Foresighted Human-Aware Robot Navigation Based on Predicted Object-Interactions

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Lilli Ophelia Bruckschen
aus
Bonn

Bonn, Oktober 2020

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachterin:	Prof. Dr. Maren Bennewitz
2. Gutachter:	PD. Dr. Volker Steinhage
Tag der Promotion:	08.10.2020
Erscheinungsjahr:	2020

Abstract

Since the beginning of the digital revolution robots have become an increasingly important part of modern live. Until recently, this trend was primarily observable in industrial robots, however, with the success of the first mass produced household assistance robots like Roomba and Pepper, a shift towards consumer robots is on the horizon. This shift also leads to the need for new research regarding the behavior of robots in domestic spaces, as principles suitable for industrial robots may not be compatible with domestic robots. This thesis is concerned with research in this field and presents a novel navigation approach for foresighted human-aware robot navigation using predicted human-object interactions. The approach is based on the core idea that the last object interaction of a human carries knowledge about the next object interaction. Given enough training data this knowledge in combination with active observations can be a powerful tool to estimate likely navigation goals of a moving human user. Such predictions can then be used to guess the path the user will take to their next navigation goal and in turn adapt the path of the robot such that it will not disturb the user. Furthermore, in case of a service robot the prediction could also be used to guess at which destinations the user might need the assistance of the robot, thereby increasing the efficiency of it. The experimental evaluations in simulated and real world environments highlight that the presented approach can outperform state-of-the-art approaches both in the path efficiency of the robot as well as in the comfort of the human. The results thereby also show the usefulness of long-term planing and human-aware movement and that path efficiency and human comfort can go hand in hand.

Zusammenfassung

Seit dem Beginn der digitalen Revolution sind Roboter zu einem immer wichtigeren Bestandteil des modernen Lebens geworden. Bis vor Kurzem war dieser Trend vor allem bei Industrierobotern zu beobachten. Mit dem Erfolg der ersten massenproduzierten Haushaltsroboter wie Roomba und Pepper zeichnet sich jedoch eine Verlagerung hin zu Robotern für Verbraucher ab. Diese Verschiebung macht auch neue Forschungsarbeiten zum Verhalten von Robotern in Wohnräumen erforderlich, da Prinzipien für Industrieroboter nicht immer für Haushaltsroboter anwendbar sind. Diese Doktorarbeit befasst sich mit der Forschung auf diesem Gebiet und präsentiert einen neuartigen Navigationsansatz für eine vorausschauende, menschenbewusste Roboternavigation unter Verwendung vorhergesagter Mensch-Objekt-Interaktionen. Der Ansatz basiert auf der Idee, dass die letzte Objektinteraktion eines Menschen Informationen über die nächste Objektinteraktion enthält. Bei ausreichenden Trainingsdaten kann dieses Wissen, in Kombination mit aktiver Beobachtung, ein leistungsfähiges Werkzeug sein, um die wahrscheinlichen Navigationsziele eines sich bewegenden menschlichen Benutzers abzuschätzen. Solche Vorhersagen können dann verwendet werden, um den Weg vorherzusagen, den Nutzer zu ihrem nächsten Navigationsziel nehmen werden, sowie um den Pfad des Roboters so anzupassen, dass er Nutzer nicht stört. Darüber hinaus könnte im Fall eines Serviceroboters die Vorhersage auch verwendet werden, um vorherzusagen, an welchen Zielen Nutzer die Unterstützung des Roboters benötigen werden, wodurch die Effizienz des Roboters erhöht wird. Die Evaluation in simulierten und realen Umgebungen zeigen, dass der vorgestellte Ansatz sowohl hinsichtlich der Pfadeffizienz des Roboters als auch des Komforts des menschlichen Nutzers vergleichbare Ansätze aus der Literatur übertreffen kann. Die Ergebnisse zeigen ferner auch die Nützlichkeit von Langzeitplanung und menschenbewusster Bewegung, sowie, dass Pfadeffizienz und menschlicher Komfort Hand in Hand gehen können.

Acknowledgements

This thesis was written at the Humanoid Robots Lab of the University of Bonn under my thesis supervisor Prof. Dr. Maren Bennewitz. Without the continuous support of many, this thesis would not have been written. Among those I am particularly grateful to Dr. Jenny Mack, Kira Bungert, M.Ed. and Sabrina Amft, M.Sc., who helped me whenever I needed help and pushed and supported me throughout the last years. I also want to express my thanks for the support of my department. Firstly to Prof. Dr. Maren Bennewitz for giving me the opportunity to write this thesis and her advice during it. I further want to especially thank Sandra Höltervennhoff, B.Sc. and Nils Dengler, B.Sc. who worked with me on many projects during my thesis, as well as to Petra Zitzmann, B.A. for her continuous support regardless of how deadline-oriented my work was. I'm also grateful for the support of my co-authors, as well as my various proofreaders over the years. Last but not least, I want to express my thanks to the computer science student council and its members throughout the last years, for their help and support during my studies.

The work on which this thesis is based has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) BE 4420/2-1 (FOR 2535 Anticipating Human Behavior).

Contents

CHAPTER 1

Introduction

Today we stand at a point in history at which ubiquitous human-aware robots are slowly becoming more a reality than a dream. Assistance and personal robots in their various forms find their place in more and more households, from robot toys aimed at helping children understand basic principles of computer sciences [1], over household robots, e.g., iRobot's Roomba [2], to mass produced anthropomorphic assistance robots like SoftBank's Pepper [3]. The resulting close proximity to humans, however, creates new challenges, which often do not occur for relatively static industrial robots of the last century, or the earlier proto-robots seen through history. Indeed by looking at the development of robots from simple mechanisms to the programmable, autonomous systems known today, we can get a sense of how meaningful the current development may be and how high the demand for further research likely will be.

Robots and automatons, as their non programmable predecessors are often called today, have fascinated humans since at least the Late Bronze Age. One of their earliest forms is found in the New Kingdom period of ancient Egypt (ca. 1550 BC to ca. 1069 BC) [4, 5]. During this time animated statues, capable of moving their heads, played an important role in religions ceremonies and were believed to have a soul [6]. Other noticeable examples are offered by Greek mythology, from Pygmalion's ivory statue Galatea, given life by Aphrodite [7], to the creations of Hephaestus, which included mechanical handmaidens

built out of gold and capable of speaking [8], and Talos, a giant bronze man tasked with guarding the island of Crete. The Egyptian and Hellenistic worlds were not isolated in their conception of artificial beings. The same ideas can be found in Norse legend, e.g. in Mokkurkálfi, a giant made of clay with the heart of a mare [9], in Jewish folklore with the Golem, a creature made out of inanimate matter brought to life and controlled by its creator [10], or in ancient Chinese lore, e.g. when an artificial man, indistinguishable from a human, is presented to king Mu of Zhou in Liezi [11]. Aside from legends we also know of several early automatons through history. The first one was a steam-powered flying pigeon built by Archytas of Tarentum around 400 BC [12], not unlike the famous Digesting Duck invented by Jacques de Vaucanson more than 2300 years later [13]. However, it would take until the late 20th century for autonomous, pre-programmed robots to appear. The Unimate, the first industrial robot on a General Motors assembly line in 1961, is one of the first examples [14, 15]. From there on progress in and possibilities generated by robotics improved drastically, as the world shifted from analogous to digital technology.

By 1986 less than 1% of the world's stored information was in digital form. In 2014, just 28 years later, this number increased to over 99.5%. The capability to store data did similarly grow from $2.6 \cdot 10^{18}$ Bytes in 1986 to $4.6 \cdot 10^{21}$ Bytes in 2014 [16], $3.3 \cdot 10^{22}$ Bytes in 2018 and is expected to grow to $1.7 \cdot 10^{23}$ Bytes in 2025 [17]. Following this prediction the data storage capability of the world did increase by a factor of 100,000 in a "historic blink of an eye" [16] of 37 years. The improvements and new possibilities that came with these developments were so significant that this shift to digital technology is commonly called a *Digital Revolution* [18] or the *4th Industrial Revolution* [19, 20, 21]. One of the improvements that this revolution brought were intelligent, or as they are more commonly called *smart*, systems [22]. Today we can find some variant of them in numerous common objects, from phones [23], over home environments [24] to early forms of service robots [3]. These systems have risen alongside the Internet and are often incorporated in it. In fact they have outnumbered human users since 2009 [25], a trend that created the iconic phrase *Internet of Things* [26, 27]. At the time of writing there are approximately 5.8 billion systems connected to the Internet [28], while there are only 4.64 billion human users [29], resulting in a systems/user rate of 1.25. This rate is expected to grow to at least 1.8 by 2028, as the number of systems rises to 15 billion [28], outnumbering the expected world population at this point by 7.5 billion [30]. As not all intelligent systems are connected to

the Internet, their true number is likely even higher.

These past trends are important to make educated guesses about the near future. As shown above we are in the midst of an ongoing digital revolution and it can be reasonably expected that intelligent systems will become an even more essential and natural part of our daily lives. Furthermore, the commercially successful systems are also becoming more and more complex, from simple RFID supply-chain helpers in the early 2000s [31] to autonomous drones and small household robots in the mid 2010s [32, 33, 3]. As both demand as well as technical feasibility are already present for household robots, it is, judging by trends of the last decades, likely just a question of time until service robots will also become a common sight in our evermore digitalized world. However, robots are not without problems and may very well present an annoyance to humans, rather than providing support, if implemented without care. Often human-aware navigation policies as well as the ability to foresee future actions of humans are as important as the actual task-solving capabilities of a robot, as users are likely to feel some form of discomfort if the robot misses these features [34]. Research into these areas will therefore be likely very useful in the coming years.

The anthropomorphic assistance robot Pepper is a good example of the current state-of-the-art. It was introduced in June 2014 as a business-to-business robot model for SoftBank Robotics. The robot quickly became a success and was adapted to business-to-consumer, business-to-academics and business-to-developers markets [3]. From December 2015 to May 2018 SoftBank sold 12,000 units in Europe alone [35]. As of 2020, Pepper is used in over a thousand households in Japan [36] with an approximate price of $14,000 per unit [37]. At the time of writing the service robot market as a whole is estimated to be worth 37 billion USD and to grow to 102.5 billion USD by 2025, a compound annual growth rate of 22.6% [38]. Regarding its software, Pepper comes equipped with a basic navigation and local obstacle-avoidance behavior. It is further possible to increase its performance by utilizing additional ROS-based modules, e.g., for navigation-oriented perception and planning [3]. This marks an interesting point about the current status of robots; while Pepper is one of the most successful commercial household robots today it only utilizes a basic navigation model and has no means of human-aware behavior on its own, aside from collision avoidance. However, as observed in the field of social robotics, basic collision avoidance is not enough if a robot is omnipresent around humans. The

robot also needs to keep human comfort (absence of annoyance and stress for humans), naturalness (similarity between robot and human behavior), and sociability (adherence to cultural conventions) in mind to avoid disturbing its users. These constraints can however only be fulfilled if the robot posesses a reliable prediction about human movements and can therefore adjust its behavior in advance [34]. The literature also shows that if this can be done, not only does the comfort of the user increase but also the performance of the robot and the trust of the user in them [39, 40].

This work is part of the above mentioned social robotic research field and presents one possible approach to achieve foresighted human-aware robot navigation using predicted human-object interactions. The approach is based on the core idea that the last object interaction of a human carries knowledge about the possible next object interaction. For example if a user interacts with an empty cup it is intuitively more likely that they will next interact with a coffee machine, dishwasher or table than with e.g. a wardrobe. Given enough training data this knowledge in combination with active observations can be a powerful tool to estimate likely navigation goals of a moving human user. Such predictions can then be used to guess the path the user will take to their next navigation goal and in turn adapt the path of the robot such that it will not disturb the user. Furthermore, in case of a service robot the prediction could also be used to guess at which destinations the user might need the assistance of the robot, thereby increasing the efficiency of the robot. A motivational example of this approach is depicted in Fig. 1.1. In this example a user interacts with a microwave. Based on this observation and learned interaction sequences the robot predicts that they will next interact with the nearby table. In turn the robot avoids the path between the user and the table and begins to set the table.

