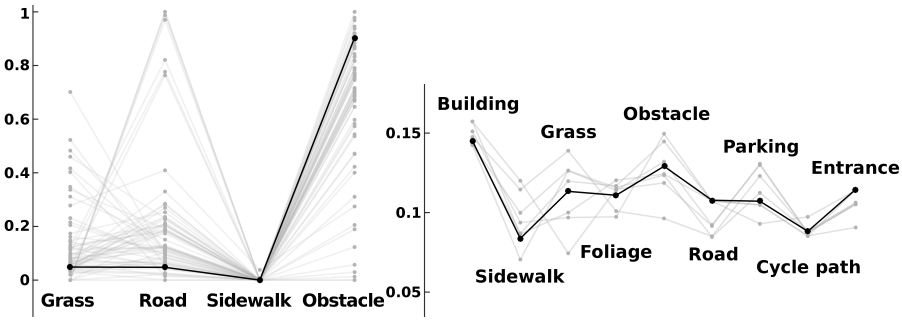


Figure 4.7: Optimal θ costs of various semantic classes, learned by IOCMM in each individual map, are shown in gray. Globally optimal weights for the entire dataset are overlaid in black. **Left:** U4 dataset. **Right:** Stanford Drone dataset.

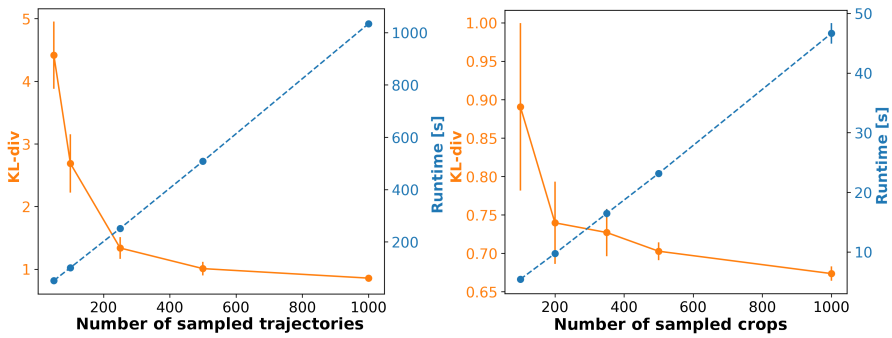


Figure 4.8: Runtime and performance of IOCMM (**left**) and semapp (**right**) in the SDD dataset as a function of the number of sampled trajectories/crops during inference. Both methods' runtime scales linearly with the number or random samples, while performance improves exponentially.

Chapter 5
Data Collection for Motion Prediction

Understanding human behavior is key for robots and intelligent systems that share a space with people. Accordingly, research that enables such systems to perceive, track, learn and predict human behavior as well as to plan and interact with humans has received increasing attention over the last years. The availability of large human motion datasets that contain relevant levels of difficulty is fundamental to this research. Existing datasets are often limited in terms of information content, annotation quality or variability of human behavior. In this chapter, we present THÖR, a new dataset with human motion trajectory and eye gaze data collected in an indoor environment with accurate ground truth for position, head orientation, gaze direction, social grouping, obstacles map and goal coordinates. THÖR also contains sensor data collected by a 3D lidar and involves a mobile robot navigating the space. We propose a set of metrics to quantitatively analyze motion trajectory datasets such as the average tracking duration, ground truth noise, curvature and speed variation of the trajectories. In comparison to prior art, our dataset has a larger variety in human motion behavior, is less noisy, and contains annotations at higher frequencies.

5.1 Introduction

In Chapter 1 we have outlined many tasks which require understanding human motion behavior in automated driving, mobile robotics, intelligent video surveillance systems and motion simulation. Human motion trajectories are a valuable learning and validation resource in these tasks. For instance, they can be used for learning safe and efficient human-aware navigation, predicting motion of people for improved interaction and service, inferring motion regularities and detecting anomalies in the environment.

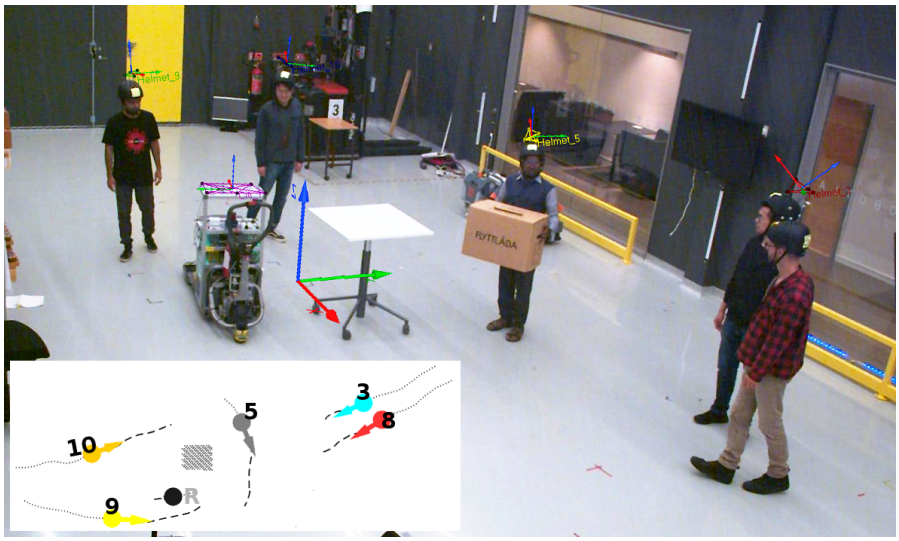


Figure 5.1: Environment configuration. Participants, wearing tracking helmets, and the robot are moving towards their goals in a shared space, tracked by the Qualisys motion capture system (recorded motion in the bottom left corner).

Datasets of ground level human trajectories, typically used for learning and benchmarking, include the ETH [236], Edinburgh [205] and the Stanford Drone [262] datasets, recorded outdoors, or the indoor ATC [48], L-CAS [349] or Central Station [366] datasets (see Table 5.1). While providing the basic input of motion trajectories, these datasets often lack relevant contextual information and the desired properties of data, e.g. the map of static obstacles, coordinates of goal locations, social information such as the grouping of agents, high variety in the recorded behaviors or long continuous tracking of each observed agent. Furthermore, most of the recordings are made outdoors, a robot is rarely present in the environment and the ground truth pose annotation, either automated or manual, is prone to artifacts and human errors.

Dataset	Location	Map	Goal positions	Groups	Head orientation	Eye gaze	Robot in the scene	Sensors for pose estimation	Frequency	Annotation
ETH [236]	Outdoors	✓	✓	✓				Camera	2.5 Hz	Manual
UCY [181]	Outdoors					✓		Camera	Continuous	Manual
VIRAT [224]	Outdoors	✓*						Camera	2,5,10 Hz	Manual
KITTI [97]	Outdoors	✓					✓	Velodyne and several cameras	10 Hz	Manual
Edinburgh [205]	Outdoors							Camera	6-10 Hz (variable)	Automated
Stanford Drone [262]	Outdoors	✓*						Camera	30 Hz	Manual
Town Center [28]	Outdoors	✓*						Camera	15 Hz	Manual
ATC [48]	Indoors				✓			Several 3D range sensors	10-30 Hz (variable)	Automated
Central station [366]	Indoors							Camera	25 Hz	Automated
L-CAS [349]	Indoors			✓			✓	3D LiDAR	10 Hz	Manual
KTH [76]	Indoors						✓	RGB-D, 2D laser scaner	10-17 Hz (variable)	Automated
THÖR	Indoors	✓	✓	✓	✓	✓	✓	Motion capture	100 Hz	Ground truth

* Unsegmented camera image.

Table 5.1: Contextual cues in the datasets of human motion trajectories

5.1.1 Contribution

In this chapter we present a human-robot interaction procedure, designed to collect motion trajectories of people in a generic indoor social setting with extensive interaction between groups of people and a robot in a spacious environment with several obstacles. The locations of the obstacles and goal positions are set up to make navigation non-trivial and produce a rich variety of behaviors. The participants are tracked with a motion capture system; furthermore, several participants are wearing eye-tracking glasses. "Tracking Human motion in the ÖRebro university" (THÖR) dataset[1], which is released public and free for non-commercial purposes, contains over 60 minutes of human motion in 395k frames, recorded at 100 Hz, 2531k people detections and over 600 individual and group trajectories between multiple resting points. In addition to the video stream from one of the eye tracking headsets, the data includes 3D Lidar scans and a video recording from stationary sensors. We quantitatively analyze the dataset using several metrics, such as tracking duration, perception noise, curvature and speed variation of the trajectories, and compare it to popular state-of-the-art datasets of human trajectories. Our analysis shows that THÖR has more variety in recorded behavior, less noise, and high duration of continuous tracking.

5.1.2 Outline

The chapter is organized as follows: in Sec. 5.2 we review the related work and in Sec. 5.3 detail the data collection procedure. In Sec. 5.4 we describe the recorded data and analyze it quantitatively and qualitatively. Sec. 5.5 concludes the chapter.

5.2 Related Work

Recordings of human trajectory motion and eye gaze are useful for a number of research areas and tasks both for machine learning and benchmarking, as we discussed In Chapter 2. Examples include person and group tracking [175, 194, 236], human-aware motion planning [19, 94, 229, 304], motion behavior learning [225], human motion prediction [63, 268], human-robot interaction [174], video surveillance [4] or collision risk assessment [196]. According to our taxonomy, state-of-the-art methods for tracking or motion prediction can incorporate information about the environment, social grouping, head orientation or personal traits. For instance, Lau et al. [175] estimate social grouping formations during tracking and our own method from Chapter 3 uses group affiliation as a contextual cue to predict future motion. Unhelkar et al. [317] use head orientation to disambiguate and recognize typical motion

[1] Available at `http://thor.oru.se`

patterns that people are following. Bera et al. [32] and Ma et al. [202] learn personal traits to determine interaction parameters between several people. To enable such research in terms of training data and benchmarking requirements, a state-of-the-art dataset should include this information.

Human trajectory data is also used for learning long-term mobility patterns [214], such as the CLiFF maps [166], to enable compliant flow-aware global motion planning and reasoning about long-term path hypotheses towards goals in distant map areas for which no observations are immediately available. Finally, eye-gaze is a critical source of non-verbal information about human task and motion intent in human-robot collaboration, traffic maneuver prediction, spatial cognition or sign placement [2, 52, 77, 146, 228].

Tables 2.2 – 2.4 in Chapter 2 review the existing datasets of human, vehicle and cyclist trajectories, commonly used in the literature. Here we extend this review, focusing on the pedestrian datasets and the recorded contextual cues in Table 5.1. With the exception of [48, 76, 349, 366], all datasets have been collected outdoors. Intuitively, patterns of human motion in indoor and outdoor environments are substantially different due to scope of the environment and typical intentions of people therein. Indoors people navigate in loosely constrained but cluttered spaces with multiple goal points and many ways (e.g. from different homotopy classes) to reach a goal. This is different from their behavior outdoors in either large obstacle-free pedestrian areas or relatively narrow sidewalks, surrounded by various kinds of walkable and non-walkable surfaces. Among the indoor recordings, only [76, 349] introduce a robot, navigating in the environment alongside humans. However, recording only from on-board sensors limits visibility and consequently restricts the perception radius. Furthermore, ground truth positions of the recorded agents in all prior datasets were estimated from RGB(-D) or laser data. On the contrary, we directly record the position of each person using a motion capture system, thus achieving higher accuracy of the ground truth data and complete coverage of the working environment at all times. Moreover, our dataset contains many additional contextual cues, such as social roles and groups of people, head orientations and gaze directions.

5.3 Data Collection Procedure

In order to collect motion data relevant for a broad spectrum of research areas, we have designed a controlled scenario that encourages social interactions between individuals, groups of people and with the robot. The interactive setup assigns social roles and tasks so as to imitate typical activities found in populated spaces such as offices, train stations, shopping malls or airports. Its goal is to motivate participants to engage into natural and purposeful motion behaviors as well as to create a rich variety of unscripted interactions. In this section we detail the system setup and the data collection procedure.

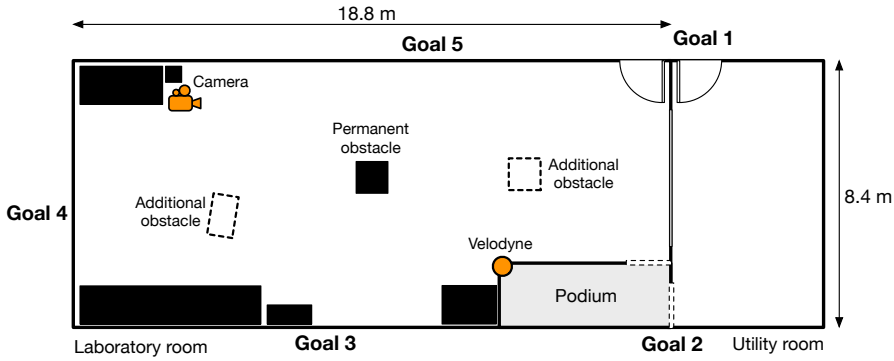


Figure 5.2: Overview of the environment. The Qualisys motion tracking system is installed in a laboratory room, which is mostly empty except for some shelves and equipment along the walls. A permanent obstacle in the middle of the room is present in all recordings, while additional obstacles are only placed in the "Three obstacles" scenario (see Sec. 5.3.2 for details). The position of the camera is shown in the top left corner, and the position of the Velodyne in the bottom right.

5.3.1 System Setup

Data collection was performed in a spacious laboratory room of 8.4×18.8 m and the adjacent utility room, separated by a glass wall (see the overview in Fig. 5.2). The laboratory room, where the motion capture system is installed, is mostly empty to allow for maneuvering of large groups, but also includes several constrained areas where obstacle avoidance and the choice of homotopy class is necessary. Goal positions are placed to force navigation along the room and generate frequent interactions in its center, while the placement of obstacles prevents walking between goals on a straight line.

To track the motion of the agents we used the Qualisys Oqus 7+ motion capture system[2] with 10 infrared cameras, mounted on the perimeter of the room. The motion capture system covers the entire room volume apart from the most right part close to the podium entrance – a negligible loss due to the focus on the central part of the room. The system tracks small reflective markers at 100 Hz with spatial discretization of 1 mm. The coordinate frame origin is on the ground level in the middle of the room. For people tracking, the markers have been arranged in distinctive 3D patterns on the bicycle helmets, shown in Fig. 5.3. The motion capture system was calibrated beforehand with an average residual tracking error of 2 mm, and each helmet, as well as the robot, was defined in the system as a unique rigid body of markers, yielding its 6D head position and orientation. Each participant was assigned an individual

[2] `https://www.qualisys.com/hardware/5-6-7/`

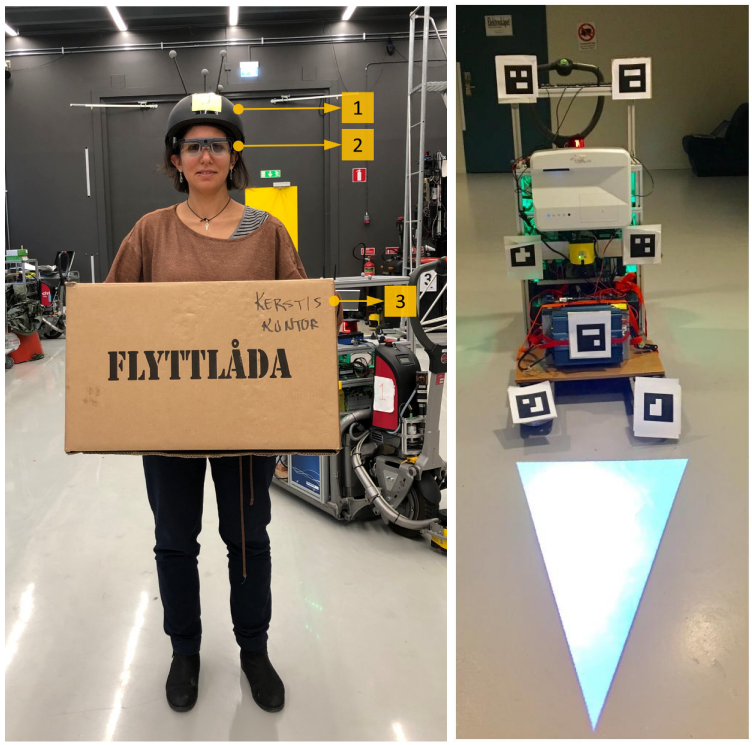


Figure 5.3: Equipment used in our data collection: **Left:** (1) bicycle helmet with mocap tracking markers, (2) Tobii Pro Glasses, (3) boxes which were carried by the participants as a part of the tasks. **Right:** Linde CitiTruck robot projecting its current motion intent on the floor.

helmet for all recording sessions, labeled 2 to 10. Helmet 1 was not used in this data collection.

For acquiring eye gaze data we used four mobile eye-tracking headsets worn by four participants (helmet numbers 3, 6, 7, and 9 respectively). However, in this dataset we only include data from one headset (Tobii Pro Glasses), worn by the participant with helmet 9. The gaze sampling frequency of Tobii Pro Glasses is 50 Hz. It also has a scene camera which records the video at 25 fps. A gaze overlaid version of this video is included in this dataset. We synchronized the clocks of each machine (the Qualisys system, the stationary Velodyne sensor and the eye-tracking glasses) with the same NTP time server. Finally, we recorded a video of the environment from a stationary camera, mounted in a corner of the room.

The robot, used in our data collection, is a small forklift Linde CitiTruck robot with a footprint of 1.56×0.55 m and 1.17 m high, shown in Fig. 5.3.

It was programmed to move in a socially unaware manner, following a predefined path around the room and adjusting neither its speed nor trajectory to account for surrounding people. For safety reasons, the robot was navigating with a maximal speed of $0.34\,\mathrm{m\,s^{-1}}$ and projecting its current motion intent on the floor in front of it using a mounted beamer [52]. A dedicated operator was constantly monitoring the environment from a remote workstation to stop the robot in case of an emergency. The participants were made aware of the emergency stop button on the robot should they be required to use it.

5.3.2 Scenario Description and Participants' Priming

During the data collection the participants performed simple tasks, which required walking between several goal positions. To increase the variety of motion, interactions and behavioral patterns, we introduced several roles for the participants and created individual tasks for each role, summarized in Fig. 5.4.

The first role is a *visitor*, navigating alone and in groups of up to 5 people between four goal positions in the room. At each goal they take a random card, indicating the next target. As each group was instructed to travel together, they only take one card at a time. We asked the visitors to talk and interact with the members of their group during the data collection, and changed the structure of groups every 4-5 minutes. There are 6 visitors in our recording. The second role is a *worker*, whose task is to receive and carry large boxes between the laboratory and the utility room. The workers wear a yellow reflective vest. There are 2 workers in our recording, one carrying the boxes from the laboratory to the unity room, and the other vice versa. The third role is the *inspector*. An inspector is navigating alone between many additional targets in the environment, indicated by a QR-code, in no particular order and stops at each target to scan the code. We have one inspector in our recording.

There are several points to motivate the introduction of the social roles. Firstly, with the motion of the visitors and the workers we introduce distinctive motion patterns in the environment, while the cards and the tasks make the motion focused, goal-oriented and prevent random wandering. However, the workers' tasks allocation is such that at some points idle standing/wandering behavior is also observed, embedded in their cyclical activity patterns. Furthermore, we expect that the visitors navigating alone, in groups and the workers who carry heavy boxes exhibit distinctive behavior, therefore the grouping information and the social role cue (reflective vest) may improve the intention and trajectory prediction. Finally, motion of the inspector introduces irregular patterns in the environment, distinct from the majority of the visitors.

We prepared three scenarios for data collection with different numbers of obstacles and motion state of the robot. In the first scenario, the robot is placed by a wall and not moving, and the environment has only one obstacle (see the layout in Fig. 5.2). The second scenario introduces the moving robot, navigat-

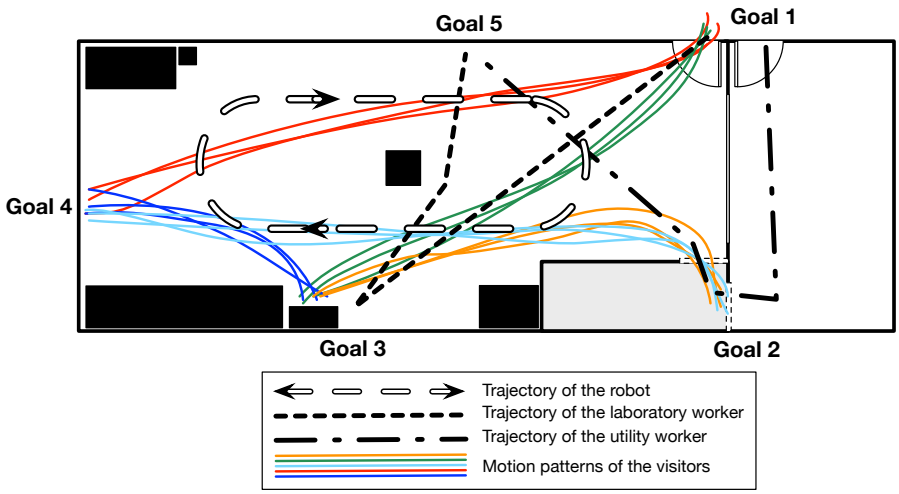


Figure 5.4: Roles of the participants and their expected motion patterns. *Visitors*, walking alone and in groups, are instructed to navigate between goals 1,2,3 and 4. Their motion patterns are shown with colored solid lines. The *laboratory worker*, whose waiting position is at goal 3, picks up an incoming box at goal 1, registers its ID at goal 3 and then places it at goal 5. The *utility worker*, whose waiting position is at goal 2, picks up the box at goal 5, registers it at goal 2 with a new ID and places it at goal 1. The patterns of both workers are shown with dotted lines. The trajectory of the *robot*, circulating around the obstacle in the middle of the room, is shown with a thick hollow line.

ing around the obstacle (the trajectory of the robot is depicted in Fig. 5.4). The third scenario features an additional obstacle and a stationary robot in the environment (see Fig. 5.2 with additional obstacles). We denote these recording scenarios as *One obstacle*, *Moving robot* and *Three obstacles*, accordingly. In each scenario the group structure for the visitors was reassigned 4-5 times. Between the scenarios, the roles were also reassigned. A summary of the scenarios and durations is given in Table 5.2.

Each round of data collection started with the participants, upon command, beginning to execute their tasks. The round lasted for approximately four minutes and ended with another call from the moderator. To avoid artificial and unnatural motion due to knowing the true purpose of the data collection, we told the participants that our goal is to validate the robot's perceptive abilities, while the motion capture data will be used to compare the perceived and actual positions of humans. Participants were asked not to communicate with us during the recording. For safety and ethical reasons, we have instructed participants to act carefully near the robot, described as "autonomous industrial

Scenario, round		Visitors, groups Helmet ID 2–10	Workers Utility,	lab	Inspector	Duration
One obstacle	1	6,7,5 + 8,2,4	3	9	10	368 sec
	2	2,5,6,7 + 8,4	3	9	10	257 sec
	3	6,7,8 + 4,5 + 2	3	9	10	275 sec
	4	2,4,5,7,8 + 6	3	9	10	315 sec
Moving robot	1	4,5,6 + 3,7,9	2	8	10	281 sec
	2	3,5,6,9 + 7,4	2	8	10	259 sec
	3	5,7,9 + 4,6 + 3	2	8	10	286 sec
	4	3,5,6,7,9 + 4	2	8	10	279 sec
	5	3,6 + 4,9 + 5,7	2	8	10	496 sec
Three obstacles	1	2,3,8 + 6,7,9	5	4	10	315 sec
	2	2,8,9 + 3,6,7	5	4	10	290 sec
	3	2,3,7 + 8,9 + 6	5	4	10	279 sec
	4	2,3,6,7,9 + 8	5	4	10	277 sec

Table 5.2: Role assignment and recording duration in the three scenarios of our data collection: (i) One obstacle, (ii) Moving robot, (iii) Three obstacles.

equipment" which does not stop if someone is in its way. An ethics approval was not required for our data collection as per institutional guidelines and the Swedish Ethical Review Act (SFS number: 2003:460). Written informed consent was obtained from all participants. Due to the relatively low weight of the robot and the safety precautions taken, there was no risk of harm to participants.