Structure-wise the approach is divided into three separated parts. The first part is the underlying human-object interaction detector (Chapter 2), the second part the navigation goal prediction framework (Chapter 3), and the final part the resulting navigation policy (Chapter 4). Each of these parts, alongside an individual experimental evaluation, is discussed in detail in a separate chapter of this thesis. In the last chapter (Chapter 5) the whole approach is summarized and possible future work is presented.

The remainder of this chapter discusses various challenges each of these parts needs to overcome and states the three key contributions of this work. It also lists all relevant publications on which this thesis is based.

Figure 1.1: Motivational example of the presented approach. The robot observed that the user interacted with a plate and a microwave. It predicts based on observations in the past that the next object the user will interact with is the table. It also predicts that the user is likely to next interact with the table to eat on it and begins to set it, all the while avoiding positions between the user and the table to minimize the chance of interfering with them.

1.1 Challenges and Contributions

Each part of the approach presented in this thesis faces unique challenges and each of these challenges must be overcome in order to guarantee correct behavior for the parts that build on it.

1.1.1 Human-Object Detection

The first part of this work is concerned with the creation of a human-object detection system. The most obvious first challenge is to find a suitable definition of what exactly a human-object interaction is in order to measurably identify it.
Once a definition is found the detector needs to be able to identify both objects and humans.

Alongside their position it is also important to identify the type of the objects, to return with which object a human has interacted, and the pose of the human, to determine if they just pass by or interact with the object. The approach should also be usable on a mobile robot and therefore should work on an ongoing video stream. Lastly we need to counter possible false positive detections, which can be done by inferring if an interaction continues over multiple frames. To ensure that all these points are sufficiently implemented an experimental comparison with similar literature approaches is also necessary. This results in the following contributions:

- A human-object interaction detector, capable of working on video streams and dealing with possible false positive detections.
- An experimental evaluation of the interaction detector's recall and precision values.

These contributions can be further summarized to one key contribution: An approach to robustly detect and extract human-object interaction in video streams based on spatio-temporal and pose information over subsequent frames.

1.1.2 Movement Prediction

Given a working human-object detector the task of the second part of this thesis is to use it alongside a long-term prediction framework to estimate the navigation goal of a moving human user.
The first challenge that presents itself is to utilize the detector to predict possible future interactions. This can be done by gathering training data over subsequent human-object interactions, in other words, building a probability distribution which encodes the likelihood that a user will interact next with an object of type B after they interacted with an object of type A. However, this *Interaction Model* on its own only yields a very rough prediction by providing weights for possible goal locations. Furthermore, there may be a multitude of objects inside the environment, which might result in a very fragmented prediction, especially if these objects are close to each other, e.g., a laptop on a table. Therefore, a way to simplify the environment by merging close objects into *Activity Regions* must be found. Once an Interaction Model is found and the environment is simplified a prediction framework needs to be found which takes both the Interaction Model as well as the observed user

movement into account. Bayesian inference with the Interaction Model as prior knowledge and current user movement as observations is the approach presented in this thesis.
Again the whole approach also needs to be evaluated against similar literature approaches to sufficiently ensure it is able to fulfill its task to predict the next navigation goal of a moving human user.
This leaves us with three contributions for this part:

- An approach to identify activity regions in indoor environments.
- A Bayesian inference framework based on transition probabilities between objects to predict the navigation goal of a moving human.
- An experimental evaluation of the prediction accuracy of the framework including a comparison to existing approaches.

These can be further summarized into one key contribution: A Bayesian inference framework based on transition probabilities between objects to predict the navigation goal of a moving human.

1.1.3 Human-Aware Navigation

The final part of the presented approach deals with the creation of a human-aware navigation policy, based on the previously discussed prediction system, which minimizes discomfort for the user and maximizes the path efficiency of the robot.
The first obvious challenge is again how to define what contributes to human discomfort and how a robot must behave to minimize it. To answer these questions user studies are needed, as the most simple way to identify user comfort or discomfort with a specific robot behavior is to ask them. Also, path efficiency needs to be defined. In our case of a service robot this can be seen at its ability to avoid unnecessary movements while still being able to provide assistance to the user when necessary. Both efficiency and human comfort can be achieved by using a prediction framework, and comfort by first establishing basic constraints for the movement of the robot and then simulating the likely movement of the human. Once this is known the robot can adapt its movement to abide by these comfort constraints for the predicted future. Similarly, the robot might predict the next object interactions of the human and ignore locations at which its help is not needed.

This, however, means that we need a way to see further into the future than just the next interaction, prior knowledge about at which objects the user might need help, and time dynamic path planing to ensure that the comfort constraints are always fulfilled. Also, as with the previous parts, an experimental evaluation is necessary, this time not only with similar approaches from the literature but also with direct human feedback to measure the comfort or discomfort of the robot behavior.
All in all this part presents the following contributions:

- An activity region based prediction system to infer the location where the user will need help next based on previous object interactions and robot observations.
- A human-aware navigation system based on long-term movement prediction, human comfort constraints, and path planning on a time-dependent cost-map.
- The conduct of a survey about human comfort to evaluate different navigation strategies.
- An evaluation of the complete system in comparison to state-of-the-art methods, in simulated environments with pre-defined metrics, and in real-world and virtual reality experiments with direct human feedback.

Again, these contributions can be summarized into one key contribution: A time-dependent navigation policy to compute a human-aware path for a service robot to the next destination at which the user will need assistance.

1.1.4 Key Contributions

In summary this work is concerned with the following three key contributions:

- An approach to robustly detect and extract human-object interactions in video streams based on spatio-temporal and pose information over subsequent frames (Chapter 2).
- A Bayesian inference framework based on transition probabilities between objects to predict the navigation goal of a moving human (Chapter 3).
- A time-dependent navigation policy to compute a human-aware path for a service robot to the next destination at which the user will need assistance (Chapter 4).

1.2 Publications

Parts of this thesis have already been published, the lead author is always listed first. In the following these publications are listed, grouped by the chapters in which they are used.

- **Chapter 2: Human-Object Interactions**
  - L. Bruckschen, S. Amft, J. Tanke, J. Gall, M. Bennewitz. Detection of Generic Human-Object Interactions in Video Streams. *In: Proceedings of the International Conference on Social Robotics (ICSR), 2019.*
- **Chapter 3: Interaction Based Movement Prediction**
  - L. Bruckschen, N. Dengler, M. Bennewitz. Human Motion Prediction Based on Object Interactions. *In: Proceedings of the European Conference on Mobile Robots (ECMR), 2019.*
  - L. Bruckschen, K. Bungert, N. Dengler, M. Bennewitz. Predicting Human Navigation Goals based on Bayesian Inference and Activity Regions. *In: Robotics and Autonomous Systems (RAS), 2020, accepted.*
- **Chapter 4: Human-Aware Robot Navigation by Long-Term Movement Prediction**
  - L. Bruckschen, K. Bungert, N. Dengler, M. Bennewitz. Human-Aware Robot Navigation by Long-Term Movement Prediction. *In: Proceedings of the International Conference on Intelligent Robots and Systems (IROS), 2020, accepted.*
  - L. Bruckschen, K. Bungert, M. Wolter, S. Krumpen, M. Weinmann, R. Klein, M. Bennewitz. Where Can I Help? Human-Aware Placement of Service Robots. *In: Proceedings of the Conference on Robot and Human Interactive Communication (RO-MAN), 2020, accepted.*

CHAPTER 2

Detecting Human-Object Interactions

This thesis is based on the ideas that by knowing past and future object interactions of users we can create a foresighted and human-aware navigation policy for a mobile service robot. However, before we can design a prediction or navigation approach we need to find a way to robustly detect human-object interactions. The following chapter is concerned with this and discusses the first key contribution of this Thesis: An approach to robustly detect and extract human-object interaction in video streams based on spatio-temporal and pose information over subsequent frames.
As stated in Chapter 1, the content of this chapter has already been published [41].

2.1 Introduction

The ability to detect interactions of humans with objects is of great use for a variety of applications, especially for service robots. Examples include the identification of customer browsing patterns in retail scenarios [42], activity recognition based on used objects alongside monitoring of daily activities [43, 44], and, as shown in the following chapters, the prediction of human movements based on subsequent object interactions. Most work regarding interaction detection focuses on well-constrained scenarios often with the goal to identify a small set of potential activities in prerecorded videos [45, 46]. In contrast to

(a) (b)

Figure 2.1: The presented detector identifies human-object interactions based on object positions (purple) as well as pose and orientation information of the human (green) (a). To deal with uncertainty in the observations, it computes for each found human-object interaction the likelihood that this interaction really occurs using previous observations (b). In this example, the human interacts with the coffee machine over several frames, resulting in a high likelihood for this interaction.

that the approach presented in this chapter does not assume specific activities but allows for the detection of arbitrary human-object interactions with 510 different entities from the *Open Image* dataset[1] [47]. While it focuses on video streams, it can also be applied to static images or pre-recorded videos. For the context of this thesis, a human-object interaction is defined as an action in which a human places at least one hand on an object while facing it, see Fig. 2.1 for a demonstration. The detector presented in this chapter identifies relevant objects inside each frame using an RGB-D Camera in combination with regional convolutional neural networks (R-CNNs) [48] and estimates humans and their body pose using the *OpenPose* system [49]. It then detects possible interactions based on the pose of the human alongside their distance to objects. To deal with uncertainty in the observations, the system computes for each found interaction the likelihood that it really occurs by tracking it over subsequent frames. The output of the framework is the set of all detected human-object interactions with a sufficiently high likelihood. Fig. 2.1 illustrates the methodology of this approach.

As the approach is constrained by the types of recognizable objects it is able to utilize a

[1] No differentiation was made between objects and animals inside the OpenImage dataset

vast amount of available training data [47]. This allows it to recognize any interaction with objects known by an interchangeable R-CNN.
As shown by an experimental evaluation in Sec. 2.4 the detector achieves recall and precision rates of 0.82 with respect to the detection of human-object-interactions. Direct comparison with the literature is difficult as the focus of most common approaches is activity recognition and not human-object interaction detection. One similar approach in the literature [45] lists recall values of 0.90 and precision values of 0.62. However, this system was only able to detect interactions with 2 types of objects, while the presented approach is able to detect interactions with 510 different objects, depending on the used CNN.

2.2 Related Work

The detection of interactions between humans and objects is closely intertwined with activity recognition, as the type of the used object is typically associated with an activity. An interesting work in this context is presented by Prest *et al.* [45]. The goal of the authors is to detect smoking and drinking activities in realistic videos. To accomplish this, Prest *et al.* train an action classifier on example interactions and use this classifier in combination with a generic, part-based human detector [50] to spot the previously learned interactions in a prerecorded video. The system tracks objects and persons in space and time and uses the action classifier on the tracked data. In contrast to the presented detector the application domain of this system is limited, as it is only able to detect interactions with cigarettes and glasses. The precision of the system is also a problem, the paper reported precision values of 0.62 for glass interactions (drinking) and 0.32 for cigarette interactions (smoking). In contrast to the presented approach this system is very limited in its detection capabilities and shows weaknesses regarding the precision of the detection. Similarly, Yang *et al.* [44] propose to use object and interaction information to assign a predefined role, in their example *kidnapper* and *hostage*, to a human. To detect human-object interactions, the authors apply depth information and R-CNNs [48] and assume that an object is in use when it is very close to a human in terms of position and depth. Object interactions are not explicitly returned but rather used as weights during the role classification process. As interaction detection is primarily done using position information obtained from an

R-CNN, detection errors can easily lead to wrong results. While the presented framework also uses an R-CNN, it additionally makes use of pose and spatio-temporal information to increase the robustness of the detection.

Several other related systems use static images rather than videos, for example the work by Yao *et al.* [43]. The authors use the assumption that objects are associated with activities with the goal to increase object detection rates in static scenes by utilizing information about pose and activities of humans. The work of Gupta *et al.* [51] follows a similar idea. The authors propose a Bayesian model that incorporates functional and spatial context for object and action recognition. Another approach that focuses on action detection in static images was presented by Gkioxari *et al.* [52]. The authors detect humans and objects with an R-CNN and estimate action-type specific densities to localize the used object. In most cases, this corresponds to the position of a hand of the human.

While these approaches focus on static images, they highlight that pose information, especially about the hands of the human, and spatial context are important for efficient interaction detection in individual frames.

A quite different approach is given by Li et al. [42]. The authors use UHF RFID tags to mark objects and track interactions. The intended applications of the system are interactive storytelling with toys, inference of daily activities in indoor environments and identification of customer browsing patterns in retail scenarios. While quite different from the presented detector this work highlights the usefulness and applications of a human-object interaction detector.

Often, video description applications also use human-object interaction detectors. Most video description approaches focus on describing short scenarios, such as cooking, with a few sentences [46, 53], or on the detection of predefined actions from short, constrained scenes [54, 55]. Usually, video description systems use some form of recurrent neural network [56] to directly map visual data to natural language. In contrast to that, the presented system applies a step to verify possible interactions based on subsequent observations.

Alternatively, for small-scale action detection trajectories of various body parts are often tracked and used as a basis for action recognition [57].

Yet another approach focusing on activity recognition is presented by Koppula *et al.* [58]. The authors propose a system to extract a descriptive labeling of a sequence of human sub-activities in RGB-D videos. To obtain this labeling, the authors use a Markov random

field, where nodes model objects as well as sub-activities and edges represent relationships between object affordances. To find the correct temporal segmentation for a single human-object interaction, the authors apply a structural support vector machine. This work focuses primarily on sub-activities and object affordances of a single main activity. As extensive background knowledge about object affordances is needed for each type of interaction the system does not scale well.
The presented detector uses pose information [43, 51], especially about the hands of the human [52], alongside R-CNNs [48] to detect possible interactions in individual frames. It then applies a verification step in the video stream to deal with false positive detections. It extends the state of the art by allowing arbitrary interactions with known objects, thereby shifting the focus from action recognition to the detection of human-object interactions, allowing the use of a large amount of freely available training data [47]. Several applications can utilize the returned information ranging from movement prediction, as shown in this thesis, to intention or activity recognition at a larger scale, as shown by the literature examples.

2.3 Detection of Human-Object Interactions

As mentioned in Sec. 2.1 our goal is to detect all human-object interactions that occur in a video stream. A human-object interaction is thereby defined as an action in which a human places at least one hand on the object while facing it, as demonstrated in Fig. 2.1.
Further let a video stream be defined as a sequence of frames $V = [f_0, ..., f_t]$ with f_0 as the first observed frame and f_t as the currently observed frame at time t. The detector uses the current frame f_t and all previously found interactions on $[f_0, ..., f_{t-1}]$ as input and returns all human-object interactions in V.
In summary, the presented system to find all human-object interactions inside V works as follows:

1. Apply an R-CNN to detect objects and the *OpenPose* system [49] to detect humans and their poses from RGB data.

2. Use position and depth data to find overlaps between object bounding boxes and human hand positions. Use pose information of the human to check whether they

are facing an object that overlaps with their hand, if so, record a possible interaction.

3. Update the likelihood of interactions based on the new observations. This step is necessary to verify that a detected interaction really occurs.

The results are all human-object interactions with a likelihood over a threshold min_L, which is determined using training data. A learning process for min_L is shown in the evaluation. An example video demonstrating the approach is available on the website of the Humanoid Robots Lab of the University of Bonn[2].
In general, the approach can deal with several persons in a frame by using unique IDs for them and their respective interactions. However, for simplicity only scenarios with one visible human are considered in the following.

2.3.1 Detection of Objects and Humans and Estimation of the Human Pose

To efficiently detect objects in the current frame the detector uses an R-CNN from Google's object detection API [59], which was trained on the Open Images dataset [47]. Note that the R-CNN is interchangeable and its object detection capabilities can be extended using transfer learning techniques [60] in case new objects need to be detected. For the detection of humans and their poses the detector applies the *OpenPose* framework [49]. The estimated pose directly contains information about the position of ears, eyes, nose, shoulders, hands, and legs of the human. Additionally, the detector analyzes RGB-D data and infers the torso orientation θ by combining estimations about the joint positions from the 2D image and depth data. Similarly to Biswas *et al.* [61], it obtains a pixelwise joint position probability map by applying an implementation of OpenPose [62] and uses thresholds to get the approximate joint positions. It then estimates the torso normal by analyzing the shoulder and hip key points. For this purpose, it computes the cross product for each edge of the rectangle formed by the shoulder and hip joints. The estimated torso normal is then the mean of these four cross products and θ can be directly derived from this torso normal. The system fails however when hips or shoulders are occluded.

[2] https://www.hrl.uni-bonn.de/icsr_interaction_demo.mp4

Figure 2.2: Example of the requirements to detect human-object interaction. The orientation of the human and the center of the hands is inferred based on the human pose (green). Objects (purple) that the human faces and touches (i.e., there is an overlap between the object bounding box and part of the hand positions) are marked as possible interactions. In this frame, the system will detect a possible interaction with a microwave.

2.3.2 Detection of Possible Human-Object Interactions

Depending on the results of the orientation estimation we can infer which objects the human is facing based on their x coordinates in the frame and depth value with respect to the human. In particular, the presented approach processes each detected object in the current frame and checks whether the following conditions are satisfied for the current frame f_t:

- Some of the 2D positions of a human hand are inside the 2D bounding box of the object.
- The human is facing the object, i.e. is oriented towards it.

– The depth values of the hand and the object are approximatively similar.

If an object fulfills all these conditions, a possible interaction of the human with this object is recorded for f_t, see Fig. 2.2 for an example.

2.3.3 Dealing With False Positive and False Negative Detections of Human-Object Interactions

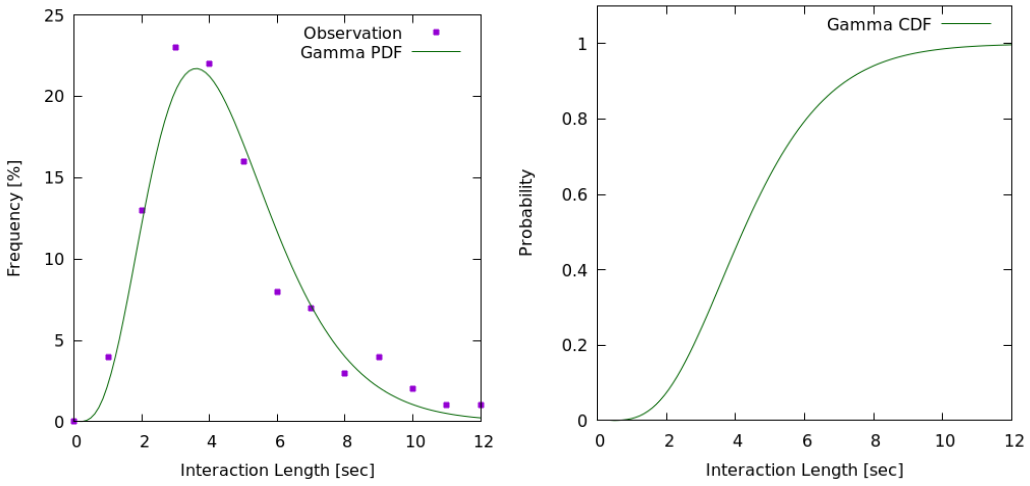
A common problem of human-object interaction systems are false positive object detections [45], e.g., when image regions are wrongly classified as objects. Furthermore, drops in the recall rates due to occlusions while the human interacted with an object, e.g., while drinking from a cup are also common. To deal with such effects, we need to explicitly consider uncertain observations and compute the likelihood of possible human-object interactions to estimate the probability that the interaction really occurs based on their detection in subsequent frames. This likelihood is first inferred based on previously found interactions with the same object, with higher values for interactions found in multiple frames. It is also updated for already observed frames based on new data, for example if an object was shortly occluded and therefore not detected for a frame inside a longer interaction.

The likelihood function was designed by evaluating the typical minimal length of human interactions with an object on a training data set collected in a university setting. Most interactions were shorter than 12 seconds. Longer interactions often last for several minutes, e.g., working with a laptop. This results in a distribution with a high amount of data points during the first 12 seconds and very scattered data points for longer durations.

$$G(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} \exp(-\frac{x}{\theta}) \tag{2.1}$$

with k=5 and θ=0.9 and $\Gamma(a)$ as the *gamma function* [63], is a close approximation to the data. The resulting distribution for the first 12 seconds is visualized in Fig. 2.3(a). As can be seen its fits the data points closely.

Given this distribution the cumulative distribution function $G_C(x)$ of $G(x)$ models the probability that an interaction has a duration of x or less seconds



(a) (b)

Figure 2.3: (a) Gamma probability density function [64] (green) approximating the observed interaction durations (purple) in the real-world data set. (b) Cumulative form of the gamma probability density function that indicates the probability that an interaction lasts for at most x seconds. Both functions were modeled with k=5 and θ=0.9.

$$G_C(x) = \frac{1}{\Gamma(k)}\gamma(k, \frac{x}{\theta}) \tag{2.2}$$

with k=5 and θ=0.9 and $\gamma(a, b)$ as the *incomplete gamma function* [65], see Fig. 2.3(b) for a visualization of this cumulative distribution function.

$G_C(x)$ then determines the likelihoods of the detected possible human-object interactions (see Sec. 2.3.2) using their estimated duration. In more detail, for an interaction with a given object we say that a frame is a *hit* if an interaction with this object was detected and a *miss* otherwise. To get the length of an interaction over several frames we can mark a *hit* as the start point of a new interaction if the elapsed time since the last *hit* for this interaction is greater than a threshold $t_{max} = 5\,s$. This value can be determined from $G_C(x)$ as 50% of the interactions are within a duration of 5 seconds. Not detecting a single *hit* for a specific human-object interaction during this time is a strong indication that no interaction with the object took place. In the beginning, each likelihood is initialized with 0 for all objects in each frame. We can then compute the likelihoods of all detected

interactions and possibly update the likelihoods of object interactions on previous frames where the interaction was not detected. Alg. 1 lists the complete approach to compute the likelihood of an observed human-object interaction, with $timeDiff(a, b)$ as the time difference between a and b.

Algorithm 1 : Likelihood computation.

Input : Possible human-object interaction I on frame f_t,
time of previous *hit* for I t_{phit}, start time of I t_{start}.
Output : Likelihood L that I really occurred.

$t_{diff} = timeDiff(t, t_{phit})$
if $t_{diff} > t_{max}$ **then**
    //new interaction with this object
    $t_{start} = t$
end
$L = G_C(timeDiff(t, t_{start}))$
if $(t-1) < t_{diff} < t_{max}$ **then**
    //false negative detections for I occurred
    set likelihood of I in frames $f_{t_{phit}+1}, \cdots, f_{t-1}$ to L
end
$t_{phit} = t$
return L

Fig. 2.4 shows an example of the likelihood computation. As can be seen, the presented approach is both able to deal with false detections as well as to merge longer interactions with detection errors in between frames.

2.4 Experimental Evaluation

The approach was tested in an experimental evaluation with respect to precision and recall and the improvement achieved by its likelihood function. The test dataset contained 195 human-object interactions of 10 different people with objects from the *Open Image* dataset [47] over 27 minutes of video data. All videos were recorded with 12 frames per second in

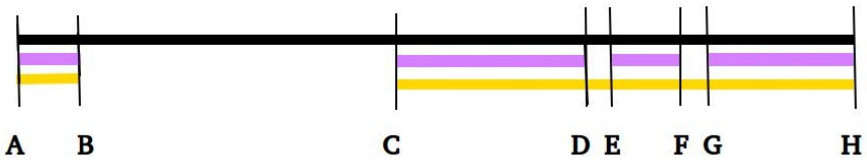


Figure 2.4: Illustration of the likelihood computation for one interaction. The black line represents the video stream, the violet frames report a *hit* for the interaction and the yellow frames illustrate the returned interaction with a certain likelihood. The duration between frame B and C is longer than t_{max}, therefore, a different interaction is happening between A and B than between C and H. Between D and E and between F and G there is no *hit*, however, as the duration until the next *hit* is less than t_{max} a detection error is likely.

indoor environments [3]. Fig. 2.5 shows 6 example interactions from the dataset in different environments.
The ground truth for each frame, i.e., the information which human-object interactions are taking place were created manually.