5.4 Results and Analysis

5.4.1 Data Description

The THÖR dataset includes over 60 minutes of motion in 13 rounds of the three scenarios. The recorded data in `.mat`, `.bag` and `.tsv` format contains over 395k frames at 100 Hz, 2531k human detections and 600+ individual and group trajectories between the goal positions. For each detected person the 6D position and orientation of the helmet in the global coordinate frame is provided. Furthermore, the dataset includes the map of the static obstacles, goal coordinates and grouping information. We also share the Matlab scripts for loading, plotting and animating the data. Additionally, the eye gaze data is available for one of the participants (Helmet 9), as well as Velodyne scans from a static sensor and the recording from the camera. We thoroughly inspected the motion capture data and manually cleaned it to remove occasional helmet ID switches and recover several lost tracks. Afterwards we applied an automated procedure to restore the lost positions of the helmets from incomplete set of recognized markers. In Fig. 5.5 we show the summary of the recorded trajectories.

5.4.2 Baselines and Metrics

The THÖR dataset is recorded using a motion capture system, which yields more consistent tracking and precise estimation of the ground truth positions and therefore higher quality of the trajectories, compared to the human detections from RGB-D or laser data, typically used in existing datasets. For the quantitative analysis of the dataset, we compare the recorded trajectories to the several datasets which are often used for training and evaluation of motion predictors for human environments, as discussed in Sec. 2.8.2. The popular ETH dataset [236] is recorded outdoors in a pedestrian zone with a stationary camera facing downwards and manually annotated at 2.5 Hz. The Hotel sequence, used in our comparison, includes the coordinates of the 4 common goals in the environment and group information for walking pedestrians. The ATC dataset [48] is recorded in a large shopping mall using multiple 3D range sensors at ~26 Hz over an area of 900 m^2. This allows for long tracking durations and potential to capture interesting interactions between people. In addition to positions it also includes facing angles. In this comparison we used the recordings from 24th and 28th of October and 14th of November. The Edinburgh dataset [205] is recorded in a university campus yard using a camera facing down with variable detection frequency, on average 9 Hz. For comparison we used the recordings from 27th of September, 16th of December, 14th of January and 22nd of June.

For evaluating the quality of recorded trajectories we propose several metrics:

1. *Tracking duration* (s): average length of continuous observations of a person, higher is better.

2. *Trajectory curvature* (m^{-1}): global curvature of the trajectory $\mathcal{T}$, caused by maneuvering of the agents in presence of static and dynamic obstacles, measured on 4 s segments based on the first $\mathcal{T}_t = (x_1, y_1)$, middle $\mathcal{T}_{t+2s} = (x_2, y_2)$ and last $\mathcal{T}_{t+4s} = (x_3, y_3)$ points of the interval: $K(\mathcal{T}_{t:t+4s}) = |\frac{2(x_2-x_1)(y_3-y_1)-(x_3-x_1)(y_2-y_1)}{||x_2-x_1,y_2-y_1||\ ||x_3-x_1,y_3-y_1||\ ||x_3-x_2,y_3-y_2||}|$. The choice of 4 s path segments is motivated by the standard motion prediction horizon in the related work [4]. Higher curvature values correspond to more challenging, non-linear paths.

3. *Perception noise* ($m\,s^{-2}$): under the assumption that people move on smooth, not jerky paths, we evaluate local distortions of the recorded trajectory $\{\mathcal{T}_t\}_{t=1...M}$ of length M, caused by the perception noise of the mocap system as the average absolute acceleration: $\frac{1}{M}\sum_{t=1}^{M}|\ddot{\mathcal{T}}_t|$. Less noise is better.

4. *Motion speed* ($m\,s^{-1}$): mean and standard deviation of velocities in the dataset, measured on 1 s intervals. If the effect of perception noise on

Metric	THÖR	ETH	ATC	Edinburgh
Tracking duration [s]	16.7 ± 14.9	9.4 ± 5.4	39.7 ± 64.7	10.1 ± 12.7
Trajectory curvature [m^{-1}]	1.9 ± 8.8	0.18 ± 1.48	0.84 ± 1.43	1 ± 3.9
Perception noise [$\mathrm{m\,s}^{-2}$]	0.12	0.19	0.48	0.81
Motion speed [$\mathrm{m\,s}^{-1}$]	0.81 ± 0.49	1.38 ± 0.46	1.04 ± 0.46	1.0 ± 0.64
Min. dist. between people [m]	1.54 ± 1.60	1.33 ± 1.39	0.61 ± 0.16	3.97 ± 3.5

Table 5.3: Comparison of the datasets

speed is negligible, higher standard deviation means more diversity in behavior of the observed agents, both in terms of individually preferred velocity and compliance with other dynamic agents.

5. *Minimal distance between people* (m): average minimal euclidean distance between two closest observed people. This metric indicates the density of the recorded scenarios, lower values correspond to more crowded environments.

5.4.3 Results

The results of the evaluation are presented in Table 5.3. Our dataset has sufficiently long trajectories (on average 16.7 s tracking duration) with high curvature values ($1.9 \pm 8.8\ \mathrm{m}^{-1}$), indicating that it includes more human-human and human-environment interactions than the existing datasets. Furthermore, despite the much higher recording frequency, e.g. 100 Hz (THÖR) vs. ~26 Hz (ATC), the amount of perception noise in the trajectories is lower than in all baselines. The speed distribution of $\pm 0.49\ \mathrm{m\,s}^{-1}$ shows that the range of observed velocities corresponds to the baselines, while the lower average velocity in combination with a high average curvature confirms higher complexity of the recorded behaviors, because comfortable navigation in straight paths with constant velocity is not possible in presence of static and dynamic obstacles. Finally, the high variance of the minimal distance between people (1.54 ± 1.60 m THÖR vs. 0.61 ± 0.16 m ATC) shows that our dataset features both dense and sparse scenarios, similarly to ETH and Edinburgh.

An important advantage of THÖR in comparison to the prior art is the availability of rich interactions between the participants and groups in presence of static obstacles and the moving robot. In this compact one hour recording we observe numerous interesting situations, such as accelerating to overtake another person; cutting in front of someone; halting to let a large group pass; queuing for the occupied goal position; group splitting and re-joining; choosing a sub-optimal motion trajectory from a different homotopy class due to a narrow passage being blocked; hindrance from walking towards each other in opposite directions. In Fig. 5.6 – 5.8 we illustrate several examples of such interactions.

5.5 Conclusions and Outlook

In this chapter we presented a novel human motion trajectories dataset, recorded in a controlled indoor environment. Aiming at applications in training and benchmarking human-aware intelligent systems, we designed the dataset to include a rich variety of human motion behaviors, interactions between individuals, groups and a mobile robot in the environment with static obstacles and several motion targets. Our dataset includes accurate motion capture data at high frequency, head orientations, eye gaze directions, data from a stationary 3D lidar sensor and an RGB camera. Using a novel set of metrics for the dataset quality estimation, we show that it is less noisy and contains higher variety of behavior than the prior art datasets.

The proposed data collection procedure has shown clear potential in generating interesting and diverse interactions. An obvious continuation of work on this topic is recording more data in various scenarios. The summary of the recorded trajectories in Fig. 5.5 shows that the nature of motion in a relatively simple indoor space is not only defined by the free and occupied space, but also by its topology, location of obstacles, goals and the presence of the robot. These are important factors, and learning to recognize their effect on human motion requires many more example scenarios with variations in each of those factors. The future recordings may include higher density of people, more eye-tracking headsets, more robots and variations in their behavior.

Furthermore, we are interested in building a benchmarking suite for trajectory prediction algorithms, rich in contextual cues and factor-conditioned experiments. The diverse nature of the recorded data allows building balanced training and validation datasets, and the length of uninterrupted observations is suitable for evaluation with long-term prediction horizons. On top of that, the accurate ground truth data can be augmented with controlled sensor noise to test robustness against imperfect sensing. These first steps are taken in Chapter 6, where we present our benchmark design and use the THÖR data in precision, robustness and transfer experiments, and in Chapter 7, where we discuss the further development of the benchmarking ideas.

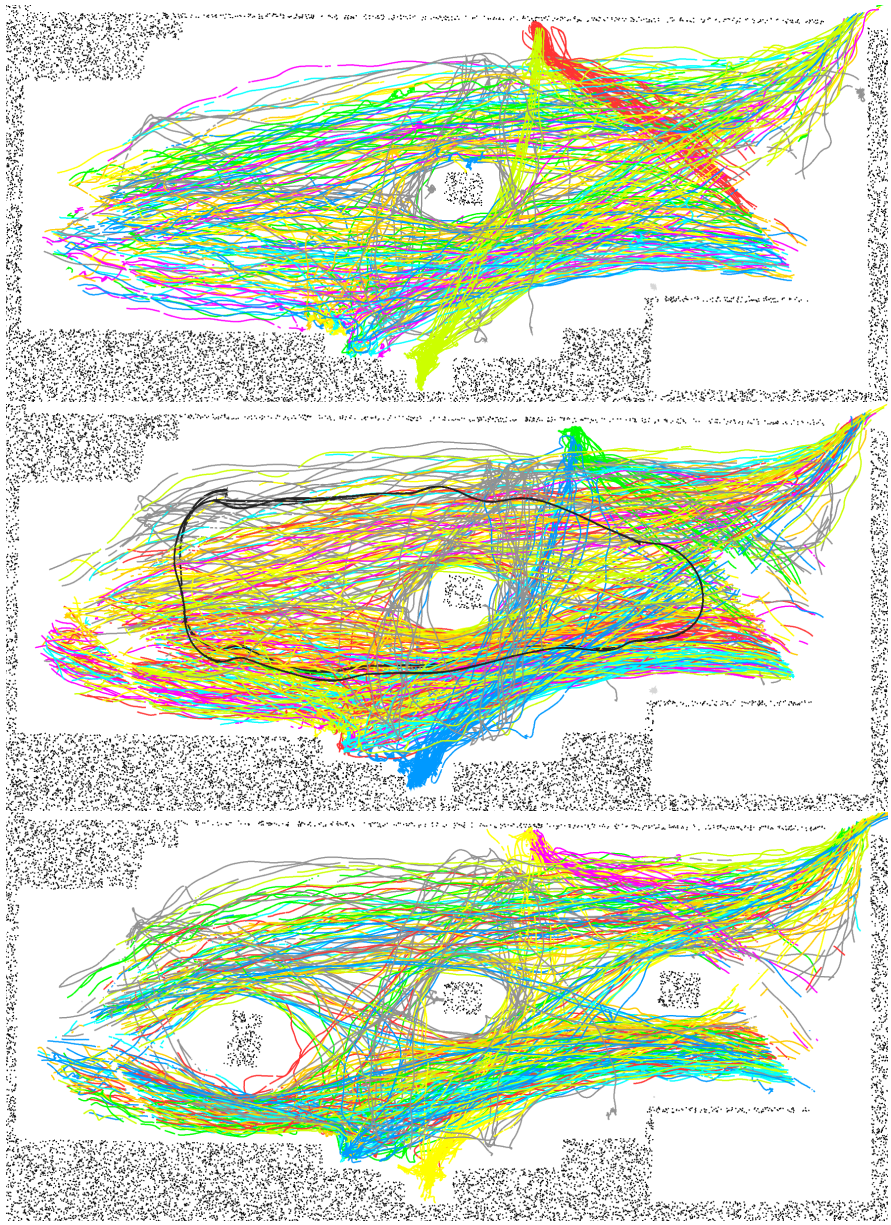


Figure 5.5: Trajectories of the participants and the robot, recorded in the "One obstacle" scenario (**top**), "Moving robot" scenario (**middle**) and "Three obstacles" scenario (**bottom**). The robot's path in the middle image is shown in **black**.

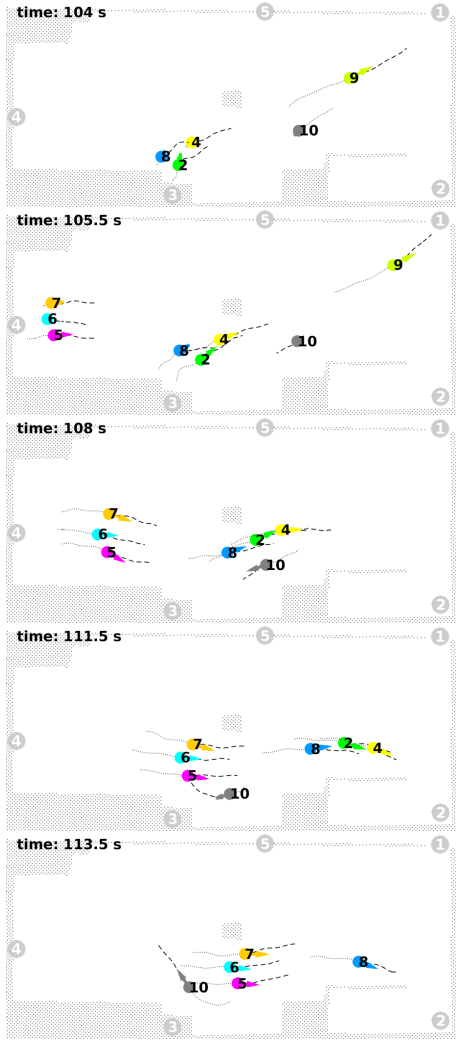


Figure 5.6: Social interactions in the THÖR dataset with color-coded positions of the observed people. The current velocity is shown with an arrow of corresponding length and direction. The past and the future 2 s trajectories are shown with dotted and dashed lines respectively. Goal locations are marked with gray circles. **"One obstacle", Round 1:** at 104 sec the group (2,4,8) starts moving from the goal point, taking the *line formation* in the constrained space due to the presence of standing person 10. Later, at 111.5 sec, person 10 has to adjust the path and slow down while the group (5,6,7) proceeds in the *V formation* [216], engaged in communication.

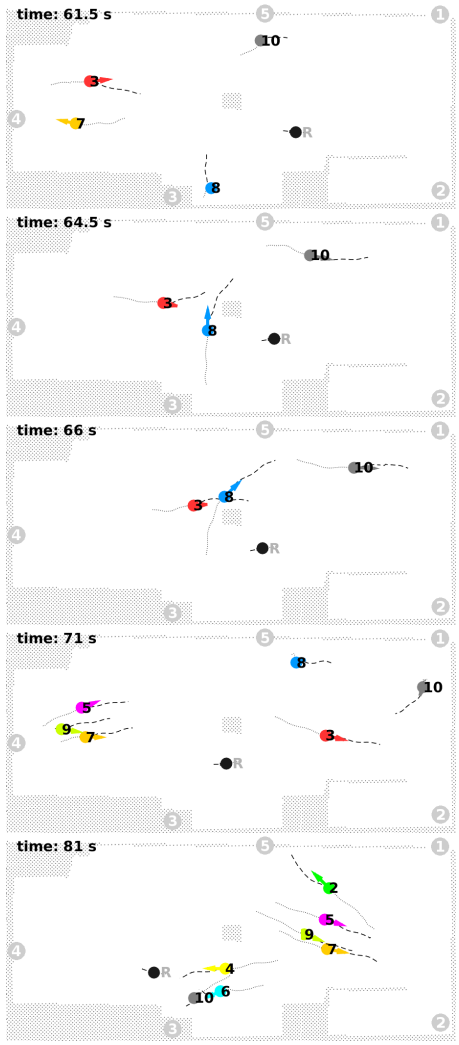


Figure 5.7: Social interactions in the THÖR dataset with color-coded positions of the observed people. The current velocity is shown with an arrow of corresponding length and direction. The past and the future 2 s trajectories are shown with dotted and dashed lines respectively. Goal locations are marked with gray circles. **"Moving robot", Round 3:** person 8 is leaving the resting position at 61.5 sec and adapts the path to account for the motion of the robot, taking a detour from the optimal way to reach the goal 5. At 66 seconds person 8 crosses person 3, who has to slow down, as compared to the velocity at time 61.5 and 71. The same maneuver of taking a detour due to the presence of the robot is performed by the group (5,7,9) at time 71.

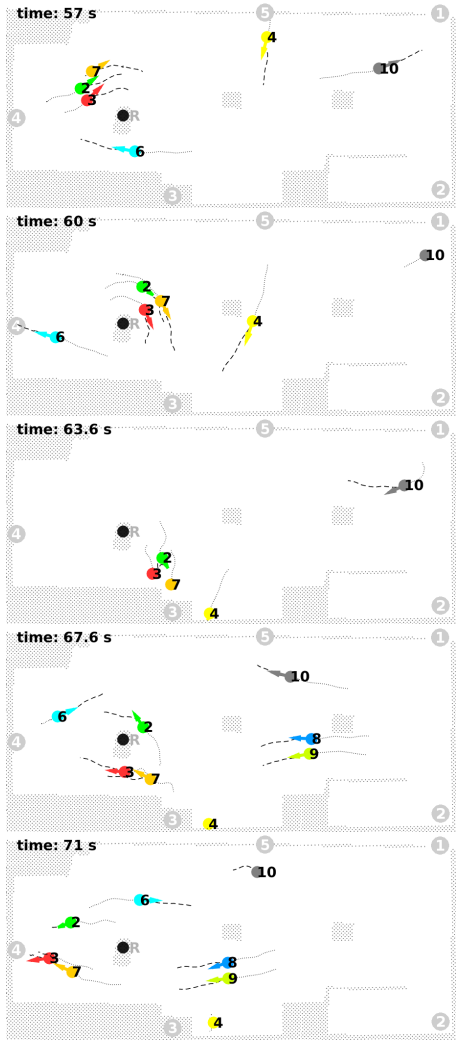


Figure 5.8: Social interactions in the THÖR dataset with color-coded positions of the observed people. The current velocity is shown with an arrow of corresponding length and direction. The past and the future 2 s trajectories are shown with dotted and dashed lines respectively. Goal locations are marked with gray circles. **"Three obstacles", Round 3:** Group (2,3,7), navigating in a constrained environment, at 57 sec has to make a detour around the obstacle while heading to goal 3. On the way back to goal 4 the group splits at 67.6 sec, and reunites later on.

Chapter 6
Benchmarking Human Motion Prediction Methods

Did I ever shorten what I ought to have lengthened? Did I only weaken what I could have strengthened?

Le Ceneri di Heliodoro
Jerome Reuter

With a multitude of new methods proposed by different communities, the lack of standardized benchmarking and objective comparison between them has been a major limitation for assessing the capabilities of the state-of-the-art systems. Existing attempts in building benchmarking infrastructure are limited in their support for relevant contextual cues, experiments and variation of prediction parameters. To advance the state of benchmarking, in this chapter we present a novel benchmark, built for thorough evaluation and comparison of the motion prediction methods along several axes. Our *Atlas benchmark* encompasses a large variety of heterogeneous datasets, representing usual human motion behaviors in different places and cultures. The benchmark offers tools, such as metrics, data preparation and filtering, calibration and visualization to overcome several limitations of the existing infrastructure, thus sustaining the enduring development of better algorithms. Using the benchmark, we investigate a hypothesis about a better local interaction model for our MDP-based motion predictor, presented in Chapter 3.

6.1 Introduction

Benchmarking motion prediction algorithms is not a trivial matter. The evaluation outcome can be affected by various factors, and the properties of the

methods can sometimes be exposed only in elaborate experiments. For instance, such factors include prediction horizon, which defines how far into the future predictions are made, and the procedure used to extract testing scenarios from raw datasets (streams of labeled detections). Even when evaluating the simplest constant velocity model using the same dataset, metrics and prediction horizon, the evaluation results still differ from each other, as we can observe in [4] and [284]. The reason here is that these two papers approach differently the testing scenarios generation and take different preprocessing steps for the raw data.

In Chapter 2 we revealed a clear demand for a good benchmark to allow systematic assessment of the prediction quality and formal comparison between methods. Among the methods, discussed in Chapter 2, evaluation has been a disorganized effort with little overlap or systematic comparison between works. Initiatives of various authors range from dropping comparison to the state of the art (or quantitative evaluation all-together) to extensive evaluation on proprietary closed access data, prohibiting a meaningful comparison to the published results by the community. The first significant milestone in benchmarking is the well-defined evaluation strategy, proposed by Alahi et al. [4] and later released as the TrajNet benchmark [276], which has been well received by the prediction community and replicated by many authors. The further release of TrajNet++ [160] marks notable progress towards the unification of benchmarking human motion predictions. This challenge-centered benchmark, however, has some limitations which we discuss and address in this chapter. TrajNet++ supports automated extraction of scenes with given observation and prediction lengths from a raw dataset. The extracted scenes, however, are guaranteed to include *only one trajectory* with the requested duration parameters. Furthermore, the challenge-based nature of the benchmark does not explicitly permit modification of important prediction parameters, restricting the evaluation to several pre-defined experiments. Crucially, as of this moment TrajNet++ does not support scenes with obstacles or semantic information about the environment, it has limited support for prediction uncertainty modeling, and can only visualize the trajectories aggregated in time, concealing the interactional aspects of human motion.

6.1.1 Contribution

In this chapter we present the Atlas benchmark as the first step towards automated benchmarking and evaluation of the motion prediction methods with systematic variation of parameters. Atlas includes heterogeneous datasets of human motion trajectories, and is capable of automatically extracting valid testing scenarios, interpolating, downsampling and smoothing the missing and noisy detections. Compared to the prior art, it offers many tunable parameters like the observation period and prediction horizon, import of semantic maps and other relevant information such as the coordinates of goals in the map,

evaluation of the probabilistic prediction results, and robustness testing with added noise to the original data. Unlike TrajNet++, our benchmark is especially suited for studying how prediction parameters influence the results, in contrast to fixing the main parameters for producing scores in a specific *challenge*.

Using this developed benchmark, and taking insight from the methodology, discussed in Chapter 3, in this chapter we set out to better investigate the collision avoidance methods, used for modeling the local interactions between people. In Sec. 2.4 we described the *reactive* and *predictive* models, the latter being hypothetically superior. In particular, the agents acting according to the classical social force model [110] would not attempt to avoid collisions until the social force takes effect in close proximity to another person. This leads to overconfident, reactive and unnatural motion. Predictive methods aim to solve this problem by the means of projecting the current dynamic state and reacting to the expected collisions in advance.

In our comparison we consider two popular predictive approaches, build on the foundation of the social forces, thus retaining all the benefits of the original approach. The model by Zanlungo et al. [356] extends the original social force with explicit collision prediction based on the repulsive potential at the time of the closest approach between two pedestrians. The model by Karamouzas et al. [141] also computes the time to a possible collision and the repulsive force based on the projected future positions, but differently accounts for the several upcoming collisions, sorted by their remoteness in time. Our thorough comparison in multiple experiments with varying observation lengths, prediction horizons, added noise, and transfer between datasets, shows no significant improvement when using the predictive models, despite their theoretical appeal.

6.1.2 Outline

In the following, we discuss the background on benchmarking motion prediction methods in Sec. 6.2 and describe our benchmark design in Sec. 6.3. We present the quantitative comparison of the local interaction models in Sec. 6.4 and conclude the chapter in Sec. 6.5.

6.2 Background

In this section we revisit the motion prediction problem formulation, define the main properties of a benchmark and analyze the existing ones.

Generally speaking, a trajectory prediction method aims to estimate future positions of a moving agent within a certain time horizon with a deterministic or stochastic state hypothesis, the latter potentially being also multi-modal. Often the problem is cast as a sequence prediction problem, solved either iteratively or jointly for several steps into the future. Typically, a motion predictor uses the current state of the agent (or a history of observed states), possibly

Benchmark	ETH/UCY [4], TrajNet [276]	TrajNet++ [160]	Atlas
Metrics	ADE, FDE	ADE, FDE, NLL, Top-k ADE and FDE, Ground truth & Prediction truth Collision	ADE, FDE, Top-k ADE and FDE, NLP
Obstacles and environment data	-	-	Semantic map, goal locations
Variable obs. and pred. lengths	-	-	✓
Dataset compatibility	-	Any dataset in json format	Any dataset in json format
Uncertainty in prediction	-	Discrete particle-based	Analytical, discrete grid- and particle-based
Robustness tests	-	-	Added noise to data
Hyperparameter optimization	-	-	SMAC3 [192] interface

Table 6.1: Benchmarks for human motion prediction

augmented with the current state of the environment (or history thereof). The most common state representation for an agent is with 2D coordinates. Similarly, the environment is represented by the states of other moving agents, a 2D map of static obstacles $\mathcal{M}$ and possibly also surface semantics $\mathbf{f}(\mathcal{M})$.

This formulation motivates the type of data, used for training and validation of the prediction models: a detection of person p at time t at position x, y (computed based on a fixed world frame) is minimally represented with a 4D (t, p, x, y) vector. Several datasets include such information, e.g. ETH [236], UCY [181], Edinburgh [205], ATC [48], SDD [262] and most recently THÖR [270], which we reviewed in Sec. 2.8.2 and 5.2. For evaluation of a motion predictor, a continuous flow of detections in a dataset is converted into *testing scenarios*, where all detections between two frames (*Observed track* in Fig. 6.1) are used as the observation history of length O, and the following T frames should be predicted and compared to the ground truth (GT) data (*Prediction horizon* in Fig. 6.1). Metrics used to this end include distances between predicted and GT positions (for instance: Average and Final Displacement Errors, ADE and FDE, Modified Hausdorff Distance, MHD) or probabilistic methods to evaluate predicted distributions over the future positions (for instance, Negative Log-Likelihood, NLL, or Negative Log-Probability, NLP). We already reviewed the commonly used metrics in Chapter 2, see Table 2.1.