2.4.1 Precision and Recall

The output of the approach on each frame was automatically compared with the ground truth to compute the recall and precision rates. Hereby the evaluation was performed with respect to the likelihood value min_L from which on the framework was certain enough to return a found interaction with an object. Fig. 2.6 shows the evolution of the precision and recall for 100 different values of min_L equally distributed in the range from 0 to 1. The results were fitted using least squares.
Using a min_L value of 0.21 the presented framework is able to achieve recall and precision rates of 0.82. Accordingly, in practice this serves as a threshold for the likelihood as both the recall and precision are relatively high. As can be seen in Fig. 2.3(b) this corresponds to a minimal interaction length of approximately 2 seconds.
Increasing min_L to a value close to 1 results in higher precision values up to 0.88 and lower recall values of 0.75. Decreasing min_L on the other hand has the opposite effect resulting in a precision rate of 0.62 and recall rate of 0.82 for a value of min_L close to zero. This evaluation highlights the usefulness of the verification step using the likelihood

[3] Videos from this dataset are availably under https://www.hrl.uni-bonn.de/icsr2019

Figure 2.5: Six example interactions from the evaluation dataset in different environments.

computation as it can improve the returned precision value while only slightly reducing the recall value. In general, values for min_L between 0.2 and 0.6 seem to be a good compromise between high precision and recall rates.

False negative detections naturally happen at the start of an interaction since the corresponding likelihood is initially low as the duration of the interaction is very short at this point in time. False positive detections typically happen if objects are very close together. In this case it is very difficult to differentiate between the interacted and the passive object.

Comparison with the literature is difficult as the focus of most approaches are on activity recognition and not human-object interaction detection. One of the most similar approaches in the literature [45] lists recall values of 0.90 and precision values of 0.62. It should also be noted that this system was only able to detect interactions with 2 types of objects, while the presented detector is able to detect interactions with 510 different objects.

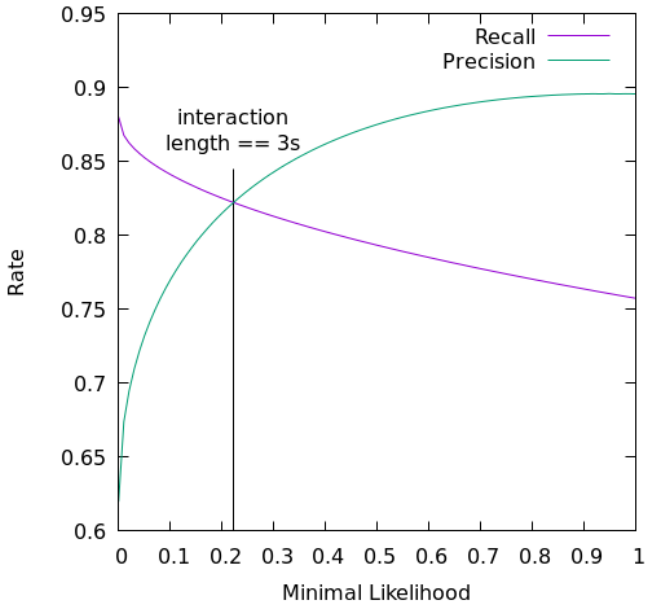


Figure 2.6: Evolution of precision and recall rates of the presented approach with respect to different values of min_L. The precision rate strongly increases with higher min_L values while the recall rate only slowly decreases. This illustrates the usefulness of the verification step based on subsequent observations, as a higher amount of false positive classifications than true positives is removed.

2.5 Discussion

As demonstrated, the framework is able to robustly detect human-object interactions in video streams. By applying the system to video streams recorded with the camera of a robot it can further be used to predict human movements. To do so, it first needs to learn a distribution from collected data to represent the probability that after an interaction with an object A the human will next interact with an object B. The training data needed to infer such a distribution could also be gathered with the detector if videos of typical human movements are available. The robot can then use this knowledge to predict future movement goals of the human based on known locations of objects in the environment. This has the added benefit that the robot must not learn map specific trajectories, as often seen in the literature [66], only the position of the object matters. Fig. 2.7 shows an application example of this. In this scenario, the robot detects a human-object interaction

(a) (b)

Figure 2.7: Example application of the detector to predict human movement goals. The robot (blue) detects an interaction of a human (violet) with a cup using the detection framework (a). Based on a pre-learned probability distribution about interaction transitions, the likelihood of possible next interaction objects is computed (b). The darker the green the higher the likelihood. Object names are abbreviated: table (T), sofa (S), refrigerator (R), coffee machine (C).

with a cup and computes based on an interaction model transition probabilities to other known objects. The most likely next objects in this example are sofas, tables, refrigerators, and coffee machines.

During the next chapters we will discuss in detail how a prediction (Chapter 3) and a human-aware navigation (Chapter 4) approach can be designed based on the detector presented in this chapter.

2.6 Conclusion

In this chapter, we saw an approach to automatically extract human-object interactions from video streams. In comparison to existing frameworks, this system focuses on the detection of general interactions with objects rather than specific activities. Furthermore, it uses spatio-temporal information to verify found interactions, an R-CNN to detect objects and the *OpenPose* pose estimator [49] to detect humans and their poses. Based on this information, it finds human-object interactions on the current frame and computes for each interaction the likelihood that it is really happening based on subsequent observations.

As the experimental evaluation demonstrates, this approach is able to robustly detect

human-object interactions with recall and precision rates of 0.82 on the used test dataset. This detector is the first step towards our overall goal of creating a foresighted human-aware robot navigation framework based on predicted object interaction. The next step is to use it to construct a long-term prediction framework based on transition probabilities between objects to infer the next navigation goal of a moving human user. The following chapter presents the second key contribution of this work: a Bayesian inference framework based on transition probabilities between objects to predict the navigation goal of a moving human.

CHAPTER 3

Interaction Based Movement Prediction

With the detector presented in Chapter 2 we can reliably detect human-object interactions in RGB-D video streams. In this chapter we use this knowledge to construct a long-term prediction framework based on transition probabilities between objects to infer the next navigation goal of a moving human user. This represents the second key contribution of this work: a Bayesian inference framework based on transition probabilities between objects to predict the navigation goal of a moving human. In the overall human-aware navigation approach this framework will serve as the foundation for the movement prediction of the human.

As stated in Chapter 1, the content of this chapter has already been published [67, 68].

3.1 Introduction

As it becomes more common for robots to operate in close proximity to humans, it is often necessary to anticipate their behavior, e.g., to avoid interferences with their daily habits [34] or predict where assistance may be needed. Previous approaches have tackled this problem by learning typical human trajectories in known environments [66, 69] or reacting dynamically to humans in close proximity [70]. However, in many cases a lot can be learned about human movements by looking at the last objects or group of objects they have

interacted with. For example, if we know that a human has interacted with objects inside a kitchen it is a likely possibility that they will next move towards a dining area. These object based areas or activity regions are often naturally defined by close-by objects, e.g., a work related area with a laptop and a cup on top of a table. These two observations are used in this chapter to present a Bayesian inference approach that predicts the navigation goal of a moving human by combining prior knowledge about such activity regions with online observations of the human's pose, resulting in the following contributions:

- An approach to identify activity regions in indoor environments.
- A Bayesian inference framework based on transition probabilities between objects to predict the navigation goal of a moving human.
- An experimental evaluation of the prediction accuracy of the framework including a comparison to existing approaches.

An online survey was used to test how humans perceive activity regions and infer which objects would be grouped together and at what distances humans would stop perceiving close objects as associated groups. Based on the results of this survey a rule-based classifier was created that identifies activity regions on a semantic map of the environment based on the proximity of objects to each other. In general, objects are grouped into the same activity regions if they are closer than two meters to each other. Following the approach discussed in Chapter 2, we can learn transition probabilities between different activity regions given the objects present in these regions. This prior knowledge can then be combined with online observations of the human's pose in a Bayesian inference framework to predict which activity regions are likely to be the next navigation goal.
This approach makes again use of an RGB-D camera system and estimates the user's pose and human-object interactions in the same way as discussed in Chapter 2. It further uses a semantic map of the environment, with marked object positions and activity regions as well as knowledge about typical human-object interaction sequences as prior knowledge. Using this it automatically infers the next activity region in which the human will interact with an object. As long as the same objects are present and the maps are known to the system it can be used in multiple environments.

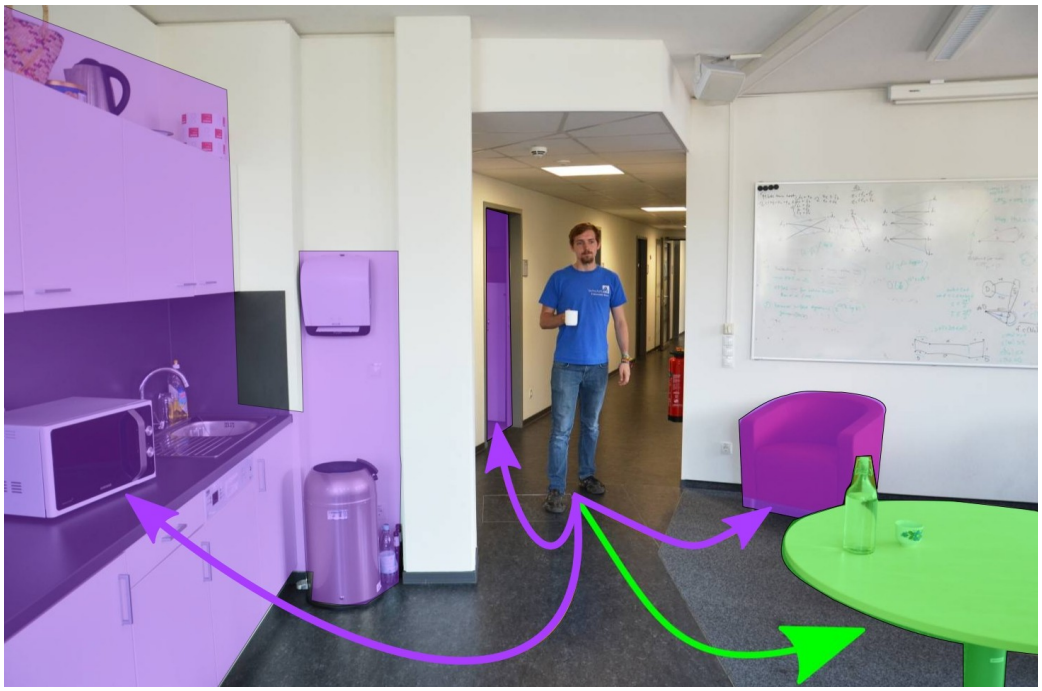


Figure 3.1: The aim of the framework is to infer the navigation goal of a moving human. This figure was segmented by hand and shows a sketch about the underlying idea of the framework. The user arrived from an office area were they interacted with a cup which they now carry. This was detected with an RGB-D camera using the approach from [41]. Four possible activity regions are likely navigation goals based on observations of the movement of the user and prior knowledge about the possible next object interaction. The green activity region, consisting of the table, bottle and cup, is the most likely one, while the violet regions are also possible with lower probabilities.

As shown in the evaluation in Sec. 3.5, the approach achieves a higher prediction accuracy than a trajectory based reinforcment learning method [71] while simplifying both the representation as well as the needed training data.

Fig. 4.1 shows a motivating example of the approach in which a human interacts with a cup. As can be seen by the different colors, the system identifies four different activity regions and determines the most likely one as the region consisting of the table and the water bottle.

The approach was designed to be used by a mobile robot, equipped with an RGB-D camera and the means to self locate. However, in principle the system can be used on any RGB-D

camera system, even without a robot.