This outlines the main parameters of the evaluation: the dataset used, the extraction strategy for a testing scenario, observation and prediction intervals O and T, and finally the adopted metrics. Variation in each of those parameters (for instance, extracting a different subset of scenarios from the same dataset) has the potential to may drastically alter the evaluation outcome. Therefore, a standardized evaluation protocol, or *benchmark*, is required to guarantee objective comparison of results published by different papers.

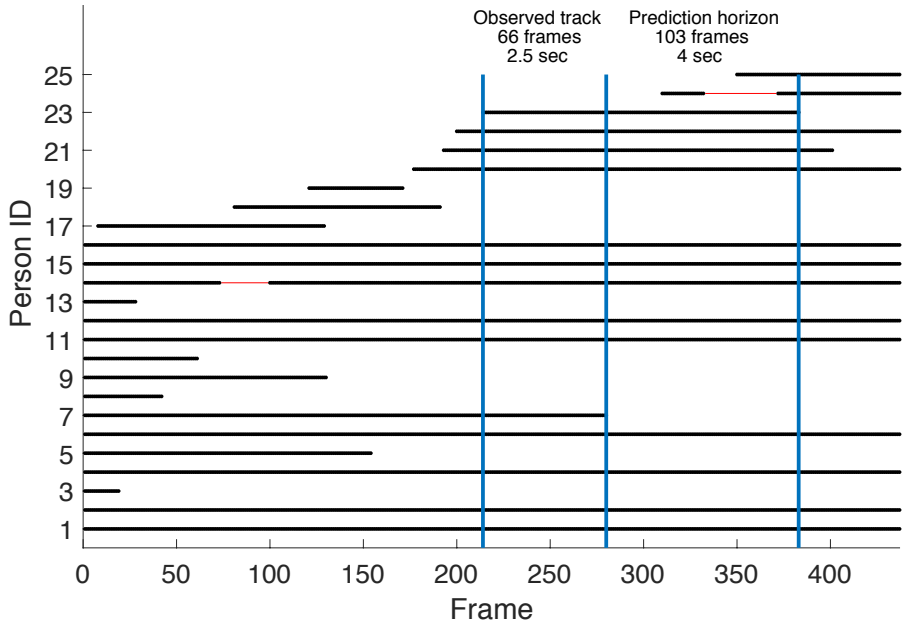


Figure 6.1: Data processing example in the ATC dataset. Horizontal tracks in **black** show available trajectory data (i.e. detections) for each frame on the x axis. Missing positions between detections, which can be interpolated, are shown with **red** lines. Vertical lines in **blue** show the observed track and prediction horizon. In this testing scenario, defined by the observed tracks, predictions are made for 14 people, but prediction for person 7 is excluded from evaluation as no ground truth positions are available.

The evaluation strategy, proposed by Alahi et al. [4], and adopted by many authors [11, 120, 123, 158, 222, 277, 345, 360, 362], fits this description to some extent. The authors propose to use the ETH and UCY datasets with fixed observation history O = 3.2 s and prediction horizon T = 4.8 s and the ADE and FDE geometric metrics. The testing scenario generation procedure is not defined, which may lead to different experiment implementations with different evaluation results, contradicting the purpose of a benchmark. For example, the UCY dataset contains trajectories in the form of continuous curves (splines), so discrete positions must be sampled. Furthermore, 8 frames of observations at 2.5 Hz may be interpreted as O = 3.2 *or* 2.8 s, as pointed out in [11]. The ETH-UCY evaluation strategy was slightly extended and formalized in the first *TrajNet* benchmark for motion prediction [276]. TrajNet does not include variability in the main parameters O and T, obstacles in the environment and any notion of prediction uncertainty or robustness.

An improved *TrajNet++* benchmark [160] by the same authors addresses the main issue – testing scenarios in this case are explicitly included in the benchmark. TrajNet++ uses several datasets, and potentially can be extended with further ones (stored in json format). It includes the possibility to predict several discrete positions for each pedestrian in each step, but does not support other probability distribution representations, such as the ones in Fig. 1.1. The main limitation here, however, is the rigidly defined testing parameters, which restrict the evaluation to the fixed observation history $O = 3.2$ s and prediction horizon $T = 4.8$ s. For proper quality assessment of a prediction method it is strictly preferable to benchmark its performance over a wide range of prediction horizons. Furthermore, the scenario extraction strategy only guarantees that in each scenario *one* target pedestrian has a complete track of requested $O + T$ consecutive positions. This contradicts the assumption, commonly made by many authors, that the history tracks for *all* pedestrians are available at the time of prediction [11, 25, 89, 308].

Based on these insights, in this chapter we present our novel *Atlas benchmark*. Atlas includes an automated procedure to extract testing scenarios from an arbitrary dataset with flexible O and T parameters, accepts occupancy and semantic maps as input, supports analytical and discrete uncertainty representation, and includes robustness experiments with added noise to the observed trajectories. We outline the properties of each benchmark in Table 6.1.

6.3 Our Benchmark Description

Fig. 6.2 outlines the design of our benchmark. The benchmark includes five main elements: data import, preprocessing, a prediction phase, evaluation and visualization tools. By explicitly interfacing the prediction module and scripting the experiments, our benchmark is suited for flexible and highly automated assessment of the motion prediction algorithms.

As the first step, the datasets and possibly additional information like the known goals in the environment, obstacle or semantic maps, are imported into the benchmark. Then, the raw data is preprocessed with downsampling to the user-defined frequency, interpolating the missing detections and trajectory smoothing. Once the dataset is ready, we can extract the testing scenarios with the user-specified observation and prediction horizons, as shown in Fig. 6.1. The observed histories of all people in the testing scenario, along with environment data, are explicitly interfaced as input to the prediction algorithm. The returned predictions are immediately evaluated against the ground truth using several metrics. Finally, the prediction results can be visualized with plots or animations. Meta-parameters to control the data processing and experiments are stored in a separate yaml file, eliminating the need to modify and re-compile the code.

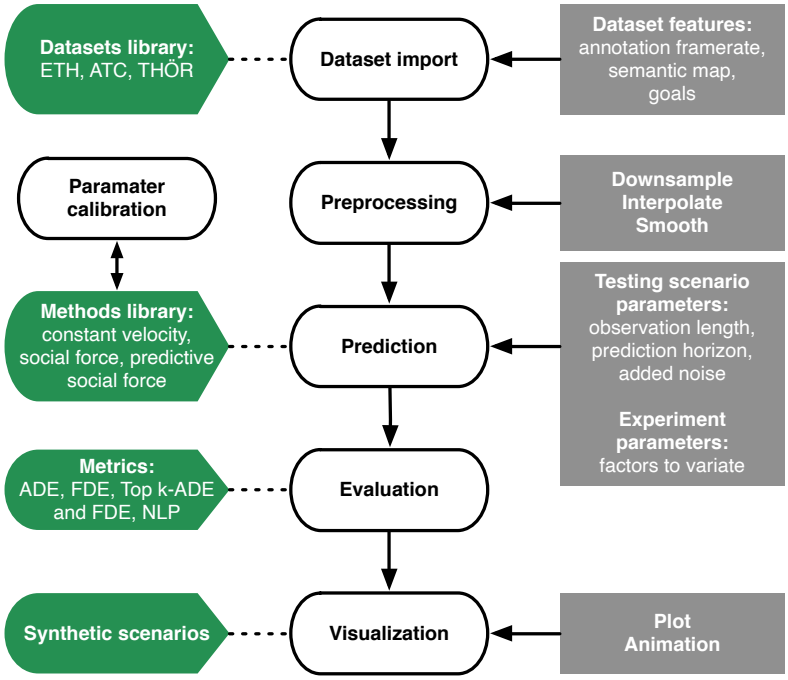


Figure 6.2: Altas benchmark design

In the following sections we describe each step of the benchmark in more detail.

6.3.1 Datasets

The benchmark users can import any dataset in the specific json file format, compatible with the TrajNet++ benchmark [160], which includes for each detection the time stamp, person id and position. The json dataset format also supports the obstacle and semantic grid maps, and the common goals in the environment, which may insight the possible destinations of people. Our benchmark currently includes the following three datasets:

i) **ETH** [236]: This dataset contains people detections from video data recorded outdoors in the ETH campus.

ii) **ATC** [48]: It is recorded in a shopping mall in Japan, representing therefore a large indoor environment with densely crowded scenes. Hence, there exist plenty of interactions.

iii) **THÖR** (Chapter 5 of this thesis): This dataset captures human motion in a room with static obstacles.

These three datasets come from different countries, taken in different environments, which increases the diversity of the data, and allows comparing the prediction methods on different social and cultural contexts.

6.3.2 Preprocessing

Raw datasets often include noise and annotation artifacts (e.g. missing detections). Hence, our benchmark offers interpolating and smoothing options in the preprocessing step. In addition, to check the robustness of implemented models, optional Gaussian noise may be added to each detection. Fig. 6.3 shows the preprocessing steps applied to one trajectory in the ATC dataset. After detecting the missing frames in the original trajectory based on the average annotation frequency, we interpolate the points in the missing part of the trajectory. Then, a moving average filter is used to smooth the noise. Finally, random noise distributed as $\mathcal{N}(0, \sigma^2)$, where σ is inversely proportional to the frame frequency, can be added to each detection.

After the data preprocessing, our benchmark generates the self-contained testing scenarios with the observation length O and ground truth for the following T frames, as shown in Fig. 6.1. As the prediction quality may strongly depend on the observation length (in particular for intention estimation and when the person detection is noisy), it is critical that all people in the testing scenario are observed in each of the O frames. A testing scenario, along with the environment information, is passed to the motion prediction step.

6.3.3 Prediction

Our benchmark offers a direct interface to the prediction module, which is called at this step for the given testing scenario. This allows highly automated evaluation with a systematic variation of parameters, defined at the previous steps.

Prior to benchmarking the prediction model on real data, the users can first validate their methods with several synthetic testing scenarios, created to stress the basic interaction modeling and obstacle avoidance showcase. These scenarios include various fundamental interactions between people and the environment, e.g. individuals and groups walking in the opposite directions, crossing paths and navigating around hindrances (see several examples in Fig. 6.4). For instance, Fig. 6.4 (top) shows two people walking on a collision course towards each other. Their velocities are $1\ \mathrm{m\,s^{-1}}$ and the initial displacement in the y axis is $0.2\ \mathrm{m}$. The frame frequency is $2.5\ \mathrm{Hz}$ and the observation length is 8 frames. Fig. 6.4 also includes example predictions made by two popular social force-based models [110, 356].

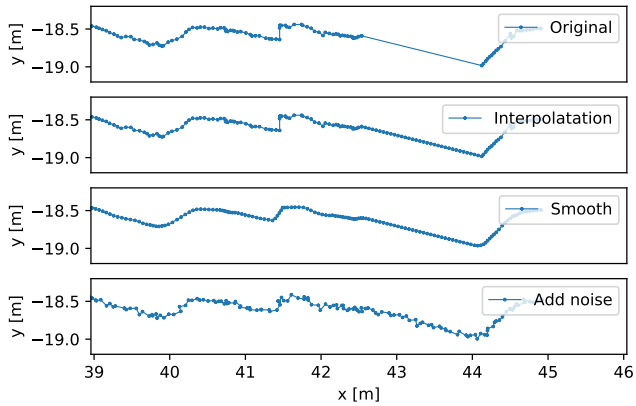


Figure 6.3: Example trajectory from the ATC dataset, which shows the noise and missing detections in the raw data (*original* trajectory on the top). Our benchmark offers interpolation and smoothing to fix this, followed by adding a controlled amount of noise to test the robustness of the prediction algorithm.

For optimizing the hyperparatemers of the prediction methods, such as [86, 110, 141, 150, 356], we implement an interface to the SMAC3 optimizer [192].

Our benchmark supports analytical, discrete and particle-based uncertainty representation for the prediction results. Discrete uncertainty is encoded in the grid map of the environment, separately for each person in each time step. Analytical uncertainty is represented with a mixture of Gaussians, as Fig. 6.5 shows. Particle-based uncertain predictions are represented with a set of discrete samples. These options allow evaluating most existing prediction algorithms.

6.3.4 Evaluation

To evaluate the performance of the various models, the benchmark offers geometric and probabilistic metrics.

Geometric metrics include the *Average Displacement Error* (ADE), which describes the error between points of predicted trajectory and the ground truth at the same time step, and the *Final Displacement Error* (FDE), which computes the error between the predicted and the ground truth position at the last prediction step.

Probabilistic metrics include the *Negative Log-Probability* (NLP), which computes the average probability of the ground truth position under the predicted distribution for the corresponding frame, and *Top-k ADE and FDE*, which compute the displacements between the ground truth position and the closest of the k samples from the probability distribution.

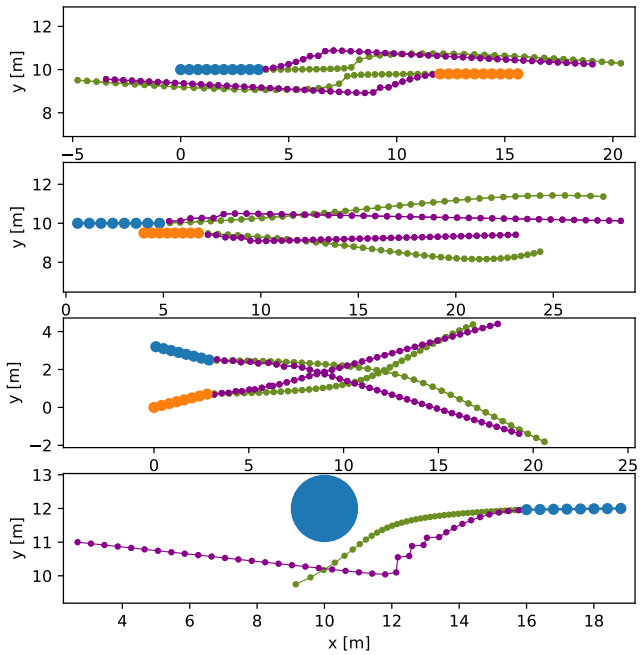


Figure 6.4: Synthetic testing scenarios. The subfigures from top to bottom show the *opposing*, *chasing*, *crossing* and *avoiding an obstacle* scenarios. The **blue** and **orange** dotted lines show the observations of the two people in each scenario, the **green** dotted lines show the social force predictions [110], and the **purple** dotted lines illustrate the predictive social force result [356].

6.3.5 Experiments

The benchmarking methodology described so far, e.g. the datasets, metrics and pre-processing steps, is fairly straightforward, although not systematically implemented in the prior art. On top of that, in our benchmark we propose a set of experiments to study the prediction performance under the influence of various factors. Under the word *experiment* we understand measuring the target evaluation metric on a set of validation scenarios under specific conditions (*parameters of the experiment*). These experiments allow systematic validation of parameters and help the users to gain a deeper insight into the applicability of the methods, in contrast to a limited insight contained in a single benchmark score. Due to the automated nature of our benchmark, the experiments are scripted with all parameters available externally in a yaml file.

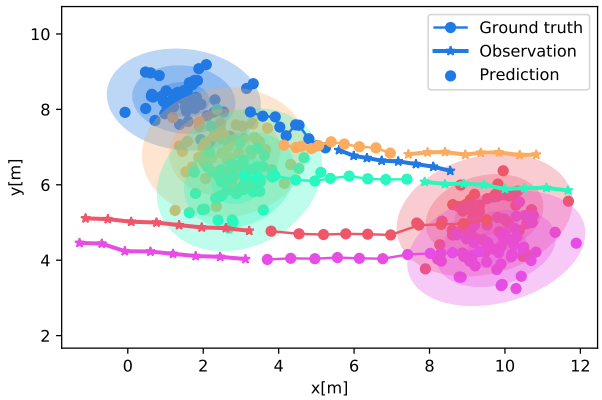


Figure 6.5: A testing scenario from the ATC dataset. This figure shows the probability distributions over the final positions of the observed people. Each person's past track and ground truth future motion is shown with a different color. Ellipses show the *analytical* and points show the *particle-based* probability distribution over the final position.

Prediction accuracy conditioned on parameters

Observation length and prediction horizon are among the main factors, associated with predicting motion. The prediction quality naturally degrades for further time instances, whereas longer observation lengths may improve it overall. In Atlas it is possible to measure the accuracy of prediction conditioned on these two main parameters. We intend to add more conditioned experiments in the future, e.g. based on the number of people in the scenario.

Transfer experiment

A crucial part of evaluating a prediction method is testing its applicability in new environments outside the training data. Surprisingly, this capability is most often overlooked in evaluation sections. In Atlas it is possible to script hyperparameter optimization in one dataset, and evaluate the resulting method in another. In the future we plan to extend this capability to training.

Robustness experiment

For a system working in the real world, perception of the positions of people is often prone to noise. Therefore the predictor must be robust to noisy input. One possible way to quantify robustness is by measuring accuracy on the testing scenarios, artificially adding increasing amounts of noise to the initially noise-free data. We implement this experiment in Atlas by adding noise to the smoothed trajectories.

6.4 Case Study: Benchmarking Local Interaction Models

Our MDP-based predictor, presented in Chapter 3, has shown promising results for predicting the interactions with the use of the social force model [216]. A reasonable and popular choice due to its reliability, performance and simple implementation, the social force model is one of many existing collision avoidance models. Among its main drawbacks is inherent reactivity: the agents engage in passive collision avoidance only when in close proximity for the social forces to take effect. In reality, people adapt their trajectories to avoid collisions in advance. To correct this sort of behavior, the social force theory was extended with explicit collision prediction by a number of authors, as we reviewed in Sec. 2.4. These improvements work well in theory, see for instance Fig. 6.4, but they were in fact never validated with real motion trajectories.

In this section we use the experiments in Atlas to compare the vanilla social force (*Sof*) with two popular predictive extensions: the model by Zanlunlgo et al. [356], abbreviated as *Zan* in plots and tables, and the model by Karamouzas et al. [141], abbreviated as *Kara*. As a baseline, we add the linear velocity model (*Lin*), implemented as average velocity in the observed track, and constant velocity model (*CVM*), implemented as forward propagating the last observed motion state. In the following sections these two prospective methods [141, 356] are briefly described.

6.4.1 Predictive Social Force Model [356]

The model by Zanlungo et al. [356] extends the original social force with explicit collision prediction based on the repulsive potential at the predicted time of the closest approach between two pedestrians. In the original social force model, as illustrated in Fig. 3.3 and 6.6, the interaction force between two pedestrians is calculated based on their current positions. Instead, in Zanlungo's model the interaction force is based on the future state where these two pedestrians are projected to come to the closest approach. The algorithm to obtain $\mathbf{f}_{i,k}^{soc}$ in Eq. 3.9 is elaborated as follows:

1. Evaluate whether the two people i and k, with $r_{i,k}$ being the sum of their radii, will collide with each other using $\theta_{i,k}$, which is the angle between the relative position $\mathbf{d}_{i,k}$ and relative velocity $\mathbf{v}_{i,k}$. If $|\theta_{i,k}|$ is smaller than $\pi/4$, we assume these two pedestrians may collide, so the time $t_{i,k}$ to the closest approach is computed with Eq. 6.1; otherwise we set $t_{i,k} = \infty$.

$$t_{i,k} = \begin{cases} \cos\theta_{i,k} \dfrac{|\mathbf{d}_{i,k}| - r_{i,k}}{|\mathbf{v}_{i,k}|} & |\mathbf{d}_{i,k}| > r_{i,k} \\ 0.2 & |\mathbf{d}_{i,k}| \leqslant r_{i,k} \end{cases} \tag{6.1}$$

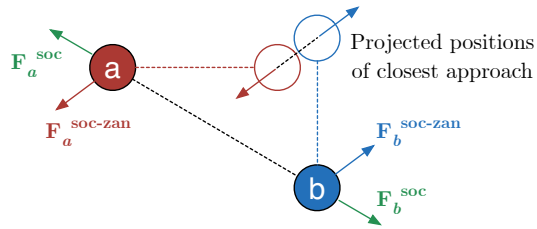


Figure 6.6: The force projection principle in the model by Zanlungo et al. [356]. Unlike the repulsion vectors F_a^{soc} and F_b^{soc} in the original social force model [110], depicted here in **green**, the model by Zanlungo et al. [356] calculates the repulsion forces $F_a^{soc-zan}$ and $F_b^{soc-zan}$ based on the point of closest approach.

2. Compute $t_{i,k}$ for each neighboring pedestrian and obtain $t_i = min_k\{t_{i,k}\}$. Similarly, we obtain $t_{i,o}$ for obstacles.

3. When the minimal t_i is obtained, we calculate the projected relative position $\mathbf{d}'_{i,k}$ for each pedestrian in Eq. 6.2. Here, all pedestrians are assumed to move forward with current velocity. The interaction force is computed based on the projected relative position and the minimal meeting time t_i with Eq. 6.3, where $\mathbf{n}'_{i,k}$ is the direction unit vector of the projected relative position.

$$\mathbf{d}'_{i,k} = \mathbf{p}_i + \mathbf{v}_i t_i - (\mathbf{p}_k + \mathbf{v}_k t_i). \tag{6.2}$$

$$\mathbf{f}^{soc}_{i,k} = \omega a_z \frac{v_i}{t_i} e^{-(d'_{i,k} - r_{i,k})/b_z} \mathbf{n}'_{i,k} \tag{6.3}$$

Compared with the Eq. 3.8 of the original social force model, the extra term v_i/t_i means the pedestrian is aiming to stop in t_i seconds, thus the force to avoid collision is proportional to v_i/t_i. In case the calculated t_i is too small, we introduce a minimal boundary for t_i to prevent enormous interaction force in our implementation (see Eq. 6.1).

The original Zanlungo's model in [356] designed for pedestrian-to-pedestrian interactions. In our implementation it is extended to account for the static environment obstacles. The total force on pedestrian i is expressed in Eq. 6.4, where the new term $\mathbf{f}^{soc}_{i,o}$ and $\mathbf{f}^{phys}_{i,o}$ respectively mean the social force and physical contact force between the nearest obstacle point and the target pedestrian.

$$\mathbf{F}_i = \mathbf{F}^{pers}_i + \mathbf{F}^{soc}_i + \mathbf{F}^{phys}_i = \mathbf{F}^{pers}_i + \sum_{j \neq i} \mathbf{f}^{soc}_{i,j} + \sum_{j \neq i} \mathbf{f}^{phys}_{i,j} + \mathbf{f}^{soc}_{i,o} + \mathbf{f}^{phys}_{i,o} \tag{6.4}$$

The social force from the obstacle is calculated in Eq. 6.5. The formulation is the same as in Eq. 6.3, but the obstacle avoidance parameters a_{zo} and b_{zo} are different from a_z and b_z (avoidance of moving people).

$$\mathbf{f}_{i,o}^{soc} = \omega a_{zo} \frac{v_i}{t_{io}} e^{-\left(d'_{i,o} - r_{i,o}\right)/b_{zo}} \mathbf{n}'_{i,o} \tag{6.5}$$

The physical force $\mathbf{f}_{i,o}^{phys}$ is calculated as in Eq. 6.6, which is similar to its formulation in the original social force. The differences are that $d'_{i,o}$ is the distance between obstacle and future position of the pedestrian, and $\mathbf{n}'_{i,o}$ is the unit vector that points from the obstacle to the future position of the pedestrian.

$$\mathbf{f}_{i,o}^{phys} = K_o g\left(r_{i,o} - d'_{i,o}\right) \mathbf{n}'_{i,o} \tag{6.6}$$

6.4.2 Predictive Collision Avoidance Model [141]

Similarly to the model, described in the previous section, the model by Karamouzas et al. [141] also computes the time to a possible collision and the repulsive force based on the projected future positions. Differently, this model constructs a collision set CP_i which stores the potential collision time $t_{i,j}$ with other pedestrians in the order of increasing collision time and accounts for the first N pedestrians. Besides, the collision time $t_{i,j}$ is calculated to a *safe distance* ρ instead of the *closest* approach, and the formula to calculate $\mathbf{f}_{i,k}^{soc}$ has a different specification. The collision set CP_i and force $\mathbf{F}_i^{soc}$ are calculated as follows:

1. Calculating the potential collision time $t_{i,j}$ requires finding the time before reaching the distance $d_{i,j}$, so we need to find the values of t subject to Eq. 6.7.

$$\begin{aligned} d_{i,j} &= \|\mathbf{p}'_j - \mathbf{p}'_i\| \\ &= \|\mathbf{p}_j + \mathbf{v}_j t - (\mathbf{p}_i + \mathbf{v}_i^{des} t)\| \\ &= \rho_i + r_j \end{aligned} \tag{6.7}$$

   Here, $\mathbf{p}'_j$ and $\mathbf{p}'_i$ are the projected future positions while $\mathbf{p}_j$ and $\mathbf{p}_i$ are present positions, ρ_i is the safe distance of pedestrian i and r_j is the radius of pedestrian j. Desired velocity $\mathbf{v}_i^{des}$ is calculated as

$$\mathbf{v}_i^{des} = \mathbf{v}_i + \frac{\mathbf{F}_g}{m} \Delta t, \tag{6.8}$$

   where v_i is the present velocity of pedestrian i, $\mathbf{F}_g$ is the intended force, m is the mass of pedestrian and Δt is the time interval of each prediction step.
   The following solutions of Eq. 6.7 are possible:

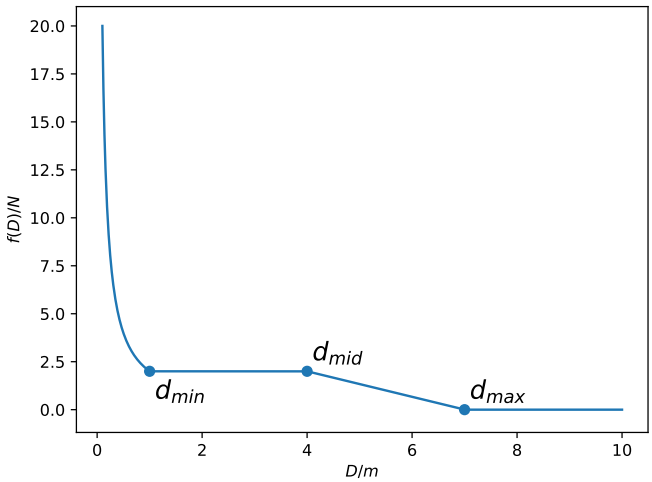


Figure 6.7: The distance-force relation in the model by Karamouzas et al. [141]

- No solution or single solution: no collision happens, therefore $t_{i,j}$ is infinite.
- Two negative solutions: this is a past collision so we can also set $t_{i,j}$ to infinity.
- One positive solution and one negative solution: this means a collision is close at hand, so we set $t_{i,j}$ to a small value, e.g. $t_{i,j} = 0.1s$, and add the time to the collision set CP_i.
- Two positive solutions: we take the smaller solution as collision time $t_{i,j}$ and add it into set CP_i.

2. Sort the collision set CP_i in the order of increasing time and choose first N pedestrians.

3. Calculate the evasive force $\mathbf{f}_{i,j}$ from pedestrian j to pedestrian i based on a piecewise function $f(D)$, where $D = \|\mathbf{p}'_i - \mathbf{p}_i\| + (\|\mathbf{p}'_i - \mathbf{p}'_j\| - r_i - r_j)$ is the displacement of pedestrian i at time $t_{i,j}$ plus the relative distance between i and j at the time $t_{i,j}$. The piecewise function is shown in Fig. 6.7, its three parameters d_{min}, d_{mid} and d_{max} control the distance-force relation.

	Prediction horizon			
Methods	1.6 s	3.2 s	4.8 s	8 s
CVM	0.10 ± 0.05	0.23 ± 0.13	0.40 ± 0.23	0.84 ± 0.57
LIN	0.17 ± 0.09	0.34 ± 0.19	0.55 ± 0.35	1.02 ± 0.69
Sof	0.10 ± 0.05	0.23 ± 0.12	0.39 ± 0.20	0.78 ± 0.45
Zan	0.10 ± 0.05	0.23 ± 0.12	0.38 ± 0.19	0.76 ± 0.44
Kara	0.11 ± 0.06	0.23 ± 0.11	0.38 ± 0.19	0.75 ± 0.44

Table 6.2: ADE in the ETH dataset with different prediction horizons

	Prediction horizon			
Methods	1.6 s	3.2 s	4.8 s	8 s
CVM	0.19 ± 0.10	0.50 ± 0.28	0.90 ± 0.54	1.96 ± 1.44
LIN	0.29 ± 0.16	0.66 ± 0.39	1.13 ± 0.73	2.25 ± 1.62
Sof	0.19 ± 0.09	0.49 ± 0.26	0.85 ± 0.45	1.72 ± 1.12
Zan	0.19 ± 0.10	0.49 ± 0.26	0.85 ± 0.44	1.67 ± 1.08
Kara	0.20 ± 0.10	0.49 ± 0.25	0.85 ± 0.44	1.67 ± 1.07

Table 6.3: FDE in the ETH dataset with different prediction horizons

	Prediction horizon			
Methods	1.6 s	3.2 s	4.8 s	8 s
CVM	0.15 ± 0.09	0.38 ± 0.24	0.71 ± 0.45	1.51 ± 0.91
LIN	0.29 ± 0.18	0.60 ± 0.38	0.99 ± 0.63	1.84 ± 1.08
Sof	0.18 ± 0.10	0.36 ± 0.20	0.60 ± 0.35	1.13 ± 0.67
Zan	0.15 ± 0.09	0.34 ± 0.20	0.59 ± 0.36	1.16 ± 0.70
Kara	0.16 ± 0.08	0.35 ± 0.19	0.60 ± 0.36	1.16 ± 0.69

Table 6.4: ADE in the THÖR dataset ("One obstacle" scenario) with different prediction horizons

4. Compute the total evasive force $\mathbf{F}_i^{soc}$ as a weighted sum of the single evasive forces, according to Eq. 6.9.

$$\mathbf{F}_i^{soc} = \sum_j^N w_{ij} \mathbf{f}_{ij} \tag{6.9}$$

The weighting factor w_{ij} is set inversely proportional to $t_{i,j}$, i.e. the force is larger for those pedestrians who have collisions closer in time.

Similarly to [356], we extend this model to consider static obstacles. The social force to the nearest obstacle is calculated in a similar way to the social force among pedestrians, except the evasive force in the third step is obtained with a different set of method parameters $d_{min,o}$, $d_{mid,o}$ and $d_{max,o}$.

	Prediction horizon			
Methods	1.6 s	3.2 s	4.8 s	8 s
CVM	0.28 ± 0.18	0.86 ± 0.54	1.64 ± 1.05	3.54 ± 2.11
LIN	0.49 ± 0.31	1.20 ± 0.75	2.07 ± 1.30	3.97 ± 2.27
Sof	0.29 ± 0.16	0.72 ± 0.42	1.27 ± 0.79	2.48 ± 1.54
Zan	0.26 ± 0.16	0.72 ± 0.43	1.31 ± 0.82	2.62 ± 1.61
Kara	0.28 ± 0.15	0.73 ± 0.42	1.31 ± 0.82	2.59 ± 1.59

Table 6.5: FDE in the THÖR dataset ("One obstacle" scenario) with different prediction horizons

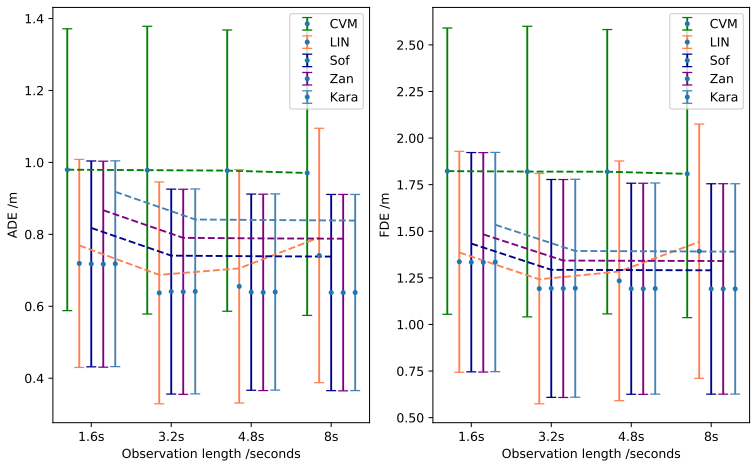


Figure 6.8: ADE/FDE in the ATC dataset with different observation lengths using Gaussian filter as initial velocity filter and without smoothing

6.4.3 Results and Discussion

Tables 6.2–6.5 show the results of experimenting with different prediction horizons. In general, and not surprisingly, the social force models outperform the linear velocity variants with lower displacement errors, and show more stable performance with lower standard deviations. However, we did not find a substantial difference between Sof, Kara and Zan in any of the datasets and on any of the prediction horizons.

Similarly, in experiments with different observation horizons we found no difference between the models. Interestingly, if the observations have low levels of noise, observing additional frames does not improve the performance, see a comparison between the noisy ATC dataset and noise-free THÖR in Fig. 6.8 and 6.9.

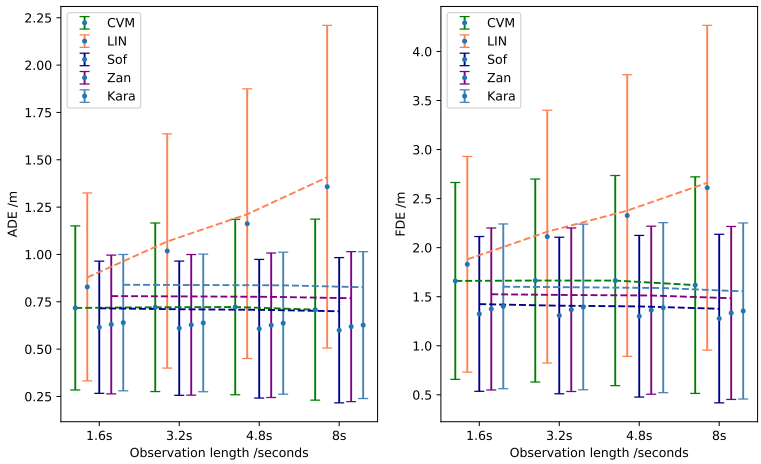


Figure 6.9: ADE/FDE in the THÖR dataset ("Three obstacles" scenario) with different observation lengths

Tables 6.6 and 6.7 summarize the transfer experiment, where the methods are calibrated on one dataset and tested on another. Also in this case we did not find that one of the three social force models exhibits superior transferability.

Finally, in Fig. 6.10 and 6.11 we show the robustness experiment, where we measure performance in presence of noise. While all social force models have excellent performance, on the level of the very simple and therefore very robust constant velocity model, the predictive variants do not outperform here either.

6.5 Conclusions and Outlook

Benchmarking motion prediction is no easy matter, and in this chapter we have shown the sheer amount of experiments required to make an in-depth analysis of performance. In reality, the summary given in Sec. 6.4 is but a fraction of experiments we conducted in Atlas to benchmark these local interaction models. In the M.Sc. thesis of Wanting Huang the reader may find further details on the design of the benchmark and its capabilities.

In the future work we plan to release Atlas implementation in Python as a tool to work with THÖR and other datasets. Among the possible extensions, as noted throughout this chapter, are additional metrics, other types of scripted experiments and an interface to train pattern-based methods in the transfer experiments. Furthermore, we revisit and extend our analysis of benchmarking in the ongoing work, outlined in Sec. 7.3.2.

		Test				
	Dataset	ETH	HOTEL	ATC	THÖR1	THÖR3
Calibrate	ETH	CVM: 0.40 ± 0.23	0.20 ± 0.16	0.50 ± 0.12	0.83 ± 0.43	0.79 ± 0.41
		LIN: 0.55 ± 0.35	0.25 ± 0.22	0.56 ± 0.16	1.14 ± 0.59	1.10 ± 0.56
		Sof: 0.39 ± 0.20	0.21 ± 0.16	0.50 ± 0.12	0.74 ± 0.37	0.75 ± 0.38
		Zan: 0.38 ± 0.19	0.20 ± 0.16	0.50 ± 0.12	0.71 ± 0.37	0.70 ± 0.38
		Kara: 0.38 ± 0.19	0.21 ± 0.16	0.50 ± 0.12	0.72 ± 0.38	0.72 ± 0.38
	HOTEL	Sof: 0.40 ± 0.23	0.20 ± 0.16	0.50 ± 0.12	0.83 ± 0.43	0.79 ± 0.41
		Zan: 0.40 ± 0.23	0.20 ± 0.16	0.50 ± 0.12	0.82 ± 0.43	0.79 ± 0.41
		Kara: 0.40 ± 0.23	0.20 ± 0.16	0.50 ± 0.12	0.83 ± 0.43	0.79 ± 0.41
	ATC	Sof: 0.40 ± 0.23	0.20 ± 0.16	0.50 ± 0.12	0.82 ± 0.43	0.79 ± 0.41
		Zan: 0.40 ± 0.23	0.20 ± 0.16	0.50 ± 0.12	0.82 ± 0.43	0.79 ± 0.41
		Kara: 0.40 ± 0.23	0.20 ± 0.16	0.50 ± 0.12	0.82 ± 0.43	0.79 ± 0.41
	THÖR1	Sof: 0.39 ± 0.20	0.21 ± 0.16	0.50 ± 0.12	0.71 ± 0.36	0.70 ± 0.37
		Zan: 0.38 ± 0.20	0.20 ± 0.16	0.50 ± 0.12	0.71 ± 0.37	0.69 ± 0.37
		Kara: 0.38 ± 0.20	0.21 ± 0.16	0.50 ± 0.12	0.71 ± 0.37	0.70 ± 0.38
	THÖR3	Sof: 0.39 ± 0.21	0.20 ± 0.16	0.50 ± 0.12	0.71 ± 0.36	0.68 ± 0.37
		Zan: 0.38 ± 0.20	0.20 ± 0.16	0.50 ± 0.12	0.71 ± 0.37	0.68 ± 0.37
		Kara: 0.38 ± 0.20	0.21 ± 0.16	0.50 ± 0.12	0.72 ± 0.38	0.70 ± 0.38

Table 6.6: ADE measured in the transfer experiments on different datasets. THÖR1 abbreviates the first "One obstacle" scenario in THÖR, and THÖR3 the third one ("Three obstacles").

		Test				
	Dataset	ETH	HOTEL	ATC	THÖR1	THÖR3
Calibrate	ETH	CVM: 0.90 ± 0.54	0.43 ± 0.40	1.03 ± 0.28	1.90 ± 0.98	1.85 ± 0.94
		LIN: 1.13 ± 0.73	0.49 ± 0.50	1.08 ± 0.33	2.37 ± 1.24	2.31 ± 1.15
		Sof: 0.86 ± 0.45	0.44 ± 0.40	1.03 ± 0.28	1.53 ± 0.82	1.54 ± 0.83
		Zan: 0.85 ± 0.44	0.43 ± 0.40	1.03 ± 0.28	1.54 ± 0.83	1.54 ± 0.84
		Kara: 0.85 ± 0.44	0.44 ± 0.40	1.03 ± 0.28	1.51 ± 0.84	1.52 ± 0.82
	HOTEL	Sof: 0.90 ± 0.54	0.43 ± 0.40	1.03 ± 0.28	1.90 ± 0.98	1.85 ± 0.94
		Zan: 0.90 ± 0.54	0.43 ± 0.40	1.03 ± 0.28	1.90 ± 0.98	1.84 ± 0.94
		Kara: 0.90 ± 0.54	0.43 ± 0.40	1.03 ± 0.28	1.90 ± 0.98	1.85 ± 0.94
	ATC	Sof: 0.90 ± 0.54	0.43 ± 0.41	1.03 ± 0.28	1.90 ± 0.98	1.85 ± 0.94
		Zan: 0.90 ± 0.54	0.43 ± 0.40	1.03 ± 0.28	1.90 ± 0.98	1.84 ± 0.94
		Kara: 0.90 ± 0.54	0.43 ± 0.40	1.03 ± 0.28	1.90 ± 0.98	1.85 ± 0.94
	THÖR1	Sof: 0.86 ± 0.45	0.44 ± 0.40	1.03 ± 0.28	1.51 ± 0.81	1.51 ± 0.81
		Zan: 0.85 ± 0.45	0.43 ± 0.40	1.03 ± 0.28	1.55 ± 0.84	1.54 ± 0.83
		Kara: 0.85 ± 0.45	0.44 ± 0.40	1.03 ± 0.28	1.55 ± 0.86	1.54 ± 0.83
	THÖR3	Sof: 0.87 ± 0.47	0.43 ± 0.40	1.03 ± 0.28	1.56 ± 0.81	1.53 ± 0.84
		Zan: 0.86 ± 0.46	0.43 ± 0.40	1.03 ± 0.28	1.56 ± 0.85	1.54 ± 0.84
		Kara: 0.86 ± 0.47	0.44 ± 0.40	1.03 ± 0.28	1.61 ± 0.88	1.58 ± 0.85

Table 6.7: FDE measured in the transfer experiments on different datasets. THÖR1 abbreviates the first "One obstacle" scenario in THÖR, and THÖR3 the third one ("Three obstacles").

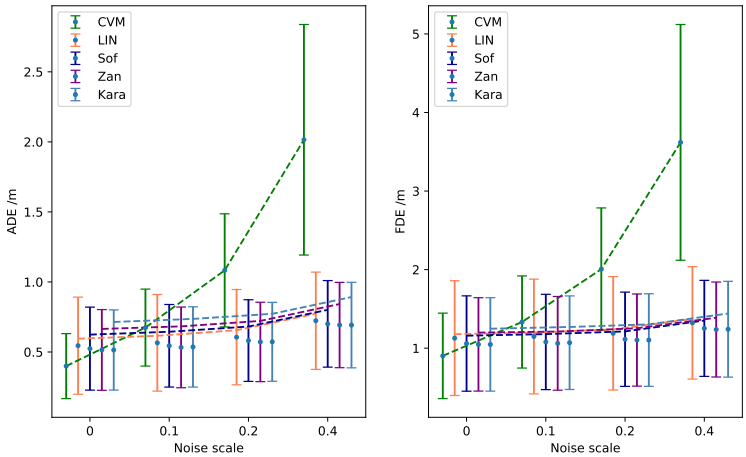


Figure 6.10: ADE/FDE in the ETH dataset with adding noise and using linear filter

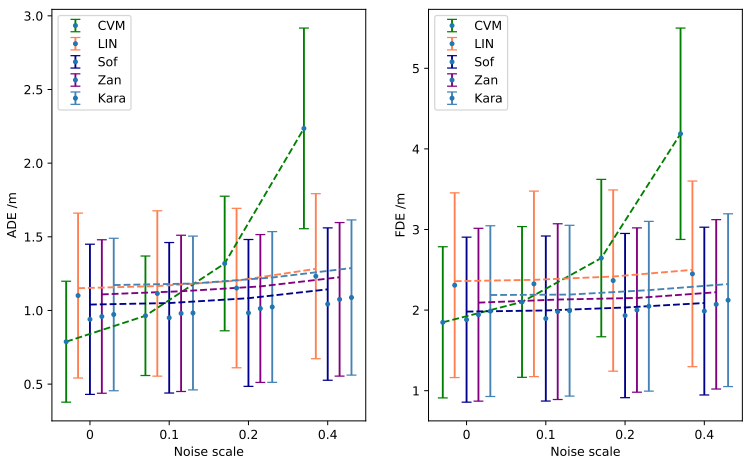


Figure 6.11: ADE/FDE in the THÖR dataset ("Three obstacles" scenario) with added noise and using linear velocity filter

Chapter 7
Conclusions

Understanding and predicting human behavior is a key skill for intelligent systems to coexist and interact with humans. An increasingly important branch of robotics research, it has attracted a remarkable amount of attention in recent years due to the rapid developments in automated driving and service robotics technology. Showing great potential, motion prediction elevates safety and efficiency in some applications, for instance in mobile robotics, while fundamentally enabling others – for instance, fully-automated self-driving vehicles are unimaginable without some form of future assessment.

The work, presented in this thesis, is an ambitious attempt to overview, structure and advance the motion prediction domain. This chapter concludes the thesis with a summary of contributions in Sec. 7.1, a review of the current trends and open challenges in the literature in Sec. 7.2, and an outlook on our ongoing and future work in Sec. 7.3.

7.1 Contributions

This thesis explores the fundamental tasks in human motion prediction for autonomous systems, ranging from surveying the complete methodology, requirements and application scenarios, aspects in data collection and method development, to evaluation, benchmarking and integration. It formulates many questions and open challenges in the area, offering solutions to several of them and proposing the directions of future research to deal with others. The discussion begins with the central question in motion prediction:

How to model motion?

To this end in Chapter 2 we review in-depth the complete methodology for motion trajectory prediction developed in the last three decades. We propose an organizational principle to classify methods based on the approach to model

motion and the level of contextual awareness. In the first category, the *physics-based*, *pattern-based* and *planning-based* approaches each have their own benefits and drawbacks. With respect to the contextual cues, such as the articulated pose of the target agent, social grouping or semantic information about the environment, usually higher levels of awareness are desirable for better performance. On the other hand, the problem of introducing certain cues to specific modeling approaches is a challenging one: for instance, very few of the highly popular pattern-based approaches can handle environments cluttered with complex non-convex obstacles, e.g. hospitals or offices. On the other hand, very few purely physics-based approaches are capable of incorporating full-body poses due to the complexity of the underlying dynamical models. Discussing the modeling approaches from the task-centered perspective, we conclude that the existing classes of approaches have their theoretical potential in different tasks.

Having reviewed the modeling approaches and prediction infrastructure, the thesis moves on to the second major question:

How to design a motion prediction approach for a service robot in cluttered and dynamic environments?

For a mobile robot, operating in such environments where both people and obstacles are present, the prediction method should be able to account for these two factors. To this end a novel prediction approach is presented in Chapter 3, which combines the strengths of the planning-based obstacle map awareness and the interaction modeling with physics-based group social forces. The proposed method incorporates many crucial aspects of the motion prediction domain exposed in this thesis: it has a high level of contextual awareness, outputs model-free uncertainty in human motion, explicitly reasons over possible goals of each person and alternative ways to reach them. Our method bridges a gap between the short-term motion prediction problem, where the dynamic environment cues are dominant, and long-term prediction, where semantics, obstacles and goals strongly influence the motion of people.

The third question we discuss in this thesis is the use of semantics for improved long-term motion prediction:

What can be learned from semantic maps for improved human-awareness in previously unseen environments?

Semantic information about the environment is a highly informative cue for the prediction of pedestrian motion or the estimation of collision risks. In Chapter 4 we explore the possibility to infer occupancy priors of human motion using only semantic environment information as input. To this end we apply and discuss a traditional Inverse Optimal Control approach, and propose a novel one based on Convolutional Neural Networks (CNN). Our CNN

method produces flexible context-aware occupancy estimations for semantically uniform map regions and generalizes well already with small amounts of training data. Evaluated on synthetic and real-world data, it shows superior results compared to several baselines, marking a qualitative step-up in semantic environment assessment. Using such occupancy priors opens the possibility of truly forward-looking predictive behavior: collision anticipation for autonomous vehicles along the planned route, less obtrusive motion planning for mobile robots, more informed search for people to assist for the service robots.

The fourth major question with considerable exposure in this thesis is the one of data and benchmarking:

How to choose data and experiment design for benchmarking and evaluation?

The existing and extensively used datasets of human motion trajectories are insufficient in the level of recorded contextual cues, imbalanced towards straight constant velocity trajectories and often suffer from severe annotation problems. Many of these issues naturally follow from the usual data collection protocol: recording natural motion in the wild with a few external sensors. This recording protocol severely limits the choice of possible environments and their configurations with restricted recording permissions and limited sensor montage possibilities. Furthermore, it impairs the annotation accuracy, achieved with position extraction in post-processing. As an alternative to this, we propose to record motion in controlled settings using a motion capture system. To generate natural and purposeful behaviors of the recorded participants, we design an elaborate weakly-scripted data collection procedure with social roles, dynamically-allocated goals, group motion, obstacles and a moving robot. As a result, all the issues listed above are addressed in our novel THÖR dataset, presented in Chapter 5, which includes diverse and very accurate trajectories with rich and non-trivial interactions in an indoor environment.

The recorded data is incorporated into a new motion prediction benchmarking suite, presented in Chapter 6. This benchmark is designed for a thorough evaluation of motion prediction methods in a variety of experiments: performance conditioned on several key factors (e.g. prediction horizon, observation length), evaluation of knowledge transfer to a new environment, testing robustness against added perception noise. The benchmark can be used with any dataset of human motion, it includes tools for data processing and formalized scenario extraction for certain observation and prediction lengths.