3.2 Related Work

As noted by Kruse *et al.* [34] the ability to predict human movements is vital for any robot that operates in the same environment as humans. Therefore a lot of research has been put into prediction frameworks that are used for the prediction of human navigation goals [66, 69] or navigation through dense crowds [72, 73].
An overview and taxonomy about recent prediction frameworks is given by Rudenko *et al.* [74]. The authors categorized approaches in regards to their modeling of the future movement in three categories: Physics-based methods (sense-predict), which predict movements by simulating the next steps of the human using dynamical models based on Newton's law of motion and observations about the current state of the user. Pattern-based methods (sense-learn-predict), which are based on motion patterns from prior observed user trajectories. Planning-based methods (sense-reason-predict), which reason about the long-term navigation goals of the user and predict path hypotheses based on this. Using this taxonomy the approach presented in this chapter falls into the planning-based category, as it infers likely long-term navigation goals of the human based on their previous behavior.
The core idea of this work evolved from more traditional pattern-based prediction approaches, as e.g. presented by Bayoumi et al. They used typical trajectories to learn a foresighted navigation policy for a service robot via Q-learning [66] and to find a user quickly if they cannot be located in the proximity of the robot [71]. In contrast to the work presented in this chapter, the authors did not use knowledge about object transitions during the learning or in the prediction. While their approach has been applied successfully to the above mentioned scenarios, its training is intensive and needs new training data for each map. In contrast to that, the presented framework utilizes the same training data for multiple environments.
Another possible approach is given by Vasquez *et al.* [75]. The authors propose to create a joint probability distribution to predict the movement of a human based on observed position changes, a pre-trained, cost-based prediction model, and a gradient-based goal prediction function. In contrast to the presented approach, this system works only for short-term motion prediction.

Ziebart *et al.* [69] presented a prediction method that uses the maximum entropy and argued that humans plan their movement according to cost functions that assign costs to environmental features, such as surfaces or available spaces. The authors aim to learn these functions based on observations and then use the learned model to predict future movements. In this method, objects are only implicitly considered, i.e., as environmental features. Note that by predicting the destination of a human as in this work, it is also possible to infer the path the human will probably take using the assumption that humans operate based on a cost function as in those approaches.

Other existing motion prediction methods include velocity-based modeling of future human movements [76, 77] or learning of social models to predict the behavior of humans in lively places [70, 70, 78]. However, those approaches have been developed for short-term prediction of human motions and trajectory readaptation of a mobile robot and not for more foresighted navigation as in our application.

Several frameworks for navigation prediction use neural networks, e.g., Alahi *et al.* developed an approach to predict the future trajectory of people based on their past positions using an LSTM for crowed spaces [72]. Pfeiffer *et al.* followed a similar approach to create a data-driven interaction aware motion prediction system using an LSTM, which was trained by demonstrating typical human motions [79]. These approaches are mostly suited for crowded spaces where the robot needs to anticipate the intermediate behavior of many humans to avoid collisions or other undesired behavior.

Another possible implementation of a long-term prediction framework is Bayesian filtering, as demonstrated by Glover *et al.* [80]. The authors proposed a method to extract the navigation goal of a user based on their walking activities for a robotic walking aid. They accomplished this by using Bayesian filtering in combination with a hierarchical Markov model trained on typical user movements. Similarly, Best and Fitch applied a Bayesian framework to estimate the navigation goal and future trajectory of a mobile agent in a static environment by assuming that the agent is traveling to predefined goal locations on the shortest path [81].

Social forces were also often seen as driving factor for movement predicting, especially in crowded scenarios. One of the earliest approaches was presented by Helbing *et al.* [82]. The main idea is to balance accelerating forces towards desirable states and decelerating forces away from obstacles and other humans. A newer approach using this idea is

presented by Rudenko *et al.* who propose a weighted random walk algorithm in which each agent is locally influenced by social forces of other agents [83]. These models are again mostly used for crowded spaces with multiple humans. A familiar idea is presented by Karaoğuz *et al.*, who proposed a human-centric partitioning of the environment by identifying objects that are commonly associated with frequent human presence and creating regions around them [84]. In contrast to activity regions the authors used the interaction frequency as classification criteria while we use proximity and composition based classification built on previously collected human feedback.

While object-based prediction has not been applied in this fashion in existing motion prediction systems, objects have often been used in the context of higher-level action prediction [85, 58]. In these frameworks, predicted actions are typically associated with objects, e.g., if a person holds a plate the next action will likely be setting the plate on some kind of surface or table. Those approaches have been used successfully in a local context but the authors did not consider general human-object interactions including moving to other places. Action recognition frameworks are also steadily improved, as they are essential for robots living and collaborating with humans on a daily basis. A recent publication in this field is given by Duckworth *et al.* [86]. The authors present a framework that uses low-dimensional representations of human observations from a mobile robot to learn and identify human activities in visual data.

All in all the framework presented in this chapter extends the approaches from the literature by exploring the previously often implicitly used relation between past and future human-object interactions, moving away from learned map specific trajectories and short term movement prediction.

3.3 Activity Regions

An important observation for this work is that humans often interact with multiple objects in close proximity before starting to move to the next navigation goal. As an example, let us consider office work, where a human typically simultaneously interacts with a chair, table, and computer before moving to some other place. If we would count each interaction individually, we would need to modeled this as three different object interactions, first an interaction with a chair, then with the table, and finally with the computer. This would

result in multiple new navigation goal predictions in a very short time. To achieve a better generalization, we can group such objects together into so-called *activity regions*. In the above example we would then only record interactions with different objects from the same activity region. This section discusses how to identify such regions. For this, mostly objects that overlap, i.e., objects with approximately equal depth values, are considered as parts of the same activity region, such as a table and objects on top of it. However, often objects that do not overlap are also used in combination for an activity, e.g., a chair and a table. To create activity regions that make intuitively sense to humans the results of an online survey on how humans perceive such regions and how specific activity regions are composed were used.
In the remainder of this section we discuss both, the design of this survey as well as its results.

3.3.1 Design of an Online Survey to Identify Activity Regions

The survey[1] was created using Qualtrics [87] and published on Clickworker [88]. It was online for one week and during this time 125 users participated. It included an attention test, which was passed by 106 german participants from a cross section of the population [89]. Responses from users that did not pass the attention test were excluded.
The survey included three different types of questions. During the first type participants were shown pictures of office environments with given, color-coded groupings for different objects and asked which of the options felt most natural to them (see Fig. 3.2 for an example), to gain insight into possible subconscious classification rules. 15 questions of this type were used, their order was randomized as well as the order of possible answers.
For the second part, participants were asked which objects they consider to be in proximity to given objects. The participants had to indicate the likelihood of ten objects on an even likert scale, which had six options ranging from very unlikely in close proximity to very likely in close proximity. Seven questions of this type including one which serves as attention test were used. Questions and possible answers were again randomized.
The last part of the survey was concerned with the association of specific objects with two example activities: *food processing* and *office work*. Participants could choose between 15

[1] The survey was published in German, a complete copy of it can be found on at the Humanoid Robots Lab website: https://www.hrl.uni-bonn.de/publications/activity_region_survey

(a) (b)

Figure 3.2: Example question of the survey. The participants were asked which one of the two object groupings appeared more natural to them, with the hope of gaining insight into possible subconscious classification rules. We recorded 106 answers to this question. 72 participants voted for option (a) while 34 voted for option (b).

different objects: table, refrigerator, PC, chair, toilet, lamp, shelf, sink, washing machine, sofa, microwave, white board, dresser, coffee maker, bed. The chosen objects should then be ordered based on their associated importance for the given task. These example tasks were chosen as they represent activities where the robot would possibly be able to provide assistance.

3.3.2 Results of the Survey

The results of the survey[2] shows a clear trend to group specific objects together. This trend is especially strong with chairs and tables, as chairs were grouped with the nearest table even if it was more than two meters away. Overall five different questions regarding the grouping of chairs to tables with table-chair distances ranging from less than one meter to more than two meters were asked. In 66% of the answers the chair was grouped with the nearest tables regardless of the distance.

No significant difference could be observed regarding the grouping of tables, as neither the grouping nor the non grouping scenario seems to be preferred by the users. Fig. 3.3 shows an example of a question regarding possible groupings of close tables.

[2] The complete results of the survey can be found on the Humanoid Robots Lab website https://www.hrl.uni-bonn.de/publications/activity_region_survey_results

(a) (b)

Figure 3.3: Example table grouping question from the survey. Participants were asked if they found grouping (a) or grouping (b) more natural. 106 answers were recorded to this question. The results were exactly split as 53 participants voted for option (a) and the remaining 53 for option (b).

The results also show a clear classification of objects into the two example activity classes, office work and food processing, as can be seen in Fig. 3.4 and Fig. 3.5, respectively. This supports the theory that grouping objects into regions related to activities matches typical human behavior.

In summary, the survey showed that the participants indeed tend to group objects based on proximity and functionality, that objects in the same group will in most cases not be further away than 2 meters from each other, and that the participants could clearly identify objects that they would expect inside two example activity regions. Using these results, we can build a proximity based rule system to automatically identify activity regions on a semantic map, as we describe in Sec. 3.4.1

These results also encourage further research in this direction, as knowledge about the composition of typical activity regions could be obtained this way, which in turn could strengthen activity region detection systems.

3.4 Prediction of Navigation Goals

As explained above, in this chapter we consider the problem of predicting the navigation goal of a moving human in an indoor environment. Thereby, the prediction is based on observations of the user's location and pose as well as on prior knowledge about a map of

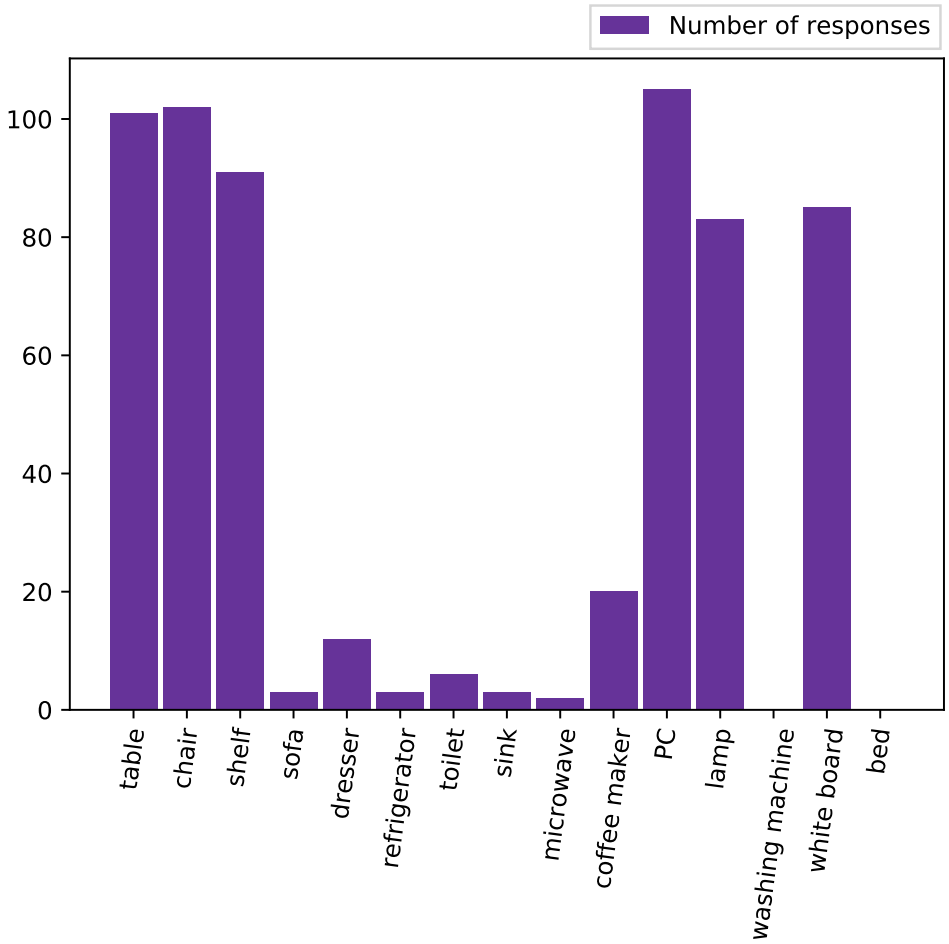


Figure 3.4: Responses of the 106 survey participants regarding the question which objects they expect in an office environment. As can be seen, there is a clear expectation towards the objects: table, chair, shelf, PC, lamp, and white board.

the environment and typical human transitions between objects. The prior knowledge can be obtained by learning a semantic map of the environment [90] and afterwards grouping of the objects into activity regions, as discussed in Sec. 3.3.2.

Furthermore, pre-recorded videos of humans acting in indoor environments were used to train a prediction model for transitions of human-object interactions. In other words, a model was created to predict how likely it is that a human who interacted with object A

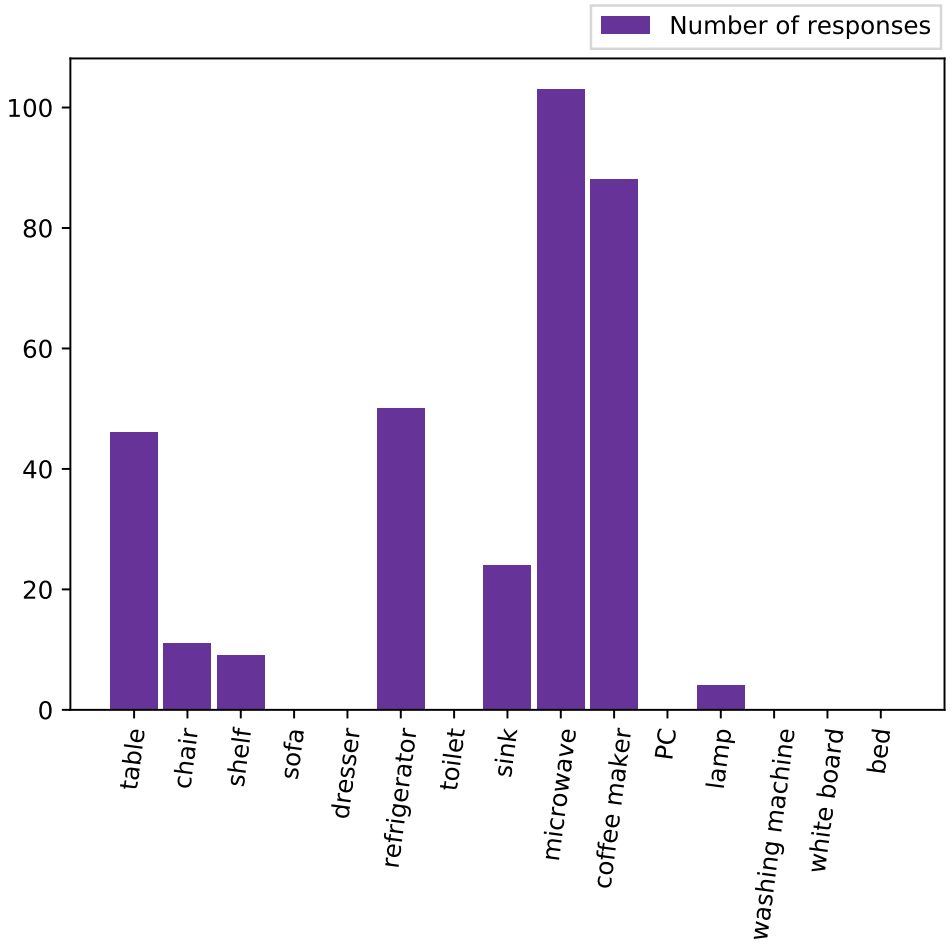


Figure 3.5: Responses of the 106 survey participants regarding the question which objects they expect in a food processing environment. As can be seen, there is a clear expectation towards the objects: table, refrigerator, microwave, and coffee maker.

will next interact with object B and called *interaction model*. The training videos for which the Humanoid Robot Lab, in which facilities the videos were recorded, holds the associated rights are published with the Bonn Activity Maps dataset [91].

Based on this prior knowledge and observations about the user's location and pose, Bayesian inference is applied to predict their next navigation goal.

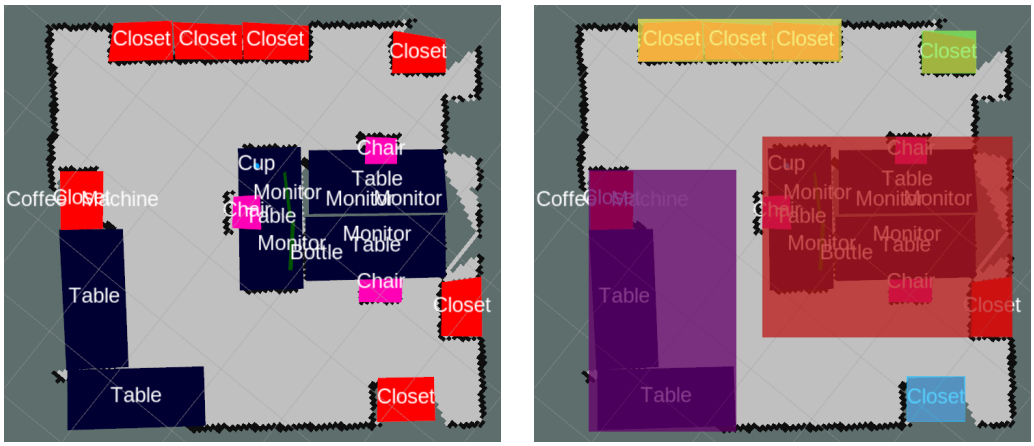
In the following, all components of the prediction framework are described in detail.

3.4.1 Semantic Environment Representation

The environment is represented as a static inflated occupancy map [92], with an additional semantic layer to encode objects and activity regions. The occupancy mapping can be realized with a common SLAM approach [93]. Object information can be added to the map through semantic mapping by using RGB-D masks from CNN object detectors and projecting them to the 2D plane [90]. To infer activity regions for a map a rule-based system is used. First, each object is assigned its own activity region. If two regions overlap, i.e., have overlapping object bounding boxes with some approximately equal depth values inside them, they are merged. The resulting region consists of all objects of the previous regions. This process is repeated until no more regions can be merged. Second, the results of the survey discussed in Sec. 3.3.2 are used by merging activity regions that are less than 2 meters apart from each other. The final environment representation consists of the inflated occupancy map M, the position $\mathcal{X}_o$ and type τ_o of each object $o = (\mathcal{X}_o, \tau_o)$ as well as the position $\mathcal{X}_R$ and object composition $C_R = \{o_a, o_b, \dots\}$ of each activity region $R = (\mathcal{X}_R, C_R)$ on M. The position of an activity region $\mathcal{X}_R$ is defined as the center of the bounding box around all objects inside the region. Fig. 3.6 shows an example of the generation of activity regions for a previously recorded semantic map. Note that the human is not part of the map.

3.4.2 Interaction Model

Let $I(\tau_a|\tau_b)$ be a distribution that describes the probability that a human who previously interacted with an object of type τ_b will next interact with an object of type τ_a, called interaction model. This model was trained on recorded object interaction sequences of humans in indoor environments. To generalize well between different environments, only object-interaction sequences were considered and not the actual trajectories of the human. As I serves as prior knowledge, it is not further updated once learned. This concept can be extended to activity regions by defining the transition probability between two regions R_a and R_b as a function of the transition probabilities between the objects of these regions. Hereby, the normalized sum of all transition probabilities between objects of different types in the individual regions is used. In other words, if two chairs are present in region A and one sofa in region B, only the transition probability of one chair towards the sofa is



(a) (b)

Figure 3.6: Example of the generation of activity regions. (a) A semantic map which contains 23 objects. (b) All objects that are less than 2 meters apart from each other are then grouped into the same activity region. This results in 5 different activity regions, shown as bounding boxes around the objects.

used. Formally, the regional interaction model $I_R(R_b, R_a)$, which encodes the probability that a human that interacted with an object from activity region R_a will next interact with an object from activity region R_b, is defined as follows:

$$I_R(R_b|R_a) = \eta \cdot \sum_{x \in T_b} \left(\sum_{y \in T_a} I(x|y) \right) \tag{3.1}$$

With η as normalizing parameter and $T_a = \bigcup_{\tau_o} C_{R_a}$ and $T_b = \bigcup_{\tau_o} C_{R_b}$ as the sets of unique object types present in R_a and R_b respectively. As the positions of objects and activity regions is given as prior knowledge, it can be directly inferred with which activity region a human is interacting by observing a specific object interaction. Thereby the actual class of the object with which the user interacted is relatively irrelevant, as we only need to detect that an interaction did occur.

3.4.3 Observations About the Human

The interaction model on its own is not sufficient for a reliable prediction of the next navigation goal of the human, as it does not consider their position and orientation after the interaction took place.
Thus, further RGB-D observations must be used to obtain additional information about the user. Therefore the framework first detects the human and their pose using a pose estimation system (OpenPose [62]). Once the human is detected, the distance towards them is computed, e.g. using data from a laser sensor. With the known position of the robot on M, we can now infer the position $\mathcal{X}_h$ of the human on M. For the orientation of the human θ_h the approach uses pose data to track the face and shoulders. It then infers the general direction the human is orientated to, as shown in Chapter 2. The complete state of the human at the current time step is defined as $S := (\mathcal{X}_h, \theta_h)$. For simplicity the approach presented in this chapter only considers scenarios in which one human is present, however it can be extended to work with multiple humans, as long as these can be distinguished. Identifying a specific human could also be a powerful cue as it would be possible to use a tailored interaction model for individual users, possibly greatly increasing the prediction accuracy of the approach. On the other side this would also greatly increase the necessary training data, as individual data for each user would be necessary.

3.4.4 Bayesian Inference

The approach uses Bayesian inference based on the prior knowledge about likely transitions of the human between activity regions in combination with the continuous observations about the human's state and the last object the human interacted with to predict the human's next navigation goal. Note that as stated in Sec. 3.4.2, knowledge about the last objects the human interacted with directly implies knowledge about the last activity region the human interacted with. The prior probability of an activity region $P(R_i)$ is given by the pre-trained regional interaction model $I_R(R_i|R_L)$ between the activity region which serves as possible navigation goal R_i and the activity region in which the last object interaction was observed R_L.
Let $\mathcal{R}$ be the set of all activity regions on M. The probability that the activity region R_i is the human's navigation goal given the current observation of their state S is given as:

$$P(R_i|S) = \frac{P(S|R_i)P(R_i)}{\sum_{R_j \in \mathcal{R}} P(S|R_j)P(R_j)} \tag{3.2}$$

Using η as a normalizing parameter Eq. (3.2) can therefore be simplified to:

$$P(R_i|S) = \eta \cdot P(S|R_i)I_R(R_i|R_L) \tag{3.3}$$

It is possible that the robot did not observe the last interaction. In this case, we can use the marginalized region interaction probability over each possible activity region:

$$I_R(R_i) = \sum_{R_j \in \mathcal{R}} I_R(R_i|R_j) \tag{3.4}$$

This modifies Eq. (3.2) in the case of no observed interaction to:

$$P(R_i|S) = \eta \cdot P(S|R_i) \cdot I(R_i). \tag{3.5}$$

In both Eq. (3.2) and Eq. (3.4), $P(S|R_i)$ corresponds to the likelihood of the human's observed state $S := (\mathcal{X}_h, \theta_h)$ given the possible navigation goal R_i. To evaluate this likelihood, we use the assumption that the user moves on the shortest A* path towards their navigation goal. Therefore, we can compute the shortest 2D A* path $\mathcal{P}_{h \to R_i}$ from the user's position $\mathcal{X}_h$ to the center of the region $\mathcal{X}_{R_i}$. Furthermore, let $\Delta a(\theta_h, \theta_{opt})$ be the 2D orientation difference of the human's current orientation θ_h and the orientation θ_{opt} the human would have if they moved to the next position on $\mathcal{P}_{h \to R_i}$. Let $dist(\mathcal{X}_h, \mathcal{X}_{R_i})$ be the A* distance between $\mathcal{X}_h$ and $\mathcal{X}_{R_i}$, i.e. the length of the path $\mathcal{P}_{h \to R_i}$. With an added value of 1 to avoid situations in which we would divide by zero, the observation likelihood $P(S|R_j)$ can then be defined as:

$$P(S|R_j) = (dist(\mathcal{X}_h, \mathcal{X}_{R_i}) + 1)^{-1} \cdot (\Delta a(\theta_h, \theta_{opt}) + 1)^{-1}. \tag{3.6}$$

Combining all equations, the probability $P(R_i|S)$ that the activity region R_i is the navigation goal of the human given the prior knowledge is defined as

$$P(R_i|S) = \eta \cdot (dist(\mathcal{X}_h, \mathcal{X}_{R_i}) + 1)^{-1} \cdot (\Delta a(\theta_h, \theta_{opt}) + 1)^{-1} \cdot I_R(R_i|R_L) \tag{3.7}$$

if the last object interaction was observed and otherwise as

$$P(R_i|S) = \eta \cdot (dist(\mathcal{X}_h, \mathcal{X}_{R_i}) + 1)^{-1} \cdot (\Delta a(\theta_h, \theta_{opt}) + 1)^{-1} \cdot I_R(R_i) \tag{3.8}$$

The belief about the navigation goal is updated in a constant interval as long as the human is visible and moving.