Our work towards better motion prediction for service robots is far from over. Hopefully, with the questions raised in this thesis and proposed future work directions, we lay foundations for the sustainable and enduring development of better methods for motion prediction.

7.2 Open Challenges and Future Research Directions

The design of motion prediction methods has come a long way since the first experiments with Kalman filleting and simple motion models. Modern techniques make extensive use of machine learning in order to better estimate context-dependent patterns in real-world data, handle more complex environment models and types of motion, or even propose end-to-end reasoning on future motion from visual input. An increasing number of methods also includes reasoning on the global structure of the environment, intentions and actions of the agent. Having these trends in mind, we see several open challenges and directions of future research:

7.2.1 Use of enhanced contextual cues

To analyze and predict human motion, as well as to plan and navigate alongside them, intelligent systems should have an in-depth semantic scene understanding. Context understanding with respect to features of the static environment and its semantics for better trajectory prediction is still a relatively unexplored area, see Sec. 2.7.3 for more details.

The same argument applies to the contextual cues of the dynamic environment. Socially-aware methods are making an important improvement over socially-unaware ones in such spaces where the target agent is not acting in isolation. However, most existing socially-aware methods still assume that all observed people are behaving similarly and that their motion can be predicted by the same model and with the same features. Capturing and reasoning on the high-level social attributes is at an early stage of development, see Sec. 2.7.1 and Sec. 2.7.2, however recent methods take initial steps to this end. Furthermore, most available approaches assume cooperative behavior, while real humans might rather optimize personal goals instead of joint strategies. In such cases, game-theoretic approaches are possibly better suited for modeling human behavior. Consequently, adopting classical AI and game-theoretic approaches in multi-agent systems is a promising research direction, that is only partly addressed in recent work, see e.g. [18, 202].

One task where contextual cues become particularly important is long-term prediction of motion trajectories. While context-agnostic motion and behavioral patterns are helpful for short prediction horizons, long-term predictions should account for intentions, based on the context and the surrounding environment. Many pattern-based methods treat agents as particles, placed in the field of learned transitions, dictating the direction of future motion. Extending these models by more goal- or intention-driven predictions, that resemble human goal-directed behavior, would be beneficial for long-term predictions.

Consequently, further research on automatic goal inference based on the semantics of the environment is important. Most planning-based methods rely on a given set of goals, which makes them unusable or imprecise in a situation where no goals are known beforehand, or the number of possible goals is too high. Alternatively, one could consider identifying on-the-fly possible goals in the environment and predicting the way the agent may reach those goals. This would allow the application of the planning-based methods in unknown environments. Additionally, semantic indicators of possible goals, coming from understanding the person's social role or current activity [47], could lead to more robust intention recognition.

Apart from the contextual cues, discussed in this thesis, there are many other factors influencing pedestrian motion, according to the recent studies [253], e.g. weather conditions, time of day, social roles of agents. Future methods could benefit from a closer connection to the studies of human motion and behavior in social spaces [16, 73, 102].

7.2.2 Robustness and Integration

Several practical aspects of deploying prediction systems in real environments should be considered in the future work.

Most of the presented methods are designed for specific tasks, scenarios or types of motion. These methods work well in certain situations, e.g. when prominent motion patterns exist in the environment, or when the spatial structure of the environment and target agent's goals are known beforehand. A conceptually interesting approach that uses a combination of multiple prediction algorithms to reason about best performance in the given situation is presented by Lasota and Shah [173]. The multiple-predictor framework opens a possibility for achieving more robust predictions when operating in undefined, changing situations, where a combination of strengths of different methods is required.

We suggest that more emphasis should be put on transfer learning and generalization of approaches to new environments. Learning and reasoning on basic, invariant rules and norms of human motion and collision avoidance is a better approach in this case. When having access to several environments, domain adaptation could be potentially used for learning generalizable models.

Integration of prediction in planning and control is another worthwhile topic for overall system robustness. Predicting human motion is usually motivated with increased safety of human-robot interaction and efficiency of operation. However, the insights on exploiting predictions in the robot's motion or action planning module are typically left out of scope in many papers. Future work would benefit from outlining possible ways to incorporate predictions in the robot control framework.

7.3 Ongoing and Future Work

In its broad attempt to overview an emerging area of active research, this thesis includes numerous suggestions on how to improve and further develop the presented methods in future work. These are concentrated in Sec. 2.9 and Sec. 7.2, where we discuss the current limitations and open challenges respectively, and in the "Conclusions and Outlook" sections of each chapter, where we outline the specific improvements to the presented methods.

In summary, there are several major directions of future work towards better motion prediction for autonomous systems. Finding new ways to exploit relevant contextual cues, such as eye-gaze and head orientation, attention to other moving actors, semantics, social relations and personal traits, is among the most important ones. To support such research, we need better datasets that include these relevant factors and have enough variability in each of them to allow systematic verification of methods under a variety of conditions. Creating better datasets in terms of data quantity and condition diversity is another major direction of future research.

In addition to that, benchmarking and integration are key research topics in the area. In the following sections we briefly present the ongoing work in these two directions.

7.3.1 Benchmarking Trajectory Forecasting Methods

The available benchmarking infrastructure, presented in state-of-the-art TrajNet++ benchmark [160] and Chapter 7 of this thesis, is hardly sufficient for the diverse range of challenging problems in motion prediction. There are many uncovered aspects: for instance, safety-critical evaluation of anomalous or outlier trajectory prediction, goal or intention inference evaluation, systematic evaluation of model adaptation for systems that actively learn from new observations. In the meanwhile, the rising interest to benchmarking manifests itself in the new challenges[1] and research efforts on specific aspects of benchmarking, for instance evaluating single isolated trajectory predictions [122], datasets of synthetic trajectories with multiple plausible futures [189], balanced verification datasets of required complexity [12, 121].

Our ongoing research in benchmarking follows the successful collaboration during the 2020 ECCV Workshop on Benchmarking Trajectory Forecasting Models[2] and builds on the results achieved in Chapter 6 of this thesis. We are interested in offering a unified review of the motion prediction evaluation aspects (e.g., scenario selection, data pre-processing, evaluation protocols and metrics), and formulate properties of a general benchmark for motion predic-

[1] `https://eval.ai/web/challenges/challenge-page/454/overview`, `https://www.nuscenes.org/prediction?externalData=all&mapData=all&modalities=Any`

[2] `https://sites.google.com/view/btfm2020`

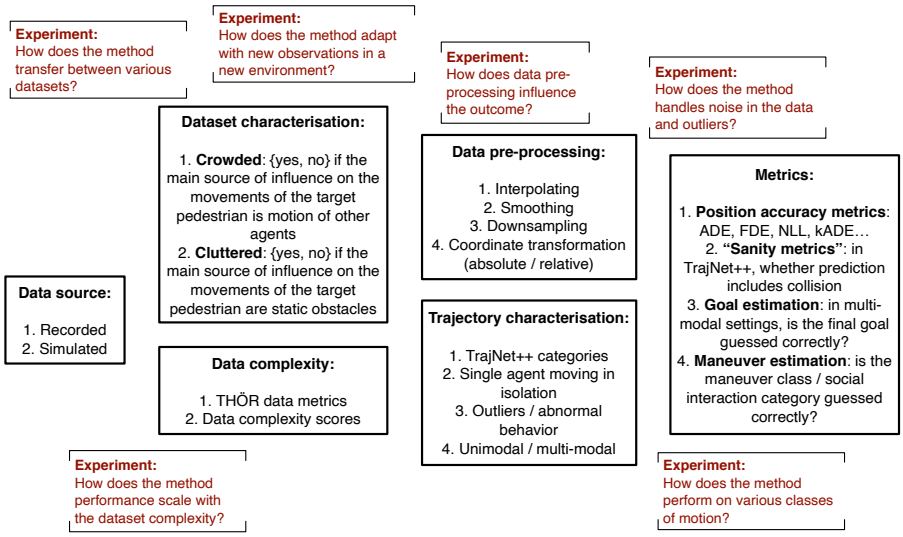


Figure 7.1: Aspects in benchmarking trajectory prediction

tion. Such material is not only useful as a thorough guideline for benchmarking with a variety of possible experiments, but also helps to identify the missing capabilities in the existing benchmarks, contributing to the creation of more meaningful challenges and enduring development of better prediction methods. A collection of ideas for our benchmark research is presented in Fig. 7.1.

7.3.2 Hierarchical Predictive Planning System

A major direction of our future work will be researching the integration of prediction methods into the control pipelines of autonomous systems, and the improvement reached thereby. As we discussed in Sec. 7.2, this question is often omitted from the prediction papers. The same is true for the planning literature: while some considerable attention is placed on researching human-like or human-aware collision avoidance policies [55, 84, 241], very few papers actually research planning with explicit predictions [19, 30, 92, 370].

One prominent idea, driving the discussion and method development in this thesis, is the one of combination: in cases where a single method or a single model fails, multiple approaches have the potential to improve performance. This idea is illustrated, for instance, by the multi-model physics-based approaches, grouped in a separate category in our taxonomy, and, importantly, the work of Lasota and Shah [173], who introduce a Multiple Predictor Approach to ensure the best performance on various prediction horizons. Our own method, presented in Chapter 3, combines a planning-based and interaction-

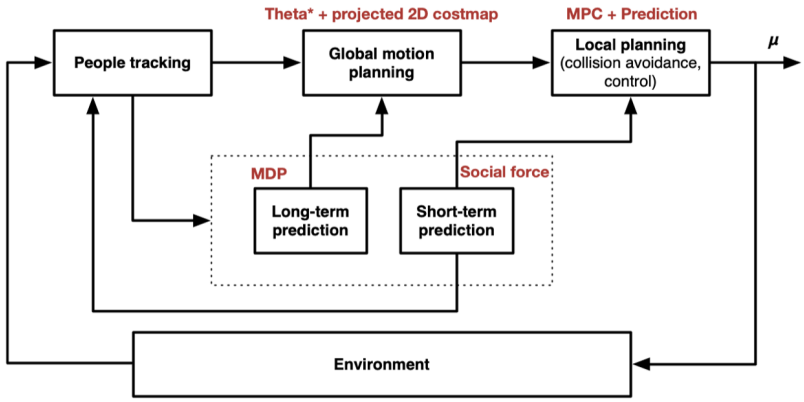


Figure 7.2: Proposed hierarchical planning architecture. An autonomous system, navigating in shared environments, should be able to detect and track other dynamic agents, plan its own navigation trajectory and execute it as a series of control inputs. Prediction module informs human tracking, path planning and control of the next probable positions.

based modules, allowing a seamless transition between the short-term and long-term predictions and higher accuracy in both cases. What is common to all these methods is that they essentially assume a single operating model at a time.

Interestingly, the applications often place contradictory requirements on this model. On one hand, predictions should be as precise as possible, especially in the short-term perspective. This is a safety-driven requirement, to prevent collisions, abrupt braking and overly conservative behavior, arising from uncertainty. On the other hand, predictions should be multi-modal and uncertainty-aware for optimal path planning. This requirement, more important in the long-term perspective, allows the robot to reach its destination in a smooth and unobtrusive manner.

This observation motivates our hypothesis that in fact several predictors need to be combined in parallel: a short-term method with the primary focus on safety and collision avoidance, and a long-term multi-modal uncertainty-aware method for global motion planning. This would resolve the necessity to control uncertainty in the short-term and long-term perspective uniformly, and allow running these two approaches at various frequencies. This is especially important since long-term prediction methods tend to be more computationally-demanding.

Fig. 7.2 outlines our design of a hierarchical predictive planning system, which combines short-term and long-term predictions. The specific ways to efficiently include predictions into the local and global planning, as well as the specifications for the prediction modules, are the subjects of our future research.

References

[1] Cityscapes pixel-level semantic labeling benchmark. URL `https://www.cityscapes-dataset.com/benchmarks/`.

[2] H. Admoni and B. Scassellati. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1):25–63, 2017.

[3] G. Agamennoni, J. I. Nieto, and E. M. Nebot. Estimation of multivehicle dynamics by considering contextual information. *IEEE Trans. on Robotics (TRO)*, 28(4):855–870, 2012.

[4] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social LSTM: Human trajectory prediction in crowded spaces. In *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, pages 961–971, 2016.

[5] A. Alahi, L. Ballan, P. Coscia, L. Palmieri, and A. Rudenko. Workshop on Benchmarking Trajectory Forecasting Models (BTFM 2020). In *Proc. of the Europ. Conf. on Comp. Vision (ECCV)*, 2020. URL `https://sites.google.com/view/btfm2020`.

[6] F. Altché and A. de La Fortelle. An LSTM network for highway trajectory prediction. In *Proc. of the IEEE Int. Conf. on Intell. Transp. Syst. (ITSC)*, pages 353–359. IEEE, 2017.

[7] M. Althoff. *Reachability analysis and its application to the safety assessment of autonomous cars*. PhD thesis, TU Munich, 2010.

[8] M. Althoff, O. Stursberg, and M. Buss. Reachability analysis of nonlinear systems with uncertain parameters using conservative linearization. In *Proc. of the IEEE Int. Conf. on Decision and Control (CDC)*, pages 4042–4048, 2008.

[9] M. Althoff, O. Stursberg, and M. Buss. Stochastic reachable sets of interacting traffic participants. In *Proc. of the IEEE Intell. Veh. Symp. (IV)*, pages 1086–1092, 2008.

[10] M. Althoff, D. Heß, and F. Gambert. Road occupancy prediction of traffic participants. In *Proc. of the IEEE Int. Conf. on Intell. Transp. Syst. (ITSC)*, pages 99–105, 2013.

[11] J. Amirian, J.-B. Hayet, and J. Pettré. Social ways: Learning multi-modal distributions of pedestrian trajectories with GANs. In *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR) Workshops*, pages 0–0, 2019.

[12] J. Amirian, B. Zhang, F. V. Castro, J. J. Baldelomar, J.-B. Hayet, and J. Pettre. Opentraj: Assessing prediction complexity in human trajectories datasets. In *Proc. of the Asian Conf. on Computer Vision (ACCV)*, 2020.

[13] G. Antonini, S. V. Martinez, M. Bierlaire, and J. P. Thiran. Behavioral priors for detection and tracking of pedestrians in video sequences. *Int. J. of Comp. Vision (IJCV)*, 69(2):159–180, 2006.

[14] G. Aoude, J. Joseph, N. Roy, and J. How. Mobile agent trajectory prediction using bayesian nonparametric reachability trees. In *Proc. of AIAA Infotech@Aerospace (I@A)*, pages 1–17. 2011.

[15] G. S. Aoude, B. D. Luders, K. K. H. Lee, D. S. Levine, and J. P. How. Threat assessment design for driver assistance system at intersections. In *Proc. of the IEEE Int. Conf. on Intell. Transp. Syst. (ITSC)*, pages 1855–1862, 2010.

[16] G. Arechavaleta, J.-P. Laumond, H. Hicheur, and A. Berthoz. An optimality principle governing human walking. *IEEE Trans. on Robotics and Automation (TRO)*, 24(1):5–14, 2008.

[17] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. on Patt. Anal. and Mach. Intell. (PAMI)*, 39(12):2481–2495, 2017.

[18] M. Bahram, A. Lawitzky, J. Friedrichs, M. Aeberhard, and D. Wollherr. A game-theoretic approach to replanning-aware interactive scene prediction and planning. *IEEE Trans. on Veh. Techn.*, 65(6):3981–3992, 2016.

[19] H. Bai, S. Cai, N. Ye, D. Hsu, and W. S. Lee. Intention-aware online POMDP planning for autonomous driving in a crowd. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 454–460, May 2015. doi: 10.1109/ICRA.2015.7139219.

[20] L. Ballan, F. Castaldo, A. Alahi, F. Palmieri, and S. Savarese. Knowledge transfer for scene-specific motion prediction. In *Proc. of the Europ. Conf. on Comp. Vision (ECCV)*, pages 697–713. Springer, 2016.

[21] S. Bandini, F. Rubagotti, G. Vizzari, and K. Shimura. An agent model of pedestrian and group dynamics: experiments on group cohesion. *AI* IA 2011: Artificial Intelligence Around Man and Beyond*, 2011.

[22] T. Bandyopadhyay, K. S. Won, E. Frazzoli, D. Hsu, W. S. Lee, and D. Rus. Intention-aware motion planning. In *Algorithmic Foundations of Robotics X*, pages 475–491. Springer, 2013.

[23] S. Bansal, A. Bajcsy, E. Ratner, A. D. Dragan, and C. J. Tomlin. A hamilton-jacobi reachability-based framework for predicting and analyzing human motion for safe planning. *arXiv:1910.13369*, 2019.

[24] A. Barth and U. Franke. Where will the oncoming vehicle be the next second? In *Proc. of the IEEE Intell. Veh. Symp. (IV)*, pages 1068–1073, 2008.

[25] F. Bartoli, G. Lisanti, L. Ballan, and A. D. Bimbo. Context-aware trajectory prediction. In *Proc. of the IEEE Int. Conf. on Pattern Recognition*, pages 1941–1946, 2018.

[26] I. Batkovic, M. Zanon, N. Lubbe, and P. Falcone. A computationally efficient model for pedestrian motion prediction. pages 374–379, 2018.

[27] T. Batz, K. Watson, and J. Beyerer. Recognition of dangerous situations within a cooperative group of vehicles. In *Proc. of the IEEE Intell. Veh. Symp. (IV)*, pages 907–912, 2009.

[28] B. Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video. In *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, pages 3457–3464, 2011.

[29] M. Bennewitz, W. Burgard, and S. Thrun. Using EM to learn motion behaviors of persons with mobile robots. In *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*, pages 502–507, 2002.

[30] M. Bennewitz, W. Burgard, G. Cielniak, and S. Thrun. Learning motion patterns of people for compliant robot motion. *Int. J. of Robotics Research*, 24(1):31–48, 2005.

[31] A. Bera, S. Kim, T. Randhavane, S. Pratapa, and D. Manocha. GLMP-realtime pedestrian path prediction using global and local movement patterns. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 5528–5535, 2016.

[32] A. Bera, T. Randhavane, and D. Manocha. Aggressive, tense, or shy? Identifying personality traits from crowd videos. In *Proc. of the Int. Conf. on Artificial Intelligence (IJCAI)*, pages 112–118, 2017.

[33] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP J. on Image and Video Proc.*, 2008(1), May 2008.

[34] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *Workshop Proc. of the AAAI Conf. on Artificial Intelligence on Knowledge Discovery and Data Mining*, AAAIWS'94, pages 359–370. AAAI Press, 1994.

[35] G. Best and R. Fitch. Bayesian intention inference for trajectory prediction with an unknown goal destination. In *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*, pages 5817–5823, 2015.

[36] R. Best and J. Norton. A new model and efficient tracker for a target with curvilinear motion. *IEEE Trans. on Aerospace and Electronic Syst. (AESS)*, 33(3):1030–1037, 1997.

[37] S. Bhattacharya, V. Kumar, and M. Likhachev. Search-based path planning with homotopy class constraints. In *Proc. of the Annual Symp. on Comb. Search*, 2010.

[38] A. Bhattacharyya, M. Hanselmann, M. Fritz, B. Schiele, and C.-N. Straehle. Conditional flow variational autoencoders for structured sequence prediction. *arXiv:1908.09008*, 2019.

[39] N. Bisagno, B. Zhang, and N. Conci. Group LSTM: Group trajectory prediction in crowded scenarios. In *Proc. of the Europ. Conf. on Comp. Vision (ECCV)*, pages 213–225. Springer, 2018.

[40] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

[41] C. Blaiotta. Learning generative socially aware models of pedestrian motion. *IEEE Robotics and Automation Letters*, 4(4):3433–3440, 2019.

[42] M. A. Brewer, K. Fitzpatrick, J. A. Whitacre, and D. Lord. Exploration of pedestrian gap-acceptance behavior at selected locations. *Transportation research record*, 1982(1):132–140, 2006.

[43] A. Broadhurst, S. Baker, and T. Kanade. Monte carlo road safety reasoning. In *Proc. of the IEEE Intell. Veh. Symp. (IV)*, pages 319–324, 2005.

[44] N. Brouwer, H. Kloeden, and C. Stiller. Comparison and evaluation of pedestrian motion models for vehicle safety systems. In *Proc. of the IEEE Int. Conf. on Intell. Transp. Syst. (ITSC)*, pages 2207–2212, 2016.

[45] G. W. Brown. Iterative solution of games by fictitious play. *Activity analysis of production and allocation*, 13(1):374–376, 1951.

[46] A. Bruce and G. Gordon. Better motion prediction for people-tracking. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2004.

[47] L. Bruckschen, N. Dengler, and M. Bennewitz. Human motion prediction based on object interactions. In *Proc. of the European Conf. on Mobile Robots (ECMR)*, pages 1–6. IEEE, 2019.

[48] D. Brščić, T. Kanda, T. Ikeda, and T. Miyashita. Person tracking in large public spaces using 3-d range sensors. *IEEE Trans. on Human-Machine Systems*, 43(6):522–534, 2013.

[49] D. Buzan, S. Sclaroff, and G. Kollios. Extraction and clustering of motion trajectories in video. In *Proc. of the IEEE Int. Conf. on Pattern Recognition*, volume 2, pages 521–524 Vol.2, Aug 2004.

[50] Y. Cai, N. de Freitas, and J. J. Little. Robust visual tracking for multiple targets. In *Proc. of the Europ. Conf. on Comp. Vision (ECCV)*, pages 107–118, 2006.

[51] J. F. Carvalho, M. Vejdemo-Johansson, F. T. Pokorny, and D. Kragic. Long-term prediction of motion trajectories using path homology clusters. In *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*, 2019.

[52] R. T. Chadalavada, H. Andreasson, M. Schindler, R. Palm, and A. J. Lilienthal. Bi-directional navigation intent communication using spatial augmented reality and eye-tracking glasses for improved safety in human–robot interaction. *Robotics and Computer-Integrated Manufacturing*, 61:101830, 2020.

[53] Y. Chai, B. Sapp, M. Bansal, and D. Anguelov. MultiPath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. *arXiv:1910.05449*, 2019.

[54] Y. F. Chen, M. Liu, and J. P. How. Augmented dictionary learning for motion prediction. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 2527–2534, 2016.

[55] Y. F. Chen, M. Everett, M. Liu, and J. P. How. Socially aware motion planning with deep reinforcement learning. In *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*, pages 1343–1350. IEEE, 2017.

[56] Y. F. Chen, M. Liu, M. Everett, and J. P. How. Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 285–292, 2017.

[57] Z. Chen, D. C. K. Ngai, and N. H. C. Yung. Pedestrian behavior prediction based on motion patterns for vehicle-to-pedestrian collision avoidance. In *Proc. of the IEEE Int. Conf. on Intell. Transp. Syst. (ITSC)*, pages 316–321, 2008.

[58] L. Cheng, R. Yarlagadda, C. Fookes, and P. K. Yarlagadda. A review of pedestrian group dynamics and methodologies in modelling pedestrian group behaviours. *World*, 1(1):002–013, 2014.

[59] S. F. Chik, C. F. Yeong, E. L. M. Su, T. Y. Lim, Y. Subramaniam, and P. J. H. Chin. A review of social-aware navigation frameworks for service robot in dynamic human environments. *J. of Telecomm., Electronic and Comp. Eng. (JTEC)*, 8(11):41–50, 2016.

[60] C. Choi, A. Patil, and S. Malla. Drogon: A causal reasoning framework for future trajectory forecast. *arXiv:1908.00024*, 2019.

[61] W. Choi and S. Savarese. Multiple target tracking in world coordinate with single, minimally calibrated camera. In *Proc. of the Europ. Conf. on Comp. Vision (ECCV)*, pages 553–567. Springer, 2010.

[62] S.-Y. Chung and H.-P. Huang. A mobile robot that understands pedestrian spatial behaviors. In *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*, pages 5861–5866, 2010.

[63] S.-Y. Chung and H.-P. Huang. Incremental learning of human social behaviors with feature-based spatial effects. In *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*, pages 2417–2422, 2012.

[64] J. Colyar and J. Halkias. US highway 80 dataset, federal highway administration (FHWA), vol. *Tech, no. Rep*, 2006.

[65] J. Colyar and J. Halkias. US highway 101 dataset. *Federal Highway Administration (FHWA), Tech. Rep. FHWA-HRT-07-030*, 2007.

[66] P. Coscia, F. Castaldo, F. A. N. Palmieri, A. Alahi, S. Savarese, and L. Ballan. Long-term path prediction in urban scenarios using circular distributions. *Image and Vision Computing*, 69:81–91, 2018.

[67] H. Cui, V. Radosavljevic, F.-C. Chou, T.-H. Lin, T. Nguyen, T.-K. Huang, J. Schneider, and N. Djuric. Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2019.

[68] S. Dai, L. Li, and Z. Li. Modeling vehicle interactions via modified LSTM models for trajectory prediction. *IEEE Access*, 7:38287–38296, 2019.

[69] N. Deo and M. M. Trivedi. Multi-modal trajectory prediction of surrounding vehicles with maneuver based LSTMs. In *Proc. of the IEEE Intell. Veh. Symp. (IV)*, pages 1179–1184, 2018.

[70] F. Diehl, T. Brunner, M. T. Le, and A. Knoll. Graph neural networks for modelling traffic participant interaction. In *Proc. of the IEEE Intell. Veh. Symp. (IV)*, pages 695–701. IEEE, 2019.

[71] W. Ding, J. Chen, and S. Shen. Predicting vehicle behaviors over an extended horizon using behavior interaction network. *arXiv:1903.00848*, 2019.

[72] N. Djuric, V. Radosavljevic, H. Cui, T. Nguyen, F.-C. Chou, T.-H. Lin, and J. Schneider. Motion prediction of traffic actors for autonomous driving using deep convolutional networks. *arXiv:1808.05819*, 2018.

[73] T. Do, M. Haghani, and M. Sarvi. Group and single pedestrian behavior in crowd dynamics. *Transportation research record*, 2540(1):13–19, 2016.

[74] J. Doellinger, M. Spies, and W. Burgard. Predicting occupancy distributions of walking humans with convolutional neural networks. *IEEE Robotics and Automation Letters*, 3(3), 2018.

[75] J. Doellinger, V. S. Prabhakaran, L. Fu, and M. Spies. Environment-aware multi-target tracking of pedestrians. *IEEE Robotics and Automation Letters*, 4(2):1831–1837, 2019.

[76] C. Dondrup, N. Bellotto, F. Jovan, and M. Hanheide. Real-time multisensor people tracking for human-robot spatial interaction. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA), Workshop on ML for Social Robo*. IEEE, 2015.

[77] A. Doshi and M. M. Trivedi. On the roles of eye gaze and head dynamics in predicting driver's intent to change lanes. *IEEE Trans. on Intell. Transp. Syst. (TITS)*, 10(3):453–462, 2009.

[78] M.-P. Dubuisson and A. K. Jain. A modified hausdorff distance for object matching. In *Proc. of the IEEE Int. Conf. on Pattern Recognition*, volume 1, pages 566–568. IEEE, 1994.

[79] S. Eiffert and S. Sukkarieh. Predicting responses to a robot's future motion using generative recurrent neural networks. *arXiv:1909.13486*, 2019.

[80] J. Elfring, R. van de Molengraft, and M. Steinbuch. Learning intentions for improved human motion prediction. *J. of Robotics and Autonomous Systems*, 62(4):591–602, 2014.

[81] D. Ellis, E. Sommerlade, and I. Reid. Modelling pedestrian trajectory patterns with gaussian processes. In *Proc. of the Int. Conf. on Comp. Vision Worksh.*, pages 1229–1234. IEEE, 2009.

[82] A. Elnagar. Prediction of moving objects in dynamic environments using kalman filters. In *Proc. of the IEEE Int. Symp. on Comp. Intel. in Robotics and Automation (CIRA)*, pages 414–419, 2001. doi: 10.1109/CIRA.2001.1013236.

[83] A. Elnagar and K. Gupta. Motion prediction of moving objects based on autoregressive model. *IEEE Trans. on Syst., Man, and Cybernetics (SMC) - Part A: Systems and Humans*, 28(6):803–810, Nov 1998. ISSN 1083-4427. doi: 10.1109/3468.725351.

[84] M. Everett, Y. F. Chen, and J. P. How. Motion planning among dynamic, decision-making agents with deep reinforcement learning. In *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*, pages 3052–3059. IEEE, 2018.

[85] Z. Fang, D. Vázquez, and A. M. López. On-board detection of pedestrian intentions. *Sensors*, 17(10):2193, 2017.

[86] F. Farina, D. Fontanelli, A. Garulli, A. Giannitrapani, and D. Prattichizzo. Walking ahead: The headed social force model. *PloS one*, 12 (1):e0169734, 2017.

[87] P. Fearnhead and Z. Liu. On-line inference for multiple changepoint problems. *J. of the Royal Stat. Soc.: Series B (Statistical Methodology)*, 69(4):589–605, 2007.

[88] S. Ferguson, B. Luders, R. C. Grande, and J. P. How. Real-time predictive modeling and robust avoidance of pedestrians with uncertain, changing intentions. In *Algorithmic Foundations of Robotics XI*, pages 161–177. Springer, 2015.

[89] T. Fernando, S. Denman, S. Sridharan, and C. Fookes. Soft+Hardwired attention: An LSTM framework for human trajectory prediction and abnormal event detection. *Neural networks*, 108:466–478, 2018.

[90] T. Fernando, S. Denman, S. Sridharan, and C. Fookes. Neighbourhood context embeddings in deep inverse reinforcement learning for predicting pedestrian motion over long time horizons. In *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV) Workshops*, pages 0–0, 2019.

[91] G. Ferrer and A. Sanfeliu. Behavior estimation for a complete framework for human motion prediction in crowded environments. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 5940–5945, 2014.

[92] G. Ferrer and A. Sanfeliu. Anticipative kinodynamic planning: multi-objective robot navigation in urban and dynamic environments. *J. of Autonomous Robots*, 43(6):1473–1488, 2019.

[93] A. F. Foka and P. E. Trahanias. Predictive autonomous robot navigation. In *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*, 2002.

[94] A. F. Foka and P. E. Trahanias. Probabilistic autonomous robot navigation in dynamic environments with human motion prediction. *Int. Journal of Social Robotics*, 2(1):79–94, 2010.

[95] E. Galceran, A. G. Cunningham, R. M. Eustice, and E. Olson. Multipolicy decision-making for autonomous driving via changepoint-based behavior prediction. In *Proc. of the Robotics: Science and Systems (RSS)*, 2015.

[96] L. Gan, R. Zhang, J. W. Grizzle, R. M. Eustice, and M. Ghaffari. Bayesian spatial kernel smoothing for scalable dense semantic mapping. *IEEE Robotics and Automation Letters*, 5(2):790–797, April 2020. ISSN 2377-3774. doi: 10.1109/LRA.2020.2965390.

[97] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, 2012.

[98] T. Gindele, S. Brechtel, and R. Dillmann. A probabilistic model for estimating driver behaviors and vehicle trajectories in traffic environments. In *Proc. of the IEEE Int. Conf. on Intell. Transp. Syst. (ITSC)*, pages 1625–1631, 2010.

[99] M. Goldhammer, K. Doll, U. Brunsmann, A. Gensler, and B. Sick. Pedestrian's trajectory forecast in public traffic with artificial neural networks. In *Proc. of the IEEE Int. Conf. on Pattern Recognition*, pages 4110–4115, Aug 2014. doi: 10.1109/ICPR.2014.704.

[100] H. Gong, J. Sim, M. Likhachev, and J. Shi. Multi-hypothesis motion planning for visual object tracking. In *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, pages 619–626, 2011.

[101] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Inf. Proc. Syst. (NIPS)*, pages 2672–2680, 2014.

[102] A. Gorrini, G. Vizzari, and S. Bandini. Age and group-driven pedestrian behaviour: from observations to simulations. *Collective Dynamics*, 1: 1–16, 2016.

[103] H. Grimmett, M. Buerki, L. Paz, P. Pinies, P. Furgale, I. Posner, and P. Newman. Integrating metric and semantic maps for vision-only automated parking. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 2159–2166. IEEE, 2015.

[104] Y. Gu, Y. Hashimoto, L.-T. Hsu, and S. Kamijo. Motion planning based on learning models of pedestrian and driver behaviors. In *Proc. of the IEEE Int. Conf. on Intell. Transp. Syst. (ITSC)*, pages 808–813, 2016.

[105] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi. Social GAN: Socially acceptable trajectories with generative adversarial networks. In *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, June 2018.

[106] D. Gupta. Semantic segmentation using keras. URL `https://github.com/divamgupta/image-segmentation-keras`.

[107] G. Habibi, N. Jaipuria, and J. P. How. Context-aware pedestrian motion prediction in urban intersections. *arXiv:1806.09453*, 2018.

[108] Y. Han, R. Tse, and M. Campbell. Pedestrian motion model using non-parametric trajectory clustering and discrete transition points. *IEEE Robotics and Automation Letters*, 4(3):2614–2621, 2019.

[109] I. Hasan, F. Setti, T. Tsesmelis, A. Del Bue, F. Galasso, and M. Cristani. MX-LSTM: mixing tracklets and vislets to jointly forecast trajectories and head poses. In *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, pages 6067–6076, 2018.

[110] D. Helbing and P. Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995.

[111] P. Henry, C. Vollmer, B. Ferris, and D. Fox. Learning to navigate through crowded environments. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 981–986. IEEE, 2010.

[112] C. Hermes, C. Wöhler, K. Schenk, and F. Kummert. Long-term vehicle motion prediction. In *Proc. of the IEEE Intell. Veh. Symp. (IV)*, pages 652–657, 2009.

[113] T. Hirakawa, T. Yamashita, T. Tamaki, and H. Fujiyoshi. Survey on vision-based path prediction. In *Distributed, Ambient and Pervasive Interactions: Technologies and Contexts*, pages 48–64. Springer International Publishing, 2018.

[114] J. Ho and S. Ermon. Generative adversarial imitation learning. In *Advances in Neural Inf. Proc. Syst. (NIPS)*, pages 4565–4573, 2016.

[115] S. Hoermann, D. Stumper, and K. Dietmayer. Probabilistic long-term prediction for autonomous vehicles. In *Proc. of the IEEE Intell. Veh. Symp. (IV)*, pages 237–243, 2017.

[116] M. W. Hofbaur and B. C. Williams. Hybrid estimation of complex systems. *IEEE Trans. on Syst., Man, and Cybernetics, Part B (SMCB)*, 34 (5):2178–2191, 2004.

[117] J. Hong, B. Sapp, and J. Philbin. Rules of the road: Predicting driving behavior with a convolutional model of semantic interactions. In *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, pages 8454–8462, 2019.

[118] Y. Hu, W. Zhan, and M. Tomizuka. Probabilistic prediction of vehicle semantic intention and motion. In *Proc. of the IEEE Intell. Veh. Symp. (IV)*, pages 307–313, 2018.

[119] S. Huang, X. Li, Z. Zhang, Z. He, F. Wu, W. Liu, J. Tang, and Y. Zhuang. Deep learning driven visual path prediction from a single image. *IEEE Trans. on Image Processing (TIP)*, 25(12):5892–5904, 2016.

[120] Y. Huang, H. Bi, Z. Li, T. Mao, and Z. Wang. STGAT: Modeling spatial-temporal interactions for human trajectory prediction. In *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, pages 6272–6281, 2019.

[121] R. Hug, S. Becker, W. Hübner, and M. Arens. Quantifying the complexity of standard benchmarking datasets for long-term human trajectory prediction. *arXiv preprint arXiv:2005.13934*, 2020.

[122] R. Hug, S. Becker, W. Hübner, and M. Arens. A short note on analyzing sequence complexity in trajectory prediction benchmarks. 2020.

[123] M. Huynh and G. Alaghband. Trajectory prediction by coupling scene-LSTM with human movement LSTM. In *Int. Symposium on Visual Computing*, pages 244–259. Springer, 2019.

[124] T. Ikeda, Y. Chigodo, D. Rea, F. Zanlungo, M. Shiomi, and T. Kanda. Modeling and prediction of pedestrian behavior based on the sub-goal concept. *Proc. of the Robotics: Science and Systems (RSS)*, 8, 2012.

[125] ISO 13482:2014. Robots and robotic devices – Safety requirements for personal care robots, 2014.

[126] ISO 15622:2018. Intelligent transport systems – Adaptive cruise control systems – Performance requirements and test procedures, 2018.

[127] B. Ivanovic and M. Pavone. The Trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, pages 2375–2384, 2019.

[128] R. Izquierdo, A. Quintanar, I. Parra, D. Fernández-Llorca, and M. Sotelo. Vehicle trajectory prediction in crowded highway scenarios using bird eye view representations and cnns. In *Proc. of the IEEE Int. Conf. on Intell. Transp. Syst. (ITSC)*, pages 1–6. IEEE, 2020.

[129] H. O. Jacobs, O. K. Hughes, M. Johnson-Roberson, and R. Vasudevan. Real-time certified probabilistic pedestrian forecasting. *IEEE Robotics and Automation Letters*, 2(4):2064–2071, Oct 2017. doi: 10.1109/LRA.2017.2719762.

[130] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena. Structural-RNN: Deep learning on spatio-temporal graphs. In *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, pages 5308–5317, 2016.

[131] A. Jain, S. Casas, R. Liao, Y. Xiong, S. Feng, S. Segal, and R. Urtasun. Discrete residual flow for probabilistic pedestrian behavior prediction. *arXiv:1910.08041*, 2019.

[132] L. Janson, B. Ichter, and M. Pavone. Deterministic sampling-based motion planning: Optimality, complexity, and performance. *Int. J. of Robotics Research*, 37(1):46–61, 2018.

[133] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR) Workshops*, pages 11–19, 2017.

[134] J. Joseph, F. Doshi-Velez, A. S. Huang, and N. Roy. A bayesian nonparametric approach to modeling motion patterns. *J. of Autonomous Robots*, 31(4):383, 2011.

[135] N. Kaempchen, K. Weiss, M. Schaefer, and K. C. J. Dietmayer. IMM object tracking for high dynamic driving maneuvers. In *Proc. of the IEEE Intell. Veh. Symp. (IV)*, pages 825–830, 2004.

[136] E. Käfer, C. Hermes, C. Wöhler, H. Ritter, and F. Kummert. Recognition of situation classes at road intersections. In *Proc. of the IEEE*

Int. Conf. on Robotics and Automation (ICRA), pages 3960–3965, 2010.

[137] M. M. Kalayeh, S. Mussmann, A. Petrakova, N. d. V. Lobo, and M. Shah. Understanding trajectory behavior: A motion pattern approach. *arXiv:1501.00614*, 2015.

[138] S. Karaman and E. Frazzoli. Sampling-based algorithms for optimal motion planning. *Int. J. of Robotics Research*, 30(7):846–894, 2011.

[139] I. Karamouzas and M. Overmars. A velocity-based approach for simulating human collision avoidance. In *Proc. of the Int. Conf. on Intelligent Virtual Agents*, pages 180–186. Springer, 2010.

[140] I. Karamouzas and M. Overmars. Simulating and evaluating the local behavior of small pedestrian groups. *IEEE Trans. on Visualization and Computer Graphics*, 18(3), 2012.

[141] I. Karamouzas, P. Heil, P. van Beek, and M. H. Overmars. A predictive collision avoidance model for pedestrian simulation. In *Int. Workshop on Motion in Games*, pages 41–52. Springer, 2009.

[142] V. Karasev, A. Ayvaci, B. Heisele, and S. Soatto. Intent-aware long-term prediction of pedestrian motion. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 2543–2549, 2016.

[143] C. G. Keller and D. M. Gavrila. Will the pedestrian cross? A study on pedestrian path prediction. *IEEE Trans. on Intell. Transp. Syst. (TITS)*, 15(2):494–506, 2014.

[144] C. G. Keller, C. Hermes, and D. M. Gavrila. Will the pedestrian cross? Probabilistic path prediction based on learned motion features. In *Pattern Recognition: DAGM Symposium*, pages 386–395, 2011.

[145] M. Khakzar, A. Rakotonirainy, A. Bond, and S. G. Dehkordi. A dual learning model for vehicle trajectory prediction. *IEEE Access*, 8:21897–21908, 2020.

[146] P. Kiefer, I. Giannopoulos, M. Raubal, and A. Duchowski. Eye tracking for spatial research: Cognition, computation, challenges. *Spatial Cognition & Computation*, 17(1-2):1–19, 2017.

[147] B. Kim, C. M. Kang, J. Kim, S. H. Lee, C. C. Chung, and J. W. Choi. Probabilistic vehicle trajectory prediction over occupancy grid map via recurrent neural network. In *Proc. of the IEEE Int. Conf. on Intell. Transp. Syst. (ITSC)*, pages 399–404, 2017.

[148] H. Kim, D. Kim, G. Kim, J. Cho, and K. Huh. Multi-head attention based probabilistic vehicle trajectory prediction. In *Proc. of the IEEE Intell. Veh. Symp. (IV)*, pages 1720–1725. IEEE, 2020.

[149] K. Kim, D. Lee, and I. Essa. Gaussian process regression flow for analysis of motion trajectories. In *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, pages 1164–1171, 2011.

[150] S. Kim, S. J. Guy, W. Liu, D. Wilkie, R. W. H. Lau, M. C. Lin, and D. Manocha. BRVO: Predicting pedestrian trajectories using velocity-space reasoning. *Int. J. of Robotics Research*, 34(2):201–217, 2015.

[151] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

[152] T. Kirubarajan, Y. Bar-Shalom, K. R. Pattipati, and I. Kadar. Ground target tracking with variable structure IMM estimator. *IEEE Trans. on Aerospace and Electronic Syst. (AESS)*, 36(1):26–46, Jan 2000.

[153] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert. Activity forecasting. In *Proc. of the Europ. Conf. on Comp. Vision (ECCV)*, pages 201–214. Springer, 2012.

[154] S. Köhler, M. Goldhammer, K. Zindler, K. Doll, and K. Dietmeyer. Stereo-vision-based pedestrian's intention detection in a moving vehicle. In *Proc. of the IEEE Int. Conf. on Intell. Transp. Syst. (ITSC)*, pages 2317–2322, 2015.

[155] D. Koller, N. Friedman, and F. Bach. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[156] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila. Context-based pedestrian path prediction. In *Proc. of the Europ. Conf. on Comp. Vision (ECCV)*, pages 618–633. Springer, 2014.

[157] J. F. P. Kooij, F. Flohr, E. A. I. Pool, and D. M. Gavrila. Context-based path prediction for targets with switching dynamics. *Int. J. of Comp. Vision (IJCV)*, 127(3):239–262, 2019.

[158] V. Kosaraju, A. Sadeghian, R. Martín-Martín, I. Reid, S. H. Rezatofighi, and S. Savarese. Social-BiGAT: Multimodal trajectory forecasting using Bicycle-GAN and graph attention networks. *arXiv:1907.03395*, 2019.

[159] M. Koschi, C. Pek, M. Beikirch, and M. Althoff. Set-based prediction of pedestrians in urban environments considering formalized traffic rules. In *Proc. of the IEEE Int. Conf. on Intell. Transp. Syst. (ITSC)*, 2018.

[160] P. Kothari, S. Kreiss, and A. Alahi. Human trajectory forecasting in crowds: A deep learning perspective. *arXiv:2007.03639*, 2020.

[161] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein. The highD dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In *Proc. of the IEEE Int. Conf. on Intell. Transp. Syst. (ITSC)*, 2018.

[162] H. Kretzschmar, M. Kuderer, and W. Burgard. Learning to predict trajectories of cooperatively navigating agents. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 4015–4020, 2014.

[163] E. Kruse and F. M. Wahl. Camera-based observation of obstacle motions to derive statistical data for mobile robot motion planning. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, volume 1, pages 662–667, 1998.

[164] T. Kruse, A. K. Pandey, R. Alami, and A. Kirsch. Human-aware robot navigation: A survey. *J. of Robotics and Autonomous Systems*, 61(12): 1726–1743, 2013.

[165] T. P. Kucner, J. Saarinen, M. Magnusson, and A. J. Lilienthal. Conditional transition maps: Learning motion patterns in dynamic environments. In *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*, pages 1196–1201, 2013.

[166] T. P. Kucner, M. Magnusson, E. Schaffernicht, V. H. Bennetts, and A. J. Lilienthal. Enabling flow awareness for mobile robots in partially observable environments. *IEEE Robotics and Automation Letters*, 2(2): 1093–1100, 2017.

[167] M. Kuderer, H. Kretzschmar, C. Sprunk, and W. Burgard. Feature-based prediction of trajectories for socially compliant navigation. In *Proc. of the Int. Conf. on Robotics: Science and Systems*, 2012.

[168] A. Kuefler, J. Morton, T. Wheeler, and M. Kochenderfer. Imitating driver behavior with generative adversarial networks. In *Proc. of the IEEE Intell. Veh. Symp. (IV)*, pages 204–211, 2017.

[169] F. Kuhnt, R. Kohlhaas, T. Schamm, and J. M. Zöllner. Towards a unified traffic situation estimation model-street-dependent behaviour and motion models. In *Proc. of the IEEE Int. Conf. on Information Fusion (Fusion)*, pages 1223–1229, 2015.

[170] F. Kuhnt, J. Schulz, T. Schamm, and J. M. Zöllner. Understanding interactions between traffic participants based on learned behaviors. In *Proc. of the IEEE Intell. Veh. Symp. (IV)*, pages 1271–1278, 2016.

[171] S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[172] Y. Kuwata, J. Teo, G. Fiore, S. Karaman, E. Frazzoli, and J. P. How. Real-time motion planning with applications to autonomous urban driving. *IEEE Trans. on Control Syst. Techn.*, 17(5):1105–1118, Sept 2009.

[173] P. A. Lasota and J. A. Shah. A multiple-predictor approach to human motion prediction. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 2300–2307, 2017.

[174] P. A. Lasota, T. Fong, and J. A. Shah. A survey of methods for safe human-robot interaction. *Foundations and Trends in Robotics*, 5(4): 261–349, 2017.

[175] B. Lau, K. O. Arras, and W. Burgard. Tracking groups of people with a multi-model hypothesis tracker. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2009.

[176] B. Lau, K. O. Arras, and W. Burgard. Multi-model hypothesis group tracking and group size estimation. *Int. Journal of Social Robotics*, 2(1), March 2010.

[177] J.-G. Lee, J. Han, and K.-Y. Whang. Trajectory clustering: A partition-and-group framework. In *Proc. of the 2007 ACM SIGMOD Int. Conf. on Management of Data*, SIGMOD '07, pages 593–604. ACM, 2007.

[178] N. Lee and K. M. Kitani. Predicting wide receiver trajectories in american football. In *Proc. of the IEEE Winter Conf. on Applications of Computer Vision (WACV)*, pages 1–9, 2016.

[179] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. S. Torr, and M. Chandraker. DESIRE: Distant future prediction in dynamic scenes with interacting agents. In *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, pages 336–345, 2017.

[180] S. Lefèvre, D. Vasquez, and C. Laugier. A survey on motion prediction and risk assessment for intelligent vehicles. *Robomech Journal*, 1(1), 2014.

[181] A. Lerner, Y. Chrysanthou, and D. Lischinski. Crowds by example. In *Computer Graphics Forum*, volume 26, pages 655–664. Wiley Online Library, 2007.

[182] S. Levine, Z. Popovic, and V. Koltun. Feature construction for inverse reinforcement learning. In *Advances in Neural Inf. Proc. Syst. (NIPS)*, pages 1342–1350, 2010.

[183] S. Levine, Z. Popovic, and V. Koltun. Nonlinear inverse reinforcement learning with gaussian processes. In *Advances in Neural Inf. Proc. Syst. (NIPS)*, pages 19–27, 2011.

[184] J. Li, H. Ma, W. Zhan, and M. Tomizuka. Coordination and trajectory prediction for vehicle interactions via bayesian generative modeling. *arXiv:1905.00587*, 2019.

[185] X. R. Li and V. P. Jilkov. Survey of maneuvering target tracking. part I: Dynamic models. *IEEE Trans. on Aerospace and Electronic Syst. (AESS)*, 39(4):1333–1364, Oct 2003.

[186] X. R. Li and V. P. Jilkov. Survey of maneuvering target tracking. part V: Multiple-model methods. *IEEE Trans. on Aerospace and Electronic Syst. (AESS)*, 41(4):1255–1321, Oct 2005.

[187] X. R. Li and V. P. Jilkov. Survey of maneuvering target tracking. part II: Motion models of ballistic and space targets. *IEEE Trans. on Aerospace and Electronic Syst. (AESS)*, 46(1):96–119, Jan 2010.

[188] Y. Li, J. Song, and A. Ermon. Infogail: Interpretable imitation learning from visual demonstrations. In *Advances in Neural Inf. Proc. Syst. (NIPS)*, pages 3812–3822, 2017.

[189] J. Liang, L. Jiang, K. Murphy, T. Yu, and A. Hauptmann. The garden of forking paths: Towards multi-future trajectory prediction. In *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, pages 10508–10518, 2020.

[190] L. Liao, D. Fox, J. Hightower, H. Kautz, and D. Schulz. Voronoi tracking: Location estimation using sparse and noisy sensor data. In *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*, volume 1, pages 723–728, 2003.