3.5 Experimental Evaluation

The framework was evaluated based on the accuracy of the prediction in regards to the human's true navigation goal. A quantitative and qualitative evaluation was performed as well as a comparison to previous approaches.

3.5.1 Data Collection

To guarantee comparability of different approaches and eliminate noise in the observations, the majority of the experiments were performed in simulation. Using the V-REP editor [94], 10 different office and home environments with sizes between 100 m^2 and 150 m^2, modeled after real-world examples, were created. Each environment contains up to 110 different objects from 16 different objects classes: bottles, cups, microwaves, chairs, tables, beds, toilets, handbasins, bathtubs, washbasins, cupboards, wardrobes, refrigerators, sofas, and laptops. Note that as explained in Chapter 2, the previously presented interaction detector is able to register interactions with the 510 different objects and animals from the Open Image dataset [95] using an R-CNN trained on the dataset. However, as some of these objects are usually not perceived inside an office or home environment or were underrepresented in the training set for the interaction model, the number of objects for the evaluation was restricted. The used interaction model was trained with a set of 161 recorded human-object transitions [41] from which the regional interaction model was determined. Furthermore, a dataset of 64 typical human transitions between two objects was collected, based on recorded movements within the buildings of the University of Bonn as well as a survey about typical indoor movements inside home environments. As mentioned in Sec. 3.4, the videos for which the Humanoid Robots Labs hold the associated rights are published with the Bonn Activity Maps dataset [91].

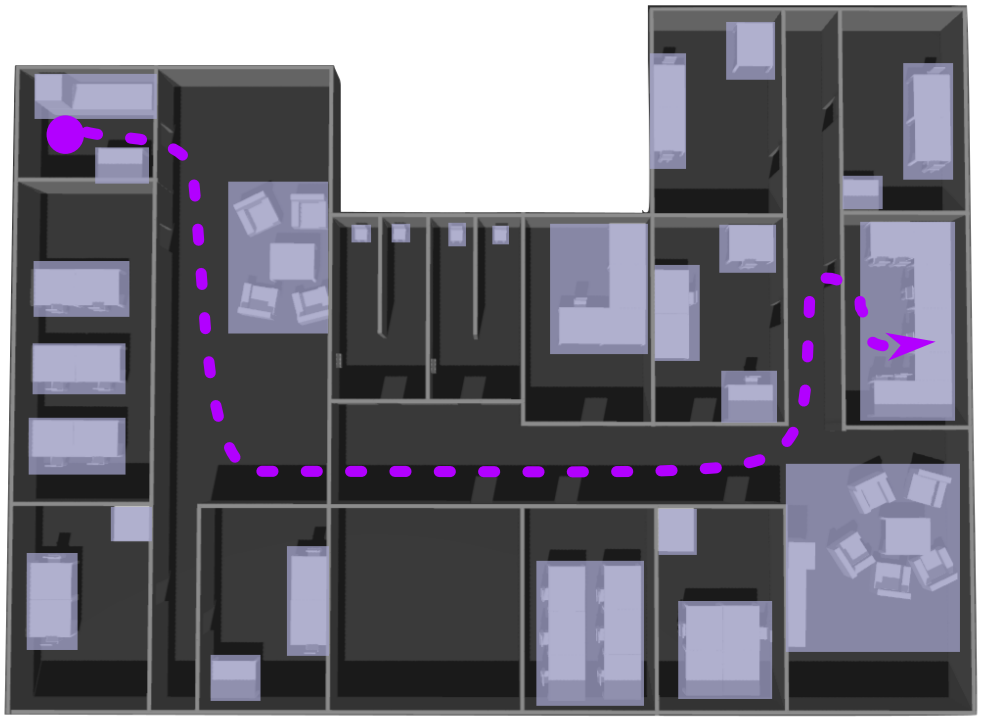


Figure 3.7: Example simulated office map. Objects are shown in gray with activity regions in light gray on top of them. An example trajectory from the starting position of the user (violet circle) near a coffee machine to an office is shown as dotted line (violet). Using activity regions, the number of navigation goals that the robot needs to consider shrinks from 93 (the number of objects present) to 27 (the number of activity regions). This map is based on a real-world office environment at the Computer Science Department of the University of Bonn.

The data was used to randomly sample 300 transitions between objects. Based on these transitions, 300 different trajectories distributed over all environments were computed using A*. Within a specific map these trajectories are sequential, in other words the trajectory number $n + 1$ begins with the object with which the trajectory number n ended. The trajectories were used as test data for all evaluations.

3.5.2 Quantitative Evaluation

For the quantitative evaluation, the approach was tested on 300 trajectories in 10 different office and home environments, as described in Sec. 3.5.1. Each trajectory had on average 64 different possible goal objects, grouped in 19 different activity regions. Fig. 3.7 shows an example of one of the office maps with a test trajectory and possible navigation goals. During the first evaluation, the human was always observable while moving and the robot had perfect observation, i.e. the probability for false positive or negative observations was 0. Once every second the system made a prediction about the human's most likely navigation goal and compared this prediction to the human's true navigation goal. This general prediction accuracy is used as standard metric in the evaluation, as well as a variant called *top 5%* accuracy. Here, a prediction is counted as correct if the true navigation goal is among the returned top 5% of most likely navigation goals. This metric was chosen to show that even if the approach temporarily predicts a wrong navigation goal, the true navigation goal is often still among the top 5%. On average three navigation goals are within this range. Additionally, the A* distance from the center of the prediction navigation goal to the true navigation goal of the user is used as metric. This metric provides insight in the severeness of false predictions, as wrongly predicted destinations might still be close to the true navigation goal.

The presented region based approach was evaluated both with and without a known initial object interaction, as specified in Eq. (3.7) and Eq. (3.8) respectively. For comparison, the experiments were also conducted with an older version of the presented approach [67], which does not use activity regions and uses objects as possible navigation goals, as well as with a trajectory-based prediction approach [71], which does not use information about object interactions and does not consider the orientation in the prediction. Activity regions were used as navigation goals for the presented approach and the trajectory-based approach [71], while objects were used as possible navigation goals for the object-based approach [67]. Both the activity region and object-based approaches used the same interaction model trained on 161 recorded human-object transitions for all maps. The trajectory based approach needed to be trained for each of the 10 maps using 30 additional sampled trajectories for each map. The results of the quantitative evaluation are given in Tab. 3.1.

As can be seen, the use of activity regions improves the average prediction accuracy by

	Avg. General Prediction Accuracy	Avg. Top 5% Prediction Accuracy	Avg. Dist to True Nav. Goal [m]
Last interaction observed	**0.68**	**0.71**	**5.06**
Last interaction **not** observed	0.64	0.66	5.74
Last interaction observed [67]	0.53	0.68	5.09
Last interaction **not** observed [67]	0.46	0.66	5.63
Trajectory-based approach [71]	0.36	0.57	6.86

Table 3.1: Results of the quantitative evaluation. The first two rows represent the results of the approach with activity regions. The third and fourth row show the results of the previous work [67] with single objects instead of activity regions and the last row shows the results of the trajectory-based apprach [71].

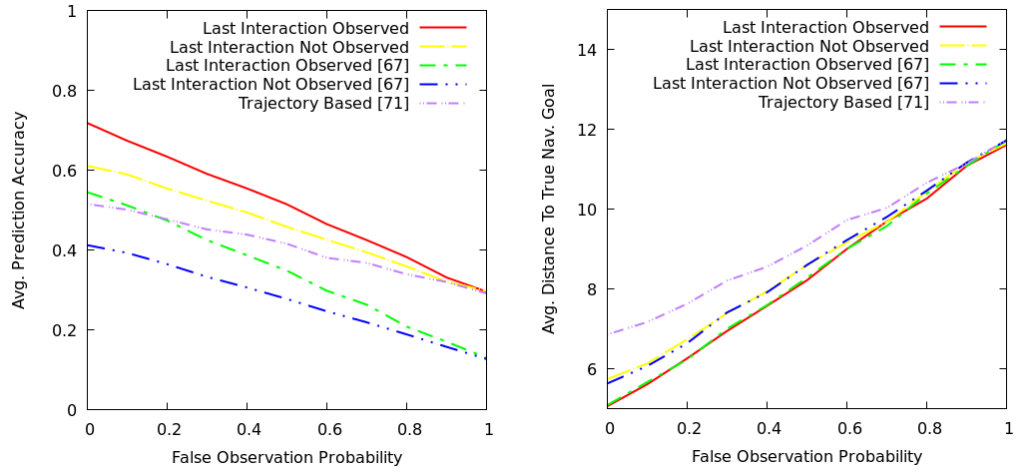
0.15 if the last interaction was observed, in comparison with older approaches without them [67]. Similar results are achieved if the last interaction was not observed, here an improvement of 0.20 was achieved. Both results are statistically significant based on the paired t-test. An interesting result is the closing of the gap between the general prediction accuracy and the top 5% prediction accuracy. In the approach without activity regions, there is a difference of 0.15 between the two metrics if the last interaction was observed, this shrinks to 0.03 if activity regions are used. The effect also exists if the last interaction was not observed. This implies that cases in which the true navigation goal was not the most likely one but is among the top 5% most likely goals, it is in many cases part of the most likely activity region. An example for this would be a resting area with a table and a chair. Both object types are commonly used together but only one can be the most likely goal object. If we use activity regions, this problem disappears, as in this case the human interacts with the region consisting of both objects. The average distance between the predicted and true navigation goal supports this theory. The resulting distances for the region- and object-based approaches are very close together. This can be interpreted as a sign that false predictions in the object-based approach correspond in most cases to objects which are now grouped inside the same activity region. Note that even if the

improvement in the actual distance is small, the significant gain in accuracy may be very important depending on the application scenario. One example would be a service robot that needs to infer the n-th destination of the user based on previous interactions. The more accurate and simplified scenario of activity regions would be much more suitable for such a scenario than the more complex approaches with objects as destinations or learned trajectories. In comparison to the trajectory-based reinforcement learning approach that does not explicitly consider object interactions [71], the presented approach achieves an improvement of 0.32 for the average accuracy and an improvement of 0.14 for the top 5% metric. Furthermore, it is able to reduce the average distance between the predicted and true navigation goal by 1.8 meters.

The next evaluation was concerned with the influence of false positive observations. Therefore a probability for an observation error, i.e. a random user pose or object interaction instead of the correct observation, was added. The results are depicted in Fig. 3.8. As can be seen, all approaches perform linearly worse with a higher likelihood of false observations. The activity region and trajectory based approaches perform slightly better, but this is likely due to a smaller number of possible goal locations in comparison with the object-based approach. However, the results also show that the presented approach is able to yield an average prediction accuracy of more than 0.6 up to a 40% probability of false observations.

3.5.3 Qualitative Evaluation

To evaluate the importance of individual observations at different times during the movement of the human, an experiment with focus on the changes in the returned probability of the true navigation goal over time was conducted. Fig. 3.9 shows the results as the evolution of the probability with respect to the position of the human on a typical test trajectory for which the last object interaction is known. As can be seen, at a position after roughly 40% of the trajectory the true navigation goal is continuously returned as the most likely navigation goal. This effect is observable within the whole dataset, once the true navigation goal is identified as the most likely goal it does not change anymore in almost all cases. This highlights an interesting property of the approach. The importance of observation decreases the closer the human gets towards their navigation goal.



(a) (b)

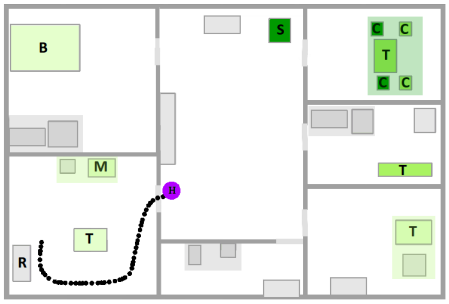
Figure 3.8: Change of the average prediction accuracy (a) and average distance between the predicted and true navigation goal (b) with false observations. As can be seen, the accuracy and distance decrease linearly for all evaluated approaches.