[191] M. Liebner, F. Klanner, M. Baumann, C. Ruhhammer, and C. Stiller. Velocity-based driver intent inference at urban intersections in the presence of preceding vehicles. *IEEE Intell. Transp. Syst. Mag.*, 5(2):10–21, 2013.

[192] M. Lindauer, K. Eggensperger, M. Feurer, S. Falkner, A. Biedenkapp, and F. Hutter. Smac v3: Algorithm configuration in python. `https://github.com/automl/SMAC3`, 2017.

[193] T. Linder and K. O. Arras. Multi-model hypothesis tracking of groups of people in rgb-d data. In *Proc. of the IEEE Int. Conf. on Information Fusion (Fusion)*, pages 1–7. IEEE, 2014.

[194] T. Linder, S. Breuers, B. Leibe, and K. O. Arras. On multi-modal people tracking from mobile platforms in very crowded and dynamic environments. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2016.

[195] M. L. Littman, T. L. Dean, and L. P. Kaelbling. On the complexity of solving markov decision problems. In *Proc. of the Conf. on Uncertainty in Artificial Intelligence (UAI)*, pages 394–402. Morgan Kaufmann Publ. Inc., 1995.

[196] S.-Y. Lo, S. Alkoby, and P. Stone. Robust motion planning and safety benchmarking in human workspaces. In *Workshop Proc. of the AAAI Conf. on Artificial Intelligence*, 2019.

[197] M. Luber and K. O. Arras. Multi-hypothesis social grouping and tracking for mobile robots. In *Proc. of the Robotics: Science and Systems (RSS)*, Berlin, Germany, 2013.

[198] M. Luber, J. A. Stork, G. D. Tipaldi, and K. O. Arras. People tracking with human motion predictions from social forces. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 464–469, 2010.

[199] M. Luber, G. D. Tipaldi, and K. O. Arras. Place-dependent people tracking. *Int. J. of Robotics Research*, 30(3):280–293, 2011.

[200] M. Luber, L. Spinello, J. Silva, and K. O. Arras. Socially-aware robot navigation: A learning approach. In *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*, pages 902–907, 2012.

[201] Y. Luo and P. Cai. GAMMA: A general agent motion prediction model for autonomous driving. *arXiv:1906.01566*, 2019.

[202] W.-C. Ma, D.-A. Huang, N. Lee, and K. M. Kitani. Forecasting interactive dynamics of pedestrians with fictitious play. In *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, pages 4636–4644, 2017.

[203] D. J. MacKay and D. J. Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[204] J. Mainprice, A. Byravan, D. Kappler, D. Fox, S. Schaal, and N. Ratliff. Functional manifold projections in deep-learch. In *Advances in Neural Inf. Proc. Syst. (NIPS)Workshop*, 2016.

[205] B. Majecka. Statistical models of pedestrian behaviour in the forum. *Master's thesis, School of Informatics, University of Edinburgh*, 2009.

[206] O. Makansi, E. Ilg, O. Cicek, and T. Brox. Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction. In *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, pages 7144–7153, 2019.

[207] D. Makris and T. Ellis. Path detection in video surveillance. *Image and Vision Computing*, 20(12):895–903, 2002.

[208] C. I. Mavrogiannis and R. A. Knepper. Decentralized multi-agent navigation planning with braids. In *Proceedings of the 2016 International Workshop on the Algorithmic Foundations of Robotics (WAFR'16)*, 2016.

[209] E. Mazor, A. Averbuch, Y. Bar-Shalom, and J. Dayan. Interacting multiple model methods in target tracking: a survey. *IEEE Trans. on Aerospace and Electronic Syst. (AESS)*, 34(1):103–123, 1998.

[210] K. Messaoud, I. Yahiaoui, A. Verroust-Blondet, and F. Nashashibi. Relational recurrent neural networks for vehicle trajectory prediction. In *Proc. of the IEEE Int. Conf. on Intell. Transp. Syst. (ITSC)*, pages 1813–1818. IEEE, 2019.

[211] A. M. Metelli, M. Pirotta, and M. Restelli. Compatible reward inverse reinforcement learning. In *Advances in Neural Inf. Proc. Syst. (NIPS)*, pages 2050–2059, 2017.

[212] R. Q. Mínguez, I. P. Alonso, D. Fernández-Llorca, and M. Á. Sotelo. Pedestrian path, pose, and intention prediction through gaussian process dynamical models and pedestrian activity recognition. In *IEEE Trans. on Intell. Transp. Syst. (TITS)*, pages 1–12, 2018.

[213] A. Møgelmose, M. M. Trivedi, and T. B. Moeslund. Trajectory analysis and prediction for improved pedestrian safety: Integrated framework and evaluations. In *Proc. of the IEEE Intell. Veh. Symp. (IV)*, pages 330–335, 2015.

[214] S. Molina, G. Cielniak, T. Krajník, and T. Duckett. Modelling and predicting rhythmic flow patterns in dynamic environments. In *Annual Conf. Towards Autonom. Rob. Syst.*, pages 135–146. Springer, 2018.

[215] B. T. Morris and M. M. Trivedi. A survey of vision-based trajectory learning and analysis for surveillance. *IEEE Trans. on Circuits and Systems for Video Technology*, 18(8):1114–1127, 2008.

[216] M. Moussaïd, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz. The walking behaviour of pedestrian social groups and its impact on crowd dynamics. *PloS one*, 5(4):e10047, 2010.

[217] C. Muench and D. M. Gavrila. Composable Q-functions for pedestrian car interactions. In *Proc. of the IEEE Intell. Veh. Symp. (IV)*. IEEE, 2019.

[218] S. Mukherjee, S. Wang, and A. Wallace. Interacting vehicle trajectory prediction with convolutional recurrent neural networks. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 4336–4342. IEEE, 2020.

[219] D. Munoz, J. A. Bagnell, and M. Hebert. Stacked hierarchical labeling. In *Proc. of the Europ. Conf. on Comp. Vision (ECCV)*. Springer, 2010.

[220] V. Murino, M. Cristani, S. Shah, and S. Savarese. *Group and Crowd Behavior for Computer Vision*. Academic Press, 2017.

[221] Q. P. Nguyen, B. K. H. Low, and P. Jaillet. Inverse reinforcement learning with locally consistent reward functions. In *Advances in Neural Inf. Proc. Syst. (NIPS)*, pages 1747–1755, 2015.

[222] N. Nikhil and B. Tran Morris. Convolutional neural network for trajectory prediction. In *Proc. of the Europ. Conf. on Comp. Vision (ECCV)*, pages 0–0, 2018.

[223] B. Noe and N. Collins. Variable structure interacting multiple-model filter (VS-IMM) for tracking targets with transportation network constraints. In *SPIE Proc. of Sign. and Data Proc. of Small Targets*, volume 4048, 2000.

[224] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai. A large-scale benchmark dataset for event recognition in surveillance video. In *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, pages 3153–3160, 2011.

[225] B. Okal and K. O. Arras. Learning socially normative robot navigation behaviors with bayesian inverse reinforcement learning. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2016.

[226] S. Oli, B. L'Esperance, and K. Gupta. Human motion behaviour aware planner (hmbap) for path planning in dynamic human environments. In *Proc. of the IEEE Int. Conf. on Adv. Robotics (ICAR)*, pages 1–7, 2013.

[227] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, and J. Peters. An algorithmic perspective on imitation learning. *Foundations and Trends in Robotics*, 7(1-2):1–179, 2018.

[228] O. Palinko, F. Rea, G. Sandini, and A. Sciutti. Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration. In *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*, 2016.

[229] L. Palmieri, T. P. Kucner, M. Magnusson, A. J. Lilienthal, and K. O. Arras. Kinodynamic motion planning on gaussian mixture fields. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 6176–6181. IEEE, 2017.

[230] L. Palmieri, A. Rudenko, J. Mainprice, and K. O. Arras. Special Issue on Long-term Human Motion Prediction. In *IEEE Robotics and Automation Letters*, 2020.

[231] X. Pan, Y. He, H. Wang, W. Xiong, and X. Peng. Mining regular behaviors based on multidimensional trajectories. *Expert Systems with Applications*, 66:106–113, 2016.

[232] B. Pannetier, K. Benameur, V. Nimier, and M. Rombaut. VS-IMM using road map information for a ground target tracking. In *Proc. of the IEEE Int. Conf. on Information Fusion (Fusion)*, 2005.

[233] S. Paris, J. Pettré, and S. Donikian. Pedestrian reactive navigation for crowd simulation: a predictive approach. In *Computer Graphics Forum*, volume 26, pages 665–674. Wiley Online Library, 2007.

[234] S. H. Park, B. Kim, C. M. Kang, C. C. Chung, and J. W. Choi. Sequence-to-sequence prediction of vehicle trajectory via LSTM encoder-decoder architecture. In *Proc. of the IEEE Intell. Veh. Symp. (IV)*, pages 1672–1678. IEEE, 2018.

[235] Z. Pei, X. Qi, Y. Zhang, M. Ma, and Y.-H. Yang. Human trajectory prediction in crowded scene using social-affinity long short-term memory. *Pattern Recognition*, 93:273–282, 2019.

[236] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, pages 261–268, 2009.

[237] S. Pellegrini, A. Ess, and L. van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *Proc. of the Europ. Conf. on Comp. Vision (ECCV)*, pages 452–465. Springer, 2010.

[238] A. Pentland and A. Liu. Modeling and prediction of human behavior. *Neural computation*, 11(1):229–242, 1999.

[239] D. Petrich, T. Dang, D. Kasper, G. Breuel, and C. Stiller. Map-based long term motion prediction for vehicles in traffic environments. In *Proc. of the IEEE Int. Conf. on Intell. Transp. Syst. (ITSC)*, pages 2166–2172, 2013.

[240] J. Pettré, J. Ondřej, A.-H. Olivier, A. Cretual, and S. Donikian. Experiment-based modeling, simulation and validation of interactions between virtual walkers. In *Proc. of the ACM SIGGRAPH/Eurographics Symp. on Comp. Anim.*, pages 189–198, 2009.

[241] M. Pfeiffer, U. Schwesinger, H. Sommer, E. Galceran, and R. Siegwart. Predicting actions to act predictably: Cooperative partial motion planning with maximum entropy models. In *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*, pages 2096–2101, 2016.

[242] M. Pfeiffer, G. Paolo, H. Sommer, J. Nieto, R. Siegwart, and C. Cadena. A data-driven model for interaction-aware pedestrian motion prediction in object cluttered environments. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 1–8, 2018.

[243] C. Piciarelli, G. L. Foresti, and L. Snidaro. Trajectory clustering and its applications for video surveillance. In *IEEE Conf. on Advanced Video and Signal Based Surveillance*, pages 40–45. IEEE, 2005.

[244] F. Poiesi and A. Cavallaro. Predicting and recognizing human interactions in public spaces. *Journal of Real-Time Image Processing*, 10(4): 785–803, 2015.

[245] E. A. Pool, J. F. Kooij, and D. M. Gavrila. Context-based cyclist path prediction using recurrent neural networks. In *Proc. of the IEEE Intell. Veh. Symp. (IV)*, pages 824–830. IEEE, 2019.

[246] E. A. I. Pool, J. F. P. Kooij, and D. M. Gavrila. Using road topology to improve cyclist path prediction. In *Proc. of the IEEE Intell. Veh. Symp. (IV)*, pages 289–296, 2017.

[247] F. Previtali, A. Bordallo, L. Iocchi, and S. Ramamoorthy. Predicting future agent motions for dynamic environments. In *Proc. of the IEEE Int. Conf. on Mach. Learning and App. (ICMLA)*, pages 94–99, 2016. doi: 10.1109/ICMLA.2016.0024.

[248] F. Qiu and X. Hu. Modeling group structures in pedestrian crowd simulation. *Simulation Modelling Practice and Theory*, 18(2):190–205, 2010.

[249] J. Quehl, H. Hu, Ö. c. Taş, E. Rehder, and M. Lauer. How good is my prediction? Finding a similarity measure for trajectory prediction evaluation. In *Proc. of the IEEE Int. Conf. on Intell. Transp. Syst. (ITSC)*, pages 1–6, 2017.

[250] R. Quintero, J. Almeida, D. F. Llorca, and M. A. Sotelo. Pedestrian path prediction using body language traits. In *Proc. of the IEEE Intell. Veh. Symp. (IV)*, pages 317–323, 2014.

[251] N. Radwan, A. Valada, and W. Burgard. Multimodal interaction-aware motion prediction for autonomous street crossing. *arXiv:1808.06887*, 2018.

[252] G. Raipuria, F. Gaisser, and P. P. Jonker. Road infrastructure indicators for trajectory prediction. In *Proc. of the IEEE Intell. Veh. Symp. (IV)*, pages 537–543, 2018.

[253] A. Rasouli and J. K. Tsotsos. Autonomous vehicles that interact with pedestrians: A survey of theory and practice. *IEEE Trans. on Intell. Transp. Syst. (TITS)*, 2019.

[254] E. Rehder and H. Klöden. Goal-directed pedestrian prediction. In *Proc. of the Int. Conf. on Comp. Vision Worksh.*, pages 139–147, 2015.

[255] E. Rehder, F. Wirth, M. Lauer, and C. Stiller. Pedestrian prediction by planning using deep neural networks. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 1–5, 2018.

[256] N. Rhinehart and K. Kitani. First-person activity forecasting from video with online inverse reinforcement learning. *IEEE Trans. on Patt. Anal. and Mach. Intell. (PAMI)*, 2018.

[257] N. Rhinehart, K. Kitani, and P. Vernaza. R2P2: A ReparameteRized Pushforward Policy for diverse, precise generative path forecasting. In *Proc. of the Europ. Conf. on Comp. Vision (ECCV)*, pages 772–788, 2018.

[258] N. Rhinehart, R. McAllister, and S. Levine. Deep Imitative Models for Flexible Inference, Planning, and Control. *arXiv:1810.06544*, Oct. 2018.

[259] N. Rhinehart, R. McAllister, K. Kitani, and S. Levine. PRECOG: Prediction conditioned on goals in visual multi-agent settings. In *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, October 2019.

[260] D. Ridel, E. Rehder, M. Lauer, C. Stiller, and D. Wolf. A literature review on the prediction of pedestrian behavior in urban scenarios. In *Proc. of*

the IEEE Int. Conf. on Intell. Transp. Syst. (ITSC), pages 3105–3112. IEEE, 2018.

[261] D. Ridel, N. Deo, D. Wolf, and M. Trivedi. Scene compliant trajectory forecast with agent-centric spatio-temporal grids. *arXiv:1909.07507*, 2019.

[262] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *Proc. of the Europ. Conf. on Comp. Vision (ECCV)*, pages 549–565. Springer, 2016.

[263] C. Rösmann, F. Hoffmann, and T. Bertram. Timed-elastic-bands for time-optimal point-to-point nonlinear model predictive control. In *Proc. of the Europ. Control Conf. (ECC)*, pages 3352–3357. IEEE, 2015.

[264] C. Rösmann, M. Oeljeklaus, F. Hoffmann, and T. Bertram. Online trajectory prediction and planning for social robot navigation. In *Proc. of the IEEE Int. Conf. on Advanced Intelligent Mechatronics (AIM)*, pages 1255–1260, 2017.

[265] M. Roth, F. Flohr, and D. M. Gavrila. Driver and pedestrian awareness-based collision risk analysis. In *Proc. of the IEEE Intell. Veh. Symp. (IV)*, pages 454–459, 2016.

[266] A. Rudenko, L. Palmieri, and K. O. Arras. Predictive planning for a mobile robot in human environments. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA), Workshop on AI Planning and Robotics*, 2017.

[267] A. Rudenko, L. Palmieri, and K. O. Arras. Joint prediction of human motion using a planning-based social force approach. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 1–7, 2018.

[268] A. Rudenko, L. Palmieri, A. J. Lilienthal, and K. O. Arras. Human motion prediction under social grouping constraints. In *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*, 2018.

[269] A. Rudenko, T. Kucner, C. Swaminathan, R. Chadalavada, K. O. Arras, and A. J. Lilienthal. Benchmarking human motion prediction methods. In *Proc. of the ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI), Workshop on Test Methods and Metrics for Effective HRI in Real World Human-Robot Teams*, 2020.

[270] A. Rudenko, T. P. Kucner, C. S. Swaminathan, R. T. Chadalavada, K. O. Arras, and A. J. Lilienthal. THÖR: Human-robot navigation data collection and accurate motion trajectories dataset. *IEEE Robotics and Automation Letters*, 5(2):676–682, 2020.

[271] A. Rudenko, L. Palmieri, A. Alahi, J. Mainprice, and K. O. Arras. 2nd Workshop on Long-term Human Motion Prediciton (LHMP 2020). In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2020. URL `https://motionpredictionicra2020.github.io`.

[272] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras. Human motion trajectory prediction: A survey. *Int. J. of Robotics Research*, 39(8):895–935, 2020.

[273] A. Rudenko, L. Palmieri, K. O. Arras, A. Bajcsy, A. Alahi, and A. J. Lilienthal. 3rd Workshop on Long-term Human Motion Prediciton (LHMP 2021). In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2021. URL `https://motionpredictionicra2021.github.io`.

[274] A. Rudenko, L. Palmieri, J. Doellinger, A. J. Lilienthal, and K. O. Arras. Learning occupancy priors of human motion from semantic maps of urban environments. *IEEE Robotics and Automation Letters*, 6(2): 3248–3255, 2021.

[275] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, 2016.

[276] A. Sadeghian, V. Kosaraju, A. Gupta, S. Savarese, and A. Alahi. TrajNet: Towards a benchmark for human trajectory prediction. *arXiv preprint*, 2018.

[277] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, and S. Savarese. SoPhie: An attentive GAN for predicting paths compliant to social and physical constraints. In *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, pages 1349–1358, 2019.

[278] K. Saleh, M. Hossny, and S. Nahavandi. Cyclist trajectory prediction using bidirectional recurrent neural networks. In *Australasian Joint Conference on Artificial Intelligence*, pages 284–295. Springer, 2018.

[279] K. Saleh, M. Hossny, and S. Nahavandi. Intent prediction of pedestrians via motion trajectories using stacked recurrent neural networks. *IEEE Trans. on Intelligent Vehicles*, 3(4):414–424, 2018.

[280] K. Saleh, M. Hossny, and S. Nahavandi. Contextual recurrent predictive model for long-term intent prediction of vulnerable road users. *IEEE Trans. on Intell. Transp. Syst. (TITS)*, 2019.

[281] E. Schmerling, K. Leung, W. Vollprecht, and M. Pavone. Multimodal probabilistic model-based planning for human-robot interaction. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 1–9, 2018.

[282] F. Schneemann and P. Heinemann. Context-based detection of pedestrian crossing intention for autonomous driving in urban environments. In *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*, pages 2243–2248, 2016.

[283] N. Schneider and D. M. Gavrila. Pedestrian path prediction with recursive Bayesian filters: A comparative study. In *Proc. of the German Conf. on Pattern Recognition*, pages 174–183. Springer, 2013.

[284] C. Schöller, V. Aravantinos, F. Lay, and A. Knoll. The simpler the better: Constant velocity for pedestrian motion prediction. *arXiv:1903.07933*, 2019.

[285] A. Schrijver. On the history of the shortest path problem. In *Documenta Mathematica*, 2012.

[286] R. Schubert, E. Richter, and G. Wanielik. Comparison and evaluation of advanced motion models for vehicle tracking. In *Proc. of the IEEE Int. Conf. on Information Fusion (Fusion)*, pages 1–6, 2008.

[287] A. T. Schulz and R. Stiefelhagen. A controlled interactive multiple model filter for combined pedestrian intention recognition and path prediction. In *Proc. of the IEEE Int. Conf. on Intell. Transp. Syst. (ITSC)*, pages 173–178, 2015.

[288] M. Seitz, G. Köster, and A. Pfaffinger. Pedestrian group behavior in a cellular automaton. In *Pedestrian and Evacuation Dynamics*, pages 807–814. 2012.

[289] J. A. Sethian. A fast marching level set method for monotonically advancing fronts. *Proc. of the National Academy of Sciences*, 93(4):1591–1595, 1996.

[290] S. Shalev-Shwartz, N. Ben-Zrihem, A. Cohen, and A. Shashua. Long-term planning by short-term prediction. *arXiv:1602.01580*, 2016.

[291] P. J. Shea, T. Zadra, D. M. Klamer, E. Frangione, and R. Brouillard. Improved state estimation through use of roads in ground tracking. In *SPIE Proc. of Sign. and Data Proc. of Small Targets*, volume 4048, pages 321–333, 2000.

[292] M. Shen, G. Habibi, and J. P. How. Transferable pedestrian motion prediction models at intersections. In *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*, pages 4547–4553. IEEE, 2018.

[293] X. Shi, X. Shao, Z. Guo, G. Wu, H. Zhang, and R. Shibasaki. Pedestrian trajectory prediction in extremely crowded scenarios. *Sensors*, 19(5): 1223, 2019.

[294] D. Simon. Kalman filtering with state constraints: a survey of linear and nonlinear algorithms. *IET Control Theory and Applications*, 4:1303–1318, August 2010.

[295] H. Singh, R. Arter, L. Dodd, P. Langston, E. Lester, and J. Drury. Modelling subgroup behaviour in crowd dynamics DEM simulation. *Applied Mathematical Modelling*, 33(12):4408–4423, 2009.

[296] J. Šochman and D. C. Hogg. Who knows who-inverting the social force model for finding groups. In *Proc. of the Int. Conf. on Comp. Vision Worksh.*, 2011.

[297] S. Srikanth, J. A. Ansari, K. R. Ram, S. Sharma, J. K. Murthy, and K. M. Krishna. INFER: INtermediate representations for FuturE pRediction. *arXiv:1903.10641*, 2019.

[298] H. Su, J. Zhu, Y. Dong, and B. Zhang. Forecast the plausible paths in crowd scenes. In *IJCAI*, volume 1, page 2, 2017.

[299] N. Sumpter and A. Bulpitt. Learning spatio-temporal patterns for predicting object behaviour. *Image and Vision Computing*, 18(9):697–704, 2000.

[300] L. Sun, Z. Yan, S. M. Mellado, M. Hanheide, and T. Duckett. 3DOF pedestrian trajectory prediction learned from long-term autonomous mobile robot deployment data. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 1–7, 2018.

[301] C. Sung, D. Feldman, and D. Rus. Trajectory clustering for motion prediction. In *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*, pages 1547–1552. IEEE, 2012.

[302] M. S. Suraj, H. Grimmett, L. Platinskỳ, and P. Ondrúška. Predicting trajectories of vehicles using large-scale motion priors. In *Proc. of the IEEE Intell. Veh. Symp. (IV)*, pages 1639–1644, 2018.

[303] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[304] C. S. Swaminathan, T. P. Kucner, M. Magnusson, L. Palmieri, and A. J. Lilienthal. Down the CLiFF: Flow-aware trajectory planning under motion pattern uncertainty. In *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*, pages 7403–7409. IEEE, 2018.

[305] S. Tadokoro, Y. Ishikawa, T. Takebe, and T. Takamori. Stochastic prediction of human motion and control of robots in the service of human. In *Proc. of the IEEE Conf. on Systems, Man, and Cybernetics (SMC)*, volume 1, pages 503–508, 1993.

[306] A. Tamar, Y. Wu, G. Thomas, S. Levine, and P. Abbeel. Value iteration networks. In *Advances in Neural Inf. Proc. Syst. (NIPS)*, pages 2154–2162, 2016.

[307] C. Tang and R. R. Salakhutdinov. Multiple futures prediction. In *Advances in Neural Inf. Proc. Syst. (NeurIPS)*, pages 15398–15408, 2019.

[308] C. Tao, Q. Jiang, L. Duan, and P. Luo. Dynamic and static context-aware lstm for multi-agent motion prediction. *arXiv:2008.00777*, 2020.

[309] M. K. C. Tay and C. Laugier. Modelling smooth paths using gaussian processes. In *Results of the Int. Conf. on Field and Service Robotics*, pages 381–390. Springer, 2008.

[310] L. A. Thiede and P. P. Brahma. Analyzing the variety loss in the context of probabilistic trajectory prediction. In *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, pages 9954–9963, 2019.

[311] S. Thompson, T. Horiuchi, and S. Kagami. A probabilistic model of human motion and navigation intent for mobile robot path planning. In *Proc. of the IEEE Int. Conf. on Autonomous Robots and Agents (ICARA)*, pages 663–668, 2009.

[312] Q. Tran and J. Firl. Online maneuver recognition and multimodal trajectory prediction for intersection assistance using non-parametric regression. In *Proc. of the IEEE Intell. Veh. Symp. (IV)*, pages 918–923, 2014.

[313] P. Trautman and A. Krause. Unfreezing the robot: Navigation in dense, interacting crowds. In *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*, pages 797–803, 2010.

[314] P. Trautman, J. Ma, R. M. Murray, and A. Krause. Robot navigation in dense human crowds: the case for cooperation. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 2153–2160, 2013.

[315] M. Treiber, A. Hennecke, and D. Helbing. Congested traffic states in empirical observations and microscopic simulations. *Physical Review E*, 62(2):1805, 2000.

[316] R. Triebel, K. Arras, R. Alami, L. Beyer, S. Breuers, R. Chatila, M. Chetouani, D. Cremers, V. Evers, M. Fiore, et al. Spencer: A socially aware service robot for passenger guidance and help in busy airports. In *Field and service robotics*, pages 607–622. Springer, 2016.

[317] V. V. Unhelkar, C. Pérez-D'Arpino, L. Stirling, and J. A. Shah. Human-robot co-navigation using anticipatory indicators of human walking motion. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 6183–6190, 2015.

[318] J. van den Berg, M. Lin, and D. Manocha. Reciprocal velocity obstacles for real-time multi-agent navigation. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 1928–1935, 2008.

[319] J. van Den Berg, S. Patil, J. Sewall, D. Manocha, and M. Lin. Interactive navigation of multiple agents in crowded environments. In *Proc. of the ACM Symp. on Interact. 3D Graphics and Games*, pages 139–147, 2008.

[320] T. van der Heiden, N. S. Nagaraja, C. Weiss, and E. Gavves. SafeCritic: Collision-aware trajectory prediction. *arXiv:1910.06673*, 2019.

[321] D. Varshneya and G. Srinivasaraghavan. Human trajectory prediction using spatially aware deep attention models. *arXiv:1705.09436*, 2017.

[322] P. Vasishta, D. Vaufreydaz, and A. Spalanzani. Natural vision based method for predicting pedestrian behaviour in urban environments. In *Proc. of the IEEE Int. Conf. on Intell. Transp. Syst. (ITSC)*, 2017.

[323] P. Vasishta, D. Vaufreydaz, and A. Spalanzani. Building prior knowledge: A Markov based pedestrian prediction model using urban environmental data. In *Proc. of the Int. Conf. on Control, Automation, Robotics and Vision (ICARCV)*, pages 1–12, 2018.

[324] D. Vasquez. Novel planning-based algorithms for human motion prediction. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 3317–3322, 2016.

[325] D. Vasquez, T. Fraichard, O. Aycard, and C. Laugier. Intentional motion on-line learning and prediction. *Machine Vision and Applications*, 19 (*5*):411–425, 2008.

[326] D. Vasquez, T. Fraichard, and C. Laugier. Incremental learning of statistical motion patterns with growing hidden markov models. *IEEE Trans. on Intell. Transp. Syst. (TITS)*, 10(3):403–416, 2009.

[327] A. Vemula, K. Muelling, and J. Oh. Modeling cooperative navigation in dense human crowds. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 1685–1692, 2017.

[328] A. Vemula, K. Muelling, and J. Oh. Social Attention: Modeling Attention in Human Crowds. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2018.

[329] T. Vintr, S. Molina, R. Senanayake, G. Broughton, Z. Yan, J. Ulrich, T. P. Kucner, C. S. Swaminathan, F. Majer, M. Stachová, A. J. Lilienthal, and T. Krajník. Time-varying pedestrian flow models for service robots. In *Proc. of the European Conf. on Mobile Robots (ECMR)*, pages 1–7. IEEE, 2019.

[330] B. Völz, H. Mielenz, R. Siegwart, and J. Nieto. Predicting pedestrian crossing using quantile regression forests. In *Proc. of the IEEE Intell. Veh. Symp. (IV)*, pages 426–432, 2016.

[331] J. Walker, A. Gupta, and M. Hebert. Patch to the future: Unsupervised visual prediction. In *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, pages 3302–3309, 2014.

[332] J. Walker, C. Doersch, A. Gupta, and M. Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *Proc. of the Europ. Conf. on Comp. Vision (ECCV)*, pages 835–851. Springer, 2016.

[333] Z. Wang, P. Jensfelt, and J. Folkesson. Modeling spatial-temporal dynamics of human movements for predicting future trajectories. In *Workshop Proc. of the AAAI Conf. on Artificial Intelligence "Knowledge, Skill, and Behavior Transfer in Autonomous Robots"*, 2015.

[334] Z. Wang, P. Jensfelt, and J. Folkesson. Building a human behavior map from local observations. In *Proc. of the IEEE Int. Symp. on Robot and Human Interactive Comm. (RO-MAN)*, pages 64–70, 2016.

[335] J. Wu, J. Ruenz, and M. Althoff. Probabilistic map-based pedestrian motion prediction taking traffic participants into consideration. In *Proc. of the IEEE Intell. Veh. Symp. (IV)*, pages 1285–1292, 2018.

[336] Y. Wu, J. Hou, G. Chen, and A. Knoll. Trajectory prediction based on planning method considering collision risk. In *Proc. of the IEEE Int. Conf. on Adv. Robotics and Mechatronics (ICARM)*, pages 466–470. IEEE, 2020.

[337] Z. Wu, C. Shen, and A. Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019.

[338] M. Wulfmeier, P. Ondruska, and I. Posner. Maximum entropy deep inverse reinforcement learning. *arXiv:1507.04888*, 2015.

[339] S. Xiao, Z. Wang, and J. Folkesson. Unsupervised robot learning to predict person motion. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 691–696, 2015.

[340] D. Xie, S. Todorovic, and S.-C. Zhu. Inferring "dark matter" and "dark energy" from videos. In *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, pages 2224–2231, Dec 2013. doi: 10.1109/ICCV.2013.277.

[341] G. Xie, H. Gao, L. Qian, B. Huang, K. Li, and J. Wang. Vehicle trajectory prediction by integrating physics-and maneuver-based approaches using interactive multiple models. *IEEE Trans. on Industrial Electronics*, 65 (7):5999–6008, 2018.

[342] K. Xu, E. Ratner, A. Dragan, S. Levine, and C. Finn. Learning a prior over intent via meta-inverse reinforcement learning. *arXiv:1805.12573*, 2018.

[343] Y. Xu, Z. Piao, and S. Gao. Encoding crowd interaction with deep neural network for pedestrian trajectory prediction. In *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, pages 5275–5284, 2018.

[344] H. Xue, D. Q. Huynh, and M. Reynolds. Bi-prediction: pedestrian trajectory prediction based on bidirectional LSTM classification. In *2017 Int. Conf. on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2017.

[345] H. Xue, D. Q. Huynh, and M. Reynolds. SS-LSTM: a hierarchical LSTM model for pedestrian trajectory prediction. In *Proc. of the IEEE Winter Conf. on Applications of Computer Vision (WACV)*, pages 1186–1194. IEEE, 2018.

[346] H. Xue, D. Huynh, and M. Reynolds. Location-velocity attention for pedestrian trajectory prediction. In *Proc. of the IEEE Winter Conf. on Applications of Computer Vision (WACV)*, pages 2038–2047. IEEE, 2019.

[347] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg. Who are you with and where are you going? In *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, pages 1345–1352, June 2011. doi: 10.1109/CVPR.2011.5995468.

[348] X. Yan, I. A. Kakadiaris, and S. K. Shah. Modeling local behavior for predicting social interactions towards human tracking. *Pattern Recognition*, 47(4):1626–1641, 2014.

[349] Z. Yan, T. Duckett, and N. Bellotto. Online learning for human classification in 3D LiDAR-based tracking. In *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*, pages 864–871, 2017.

[350] C. Yang and E. Blasch. Fusion of tracks with road constraints. *J. of Advances in Information Fusion*, 3(1), June 2008.

[351] C. Yang, M. Bakich, and E. Blasch. Nonlinear constrained tracking of targets on roads. In *Proc. of the IEEE Int. Conf. on Information Fusion (Fusion)*, 2005.

[352] H. C. Yen, H. P. Huang, and S. Y. Chung. Goal-directed pedestrian model for long-term motion prediction with application to robot motion planning. In *Proc. of the IEEE Workshop on Advanced Robotics and Its Social Impacts*, pages 1–6, 2008.

[353] S. Yi, H. Li, and X. Wang. Pedestrian behavior modeling from stationary crowds with applications to intelligent surveillance. *IEEE Trans. on Image Processing (TIP)*, 25(9):4354–4368, Sept 2016. ISSN 1057-7149. doi: 10.1109/TIP.2016.2590322.

[354] Y. Yoo, K. Yun, S. Yun, J. Hong, H. Jeong, and J. Young Choi. Visual path prediction in complex scenes with crowded moving objects. In *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, pages 2668–2677, 2016.

[355] L. Yu, T. Yu, C. Finn, and S. Ermon. Meta-inverse reinforcement learning with probabilistic context variables. *arXiv:1909.09314*, 2019.

[356] F. Zanlungo, T. Ikeda, and T. Kanda. Social force model with explicit collision prediction. *EPL (Europhysics Letters)*, 93(6):68005, 2011.

[357] P. Zechel, R. Streiter, K. Bogenberger, and U. Göhner. Pedestrian occupancy prediction for autonomous vehicles. In *IEEE Int. Conf. on Robotic Computing (IRC)*, pages 230–235. IEEE, 2019.

[358] S. Zernetsch, S. Kohnen, M. Goldhammer, K. Doll, and B. Sick. Trajectory prediction of cyclists using a physical model and an artificial neural network. In *Proc. of the IEEE Intell. Veh. Symp. (IV)*, pages 833–838, 2016.

[359] E. Zhan, S. Zheng, Y. Yue, and P. Lucey. Generative multi-agent behavioral cloning. *arXiv:1803.07612*, 2018.

[360] P. Zhang, W. Ouyang, P. Zhang, J. Xue, and N. Zheng. SR-LSTM: State refinement for LSTM towards pedestrian trajectory prediction. In *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, pages 12085–12094, 2019.

[361] Z. Zhang, K. Huang, and T. Tan. Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 1135–1138. IEEE, 2006.

[362] T. Zhao, Y. Xu, M. Monfort, W. Choi, C. Baker, Y. Zhao, Y. Wang, and Y. N. Wu. Multi-agent tensor fusion for contextual trajectory prediction. In *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, pages 12126–12134, 2019.

[363] S. Zheng, Y. Yue, and J. Hobbs. Generating long-term trajectories using deep hierarchical networks. In *Advances in Neural Inf. Proc. Syst. (NIPS)*, pages 1543–1551, 2016.

[364] Y. Zheng. Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(3):29, 2015.

[365] W. Zhi, T. Lai, L. Ott, and F. Ramos. Anticipatory navigation in crowds by probabilistic prediction of pedestrian future movements. *arXiv:2011.06235*, 2020.

[366] B. Zhou, X. Wang, and X. Tang. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, pages 2871–2878. IEEE, 2012.

[367] B. Zhou, X. Tang, and X. Wang. Learning collective crowd behaviors with dynamic pedestrian-agents. *Int. J. of Comp. Vision (IJCV)*, 111(1): 50–68, 2015.

[368] Q. Zhu. Hidden Markov Model for dynamic obstacle avoidance of mobile robot navigation. *IEEE Trans. on Robotics and Automation (TRO)*, 7(3):390–397, 1991.

[369] B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *Proc. of the AAAI Conf. on Artificial Intelligence (AAAI)*, 2008.

[370] B. D. Ziebart, N. Ratliff, G. Gallagher, C. Mertz, K. Peterson, J. A. Bagnell, M. Hebert, A. K. Dey, and S. Srinivasa. Planning-based prediction for pedestrians. In *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*, pages 3931–3936, 2009.

Publications *in the series*
Örebro Studies in Technology

1. Bergsten, Pontus (2001) *Observers and Controllers for Takagi – Sugeno Fuzzy Systems*. Doctoral Dissertation.
2. Iliev, Boyko (2002) *Minimum-time Sliding Mode Control of Robot Manipulators*. Licentiate Thesis.
3. Spännar, Jan (2002) *Grey box modelling for temperature estimation*. Licentiate Thesis.
4. Persson, Martin (2002) *A simulation environment for visual servoing*. Licentiate Thesis.
5. Boustedt, Katarina (2002) *Flip Chip for High Volume and Low Cost – Materials and Production Technology*. Licentiate Thesis.
6. Biel, Lena (2002) *Modeling of Perceptual Systems – A Sensor Fusion Model with Active Perception*. Licentiate Thesis.
7. Otterskog, Magnus (2002) *Produktionstest av mobiltelefonantenner i mod-växlande kammare*. Licentiate Thesis.
8. Tolt, Gustav (2003) *Fuzzy-Similarity-Based Low-level Image Processing*. Licentiate Thesis.
9. Loutfi, Amy (2003) *Communicating Perceptions: Grounding Symbols to Artificial Olfactory Signals*. Licentiate Thesis.
10. Iliev, Boyko (2004) *Minimum-time Sliding Mode Control of Robot Manipulators*. Doctoral Dissertation.
11. Pettersson, Ola (2004) *Model-Free Execution Monitoring in Behavior-Based Mobile Robotics*. Doctoral Dissertation.
12. Överstam, Henrik (2004) *The Interdependence of Plastic Behaviour and Final Properties of Steel Wire, Analysed by the Finite Element Metod*. Doctoral Dissertation.
13. Jennergren, Lars (2004) *Flexible Assembly of Ready-to-eat Meals*. Licentiate Thesis.
14. Jun, Li (2004) *Towards Online Learning of Reactive Behaviors in Mobile Robotics*. Licentiate Thesis.
15. Lindquist, Malin (2004) *Electronic Tongue for Water Quality Assessment*. Licentiate Thesis.
16. Wasik, Zbigniew (2005) *A Behavior-Based Control System for Mobile Manipulation*. Doctoral Dissertation.

17. Berntsson, Tomas (2005) *Replacement of Lead Baths with Environment Friendly Alternative Heat Treatment Processes in Steel Wire Production*. Licentiate Thesis.

18. Tolt, Gustav (2005) *Fuzzy Similarity-based Image Processing*. Doctoral Dissertation.

19. Munkevik, Per (2005) *"Artificial sensory evaluation – appearance-based analysis of ready meals"*. Licentiate Thesis.

20. Buschka, Pär (2005) *An Investigation of Hybrid Maps for Mobile Robots*. Doctoral Dissertation.

21. Loutfi, Amy (2006) *Odour Recognition using Electronic Noses in Robotic and Intelligent Systems*. Doctoral Dissertation.

22. Gillström, Peter (2006) *Alternatives to Pickling; Preparation of Carbon and Low Alloyed Steel Wire Rod*. Doctoral Dissertation.

23. Li, Jun (2006) *Learning Reactive Behaviors with Constructive Neural Networks in Mobile Robotics*. Doctoral Dissertation.

24. Otterskog, Magnus (2006) *Propagation Environment Modeling Using Scattered Field Chamber.* Doctoral Dissertation.

25. Lindquist, Malin (2007) *Electronic Tongue for Water Quality Assessment*. Doctoral Dissertation.

26. Cielniak, Grzegorz (2007) *People Tracking by Mobile Robots using Thermal and Colour Vision*. Doctoral Dissertation.

27. Boustedt, Katarina (2007) *Flip Chip for High Frequency Applications – Materials Aspects*. Doctoral Dissertation.

28. Soron, Mikael (2007) *Robot System for Flexible 3D Friction Stir Welding*. Doctoral Dissertation.

29. Larsson, Sören (2008) *An industrial robot as carrier of a laser profile scanner: Motion control, data capturing and path planning.* Doctoral Dissertation.

30. Persson, Martin (2008) *Semantic Mapping Using Virtual Sensors and Fusion of Aerial Images with Sensor Data from a Ground Vehicle*. Doctoral Dissertation.

31. Andreasson, Henrik (2008) *Local Visual Feature based Localisation and Mapping by Mobile Robots*. Doctoral Dissertation.

32. Bouguerra, Abdelbaki (2008) *Robust Execution of Robot Task-Plans: A Knowledge-based Approach*. Doctoral Dissertation.

33. Lundh, Robert (2009) *Robots that Help Each Other: Self-Configuration of Distributed Robot Systems.* Doctoral Dissertation.

34. Skoglund, Alexander (2009) *Programming by Demonstration of Robot Manipulators.* Doctoral Dissertation.

35. Ranjbar, Parivash (2009) *Sensing the Environment: Development of Monitoring Aids for Persons with Profound Deafness or Deafblindness.* Doctoral Dissertation.

36. Magnusson, Martin (2009) *The Three-Dimensional Normal-Distributions Transform – an Efficient Representation for Registration, Surface Analysis, and Loop Detection.* Doctoral Dissertation.

37. Rahayem, Mohamed (2010) *Segmentation and fitting for Geometric Reverse Engineering. Processing data captured by a laser profile scanner mounted on an industrial robot.* Doctoral Dissertation.

38. Karlsson, Alexander (2010) *Evaluating Credal Set Theory as a Belief Framework in High-Level Information Fusion for Automated Decision-Making.* Doctoral Dissertation.

39. LeBlanc, Kevin (2010) *Cooperative Anchoring – Sharing Information About Objects in Multi-Robot Systems.* Doctoral Dissertation.

40. Johansson, Fredrik (2010) *Evaluating the Performance of TEWA Systems.* Doctoral Dissertation.

41. Trincavelli, Marco (2010) *Gas Discrimination for Mobile Robots.* Doctoral Dissertation.

42. Cirillo, Marcello (2010) *Planning in Inhabited Environments: Human-Aware Task Planning and Activity Recognition.* Doctoral Dissertation.

43. Nilsson, Maria (2010) *Capturing Semi-Automated Decision Making: The Methodology of CASADEMA.* Doctoral Dissertation.

44. Dahlbom, Anders (2011) *Petri nets for Situation Recognition.* Doctoral Dissertation.

45. Ahmed, Muhammad Rehan (2011) *Compliance Control of Robot Manipulator for Safe Physical Human Robot Interaction.* Doctoral Dissertation.

46. Riveiro, Maria (2011) *Visual Analytics for Maritime Anomaly Detection.* Doctoral Dissertation.

47. Rashid, Md. Jayedur (2011) *Extending a Networked Robot System to Include Humans, Tiny Devices, and Everyday Objects*. Doctoral Dissertation.

48. Zain-ul-Abdin (2011) *Programming of Coarse-Grained Reconfigurable Architectures*. Doctoral Dissertation.

49. Wang, Yan (2011) *A Domain-Specific Language for Protocol Stack Implementation in Embedded Systems*. Doctoral Dissertation.

50. Brax, Christoffer (2011) *Anomaly Detection in the Surveillance Domain*. Doctoral Dissertation.

51. Larsson, Johan (2011) *Unmanned Operation of Load-Haul-Dump Vehicles in Mining Environments*. Doctoral Dissertation.

52. Lidström, Kristoffer (2012) *Situation-Aware Vehicles: Supporting the Next Generation of Cooperative Traffic Systems*. Doctoral Dissertation.

53. Johansson, Daniel (2012) *Convergence in Mixed Reality-Virtuality Environments. Facilitating Natural User Behavior.* Doctoral Dissertation.

54. Stoyanov, Todor Dimitrov (2012) *Reliable Autonomous Navigation in Semi-Structured Environments using the Three-Dimensional Normal Distributions Transform (3D-NDT)*. Doctoral Dissertation.

55. Daoutis, Marios (2013) *Knowledge Based Perceptual Anchoring: Grounding percepts to concepts in cognitive robots*. Doctoral Dissertation.

56. Kristoffersson, Annica (2013) *Measuring the Quality of Interaction in Mobile Robotic Telepresence Systems using Presence, Spatial Formations and Sociometry.* Doctoral Dissertation.

57. Memedi, Mevludin (2014) *Mobile systems for monitoring Parkinson's disease*. Doctoral Dissertation.

58. König, Rikard (2014) *Enhancing Genetic Programming for Predictive Modeling*. Doctoral Dissertation.

59. Erlandsson, Tina (2014) *A Combat Survivability Model for Evaluating Air Mission Routes in Future Decision Support Systems*. Doctoral Dissertation.

60. Helldin, Tove (2014) *Transparency for Future Semi-Automated Systems. Effects of transparency on operator performance, workload and trust.* Doctoral Dissertation.

61. Krug, Robert (2014) *Optimization-based Robot Grasp Synthesis and Motion Control*. Doctoral Dissertation.

62. Reggente, Matteo (2014) *Statistical Gas Distribution Modelling for Mobile Robot Applications*. Doctoral Dissertation.

63. Längkvist, Martin (2014) *Modeling Time-Series with Deep Networks*. Doctoral Dissertation.

64. Hernández Bennetts, Víctor Manuel (2015) *Mobile Robots with In-Situ and Remote Sensors for Real World Gas Distribution Modelling*. Doctoral Dissertation.

65. Alirezaie, Marjan (2015) *Bridging the Semantic Gap between Sensor Data and Ontological Knowledge*. Doctoral Dissertation.

66. Pashami, Sepideh (2015) *Change Detection in Metal Oxide Gas Sensor Signals for Open Sampling Systems*. Doctoral Dissertation.

67. Lagriffoul, Fabien (2016) *Combining Task and Motion Planning*. Doctoral Dissertation.

68. Mosberger, Rafael (2016) *Vision-based Human Detection from Mobile Machinery in Industrial Environments*. Doctoral Dissertation.

69. Mansouri, Masoumeh (2016) *A Constraint-Based Approach for Hybrid Reasoning in Robotics*. Doctoral Dissertation.

70. Albitar, Houssam (2016) *Enabling a Robot for Underwater Surface Cleaning*. Doctoral Dissertation.

71. Mojtahedzadeh, Rasoul (2016) *Safe Robotic Manipulation to Extract Objects from Piles: From 3D Perception to Object Selection*. Doctoral Dissertation.

72. Köckemann, Uwe (2016) *Constraint-based Methods for Human-aware Planning*. Doctoral Dissertation.

73. Jansson, Anton (2016) *Only a Shadow. Industrial Computed Tomography Investigation, and Method Development, Concerning Complex Material Systems*. Licentiate Thesis.

74. Sebastian Hällgren (2017) *Some aspects on designing for metal Powder Bed Fusion*. Licentiate Thesis.

75. Junges, Robert (2017) *A Learning-driven Approach for Behavior Modeling in Agent-based Simulation*. Doctoral Dissertation.

76. Ricão Canelhas, Daniel (2017) *Truncated Signed Distance Fields Applied To Robotics*. Doctoral Dissertation.

77. Asadi, Sahar (2017) *Towards Dense Air Quality Monitoring: Time-Dependent Statistical Gas Distribution Modelling and Sensor Planning*. Doctoral Dissertation.

78. Banaee, Hadi (2018) *From Numerical Sensor Data to Semantic Representations: A Data-driven Approach for Generating Linguistic Descriptions*. Doctoral Dissertation.

79. Khaliq, Ali Abdul (2018) *From Ants to Service Robots: an Exploration in Stigmergy-Based Navigation Algorithms*. Doctoral Dissertation.

80. Kucner, Tomasz Piotr (2018) *Probabilistic Mapping of Spatial Motion Patterns for Mobile Robots*. Doctoral Dissertation.

81. Dandan, Kinan (2019) *Enabling Surface Cleaning Robot for Large Food Silo*. Doctoral Dissertation.

82. El Amine, Karim (2019) *Approaches to increased efficiency in cold drawing of steel wires*. Licentiate Thesis.

83. Persson, Andreas (2019) *Studies in Semantic Modeling of Real-World Objects using Perceptual Anchoring*. Doctoral Dissertation.

84. Jansson, Anton (2019) *More Than a Shadow. Computed Tomography Method Development and Applications Concerning Complex Material Systems*. Doctoral Dissertation.

85. Zekavat, Amir Reza (2019) *Application of X-ray Computed Tomography for Assessment of Additively Manufactured Products*. Doctoral Dissertation.

86. Mielle, Malcolm (2019) *Helping robots help us—Using prior information for localization, navigation, and human-robot interaction*. Doctoral Dissertation.

87. Grosinger, Jasmin (2019) *On Making Robots Proactive*. Doctoral Dissertation.

88. Arain, Muhammad Asif (2020) *Efficient Remote Gas Inspection with an Autonomous Mobile Robot*. Doctoral Dissertation.

89. Wiedemann, Thomas (2020) *Domain Knowledge Assisted Robotic Exploration and Source Localization*. Doctoral Dissertation.

90. Giaretta, Alberto (2021) *Securing the Internet of Things with Security-by-Contract*. Doctoral Dissertation.

91. Rudenko, Andrey (2021) *Context-aware Human Motion Prediction for Robots in Complex Dynamic Environments*. Doctoral Dissertation.