Once the last 60% of the trajectory are reached the predicted goal region usually doesn't change anymore. As we do not know how long the final trajectory of the user will be, we cannot directly use this knowledge. However, it may be possible to add a function which decreases the impact of the prediction during the first update steps, but further research in this direction would be necessary.

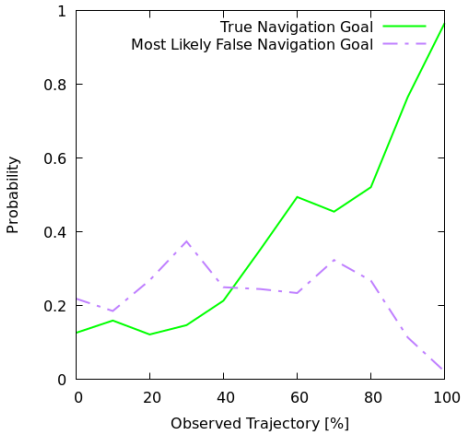
Fig. 3.10 shows another example of a typical human trajectory and the evolution of the returned probability of the true navigation goal over time.

3.5.4 Application

A framework for predicting human movements is essential for foresighted robot navigation [34]. To highlight this use case, the presented approach can be combined with a positioning approach. The framework was tesed on a Robotino mobile platform [96] with an RGB-D camera and a laser scanner in a university environment. The robot uses a 2D grid map representation of the environment with a discretization of 0.75 meters to decide where to place itself in order to provide assistance to the human if needed while



(a) Example trajectory for which the evolution of the goal probabilities is shown to the right. Object names are abbreviated: dining table (T), microwave (M), bed (B), chair (C), and sofa (S). Activity regions are given as colored shades, the darker the color the higher the likelihood.



(b) Evolution of the belief about the navigation goal with respect to the percentage of the observed length of a typical trajectory, which is depicted on the left.

Figure 3.9: Example of a typical trajectory in the simulated environment. In this case, the human (violet) first interacts with a refrigerator and then moves to a chair-table activity region. (a) Trajectory observed up to the point where the prediction is correct for the first time. (b) Corresponding evolution of the belief about the navigation goal.

simultaneously avoiding interferences. The prediction is updated every 5 seconds. To compute the optimal position of the robot, it uses the average distance between all possible navigation goals weighted by their probability while avoiding positions in a $1.5m$ radius around the human. It further uses a function $C(\mathcal{X}, S)$ to compute the costs of each possible position $\mathcal{X}$ on M based on the current observation of the human $S = (\mathcal{X}_h, \theta_h)$ and prior knowledge:

$$C(\mathcal{X}, S) = \begin{cases} \infty \text{ if } dist(\mathcal{X}, \mathcal{X}_h) < 1.5m \\ \sum_{R_i \in \mathcal{R}} (P(R_i|S) \cdot \texttt{dist}(\mathcal{X}, \mathcal{X}_h)) \text{ else} \end{cases} \tag{3.9}$$

With $dist(\mathcal{X}_a, \mathcal{X}_b)$ as the A* distance between the positions $\mathcal{X}_a$ and $\mathcal{X}_b$ in meters. The position with the lowest cost is then used as new destination for the robot.

Fig. 3.11 illustrates an example experiment. Here, the robot was in a corridor where

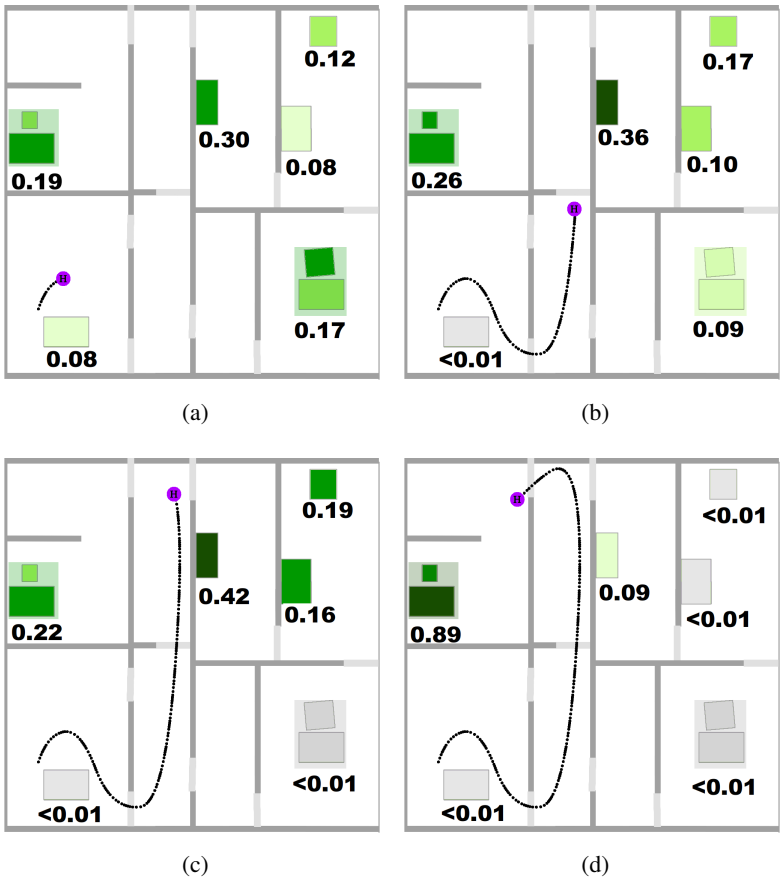


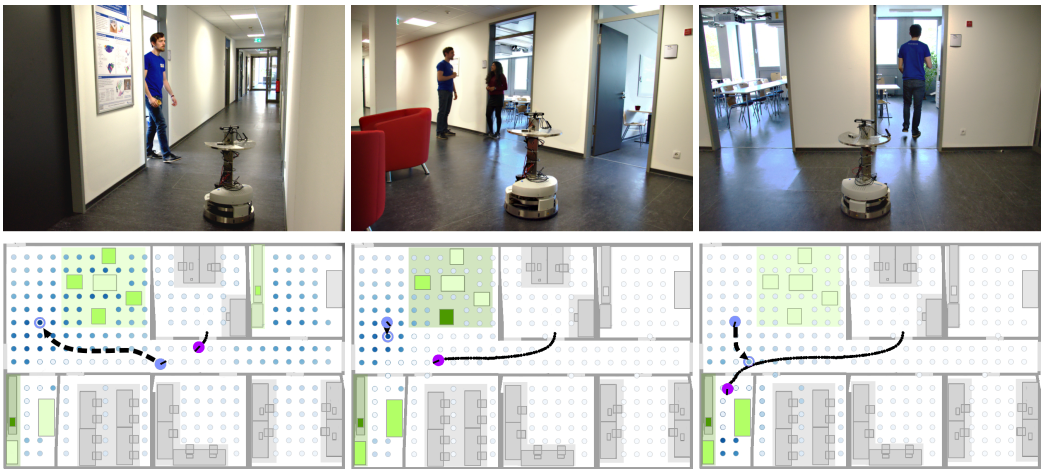
Figure 3.10: Evolution of the belief over time for an example trajectory of a user (violet) in a simulated environment. As can be seen, the initial belief based on the interaction model (a) is continuously updated with new observations (b), (c), (d). Objects are shown in green and activity regions in coloured shades with their probabilities to be the navigation goal: the darker the green the higher the probability. Doors are colored in gray and walls in dark gray. The human is depicted as a violet dot with their trajectory as dashed line.

it observed a human-object interaction with a cup. The robot then updated the prediction about the navigation goal and computed a new position for itself based on Eq. (3.9) (see Fig. 3.11 (a)). During the movement, the robot regularly updates both, the prediction of the navigation goal as well as the weight values of possible positions (see Fig. 3.11 (b)). In Fig. 3.11 (c), the human has entered the room containing the

navigation goal and the robot correctly updated its prediction. The robot does not enter the room itself since positions near the human have infinite weight. However, if the human called the robot to help them, the robot would be there immediately due to its foresighted positioning. As a result, the robot can avoid interference but is still close to the human to react quickly if needed.
This use case was further tested in the simulated environment, using the same environments and trajectories as in Sec. 3.5.2, thereby assuming that the human is always observable for the robot as well as perfect sensors, i.e. a false positive observation chance of 0. The user was configured to travel with an average velocity of $1\frac{m}{s}$ and the robot to travel with $2\frac{m}{s}$. As metric the average arrival time difference between the user and the robot when they first entered an area $1.5m$ around the goal activity region was used. The results showed that the robot did on average arrive $8.5s$ before the human.
We will further investigate such applications and possible navigation strategies in the next chapter.

3.6 Discussion

As shown in Sec. 3.5, the presented framework reliably predicts the navigation goal of a moving human on the evaluation dataset. It can be used in any indoor environment without the need of specific training, as long as the relevant objects and their transition probabilities are known. However, there are also typical cases where the method leads to ambiguities, e.g., if two likely goal regions lie on the same path and the latter is the true navigation goal. In order to approach the true navigation goal, the human naturally also approaches the first equally likely navigation goal. Therefore, the framework has no means to rule out the wrong navigation goal until the human passes by. Such cases can be solved in the future by further enhancing the prior knowledge, e.g. with a better trained interaction model. Another possible improvement could be to distinguish objects based on the activity that the user performs at them, for example by ignoring objects in our prediction at which the user does not need help of a service robot and instead directly predicting the next destination at which the user will need assistance.
Other possible improvements of the system could involve an improved detection of human-object interactions, e.g. by the utilization of smart home systems and their detection



(a) (b) (c)

Figure 3.11: Application example of the presented approach for foresighted navigation. Objects are shown in green and possible placement positions for the robot are shown in blue. Activity regions are given in coloured shades. The orientation of the robot (blue circle) and user (violet circle) is indicated by a black line. The darker the color the lower the costs. (a) The robot observes a moving human that has previously interacted with an object inside an office activity region. Based on this information, it updates the belief about the navigation goal and computes a new position for itself (blue ring) taking into account the human's most likely navigation goals while avoiding interference with the human and their predicted path. (b) The prediction as well as the robot's placement position are updated with new observations. (c) The human enters a room and the robot adapts its position to be close to the human in order to enable quick reaction when called for assistance. The robot ignores positions in direct proximity to the human even if they lead to a higher utility to minimize interferences.

capabilities. As demonstrated in Sec. 4.6, knowledge about the last object interaction improves the accuracy of the prediction and knowing the position of the human is essential to the entire approach. Another possibility would be the use of a real-time localization system, e.g., an ultra wide-band system, as interaction information would then be available at all time.

3.7 Conclusion

In this chapter, we discussed an approach to predict the navigation goal of a moving human in indoor environments by utilizing knowledge about human-object interaction sequences. The approach uses knowledge about typical human-object interaction sequences and groups close-by objects into activity regions to achieve a better generalization. To get information about typical activity regions in indoor environments, an online survey with 125 participants was performed and evaluated.

To learn the object interaction model the presented system uses for the prediction, humans in indoor environments were observed and transition probabilities between object interactions were learned. This information was then utilized in combination with observations about the human's movement and pose to infer the navigation goal using Bayesian inference.

As demonstrated in various experiments, the framework reliably predicts the navigation goal of a moving human. By using transitions between the activity regions in contrast to single object transitions, it achieves a significant increase in the prediction accuracy. Furthermore, the system outperforms a trajectory-based prediction approach that relies on previously learned trajectories between fixed destinations. Finally, the presented approach was used in an experiment in which a mobile robot uses the new framework for foresighted navigation by computing favorable positions for itself taking into account the computed navigation goal prediction of the user.

In the next chapter we combine the interaction detector presented in Chapter 2 and the interaction based movement prediction presented in this chapter into a foresighted, human-aware navigation policy for a service robot. Thereby, a solution for the third key contribution of this thesis is presented: a time-dependent navigation policy to compute a human-aware path for a service robot to the next destination at which the user will need assistance.

CHAPTER 4

Human-Aware Robot Navigation by Long-Term Movement Prediction

This chapter combines the interaction detector presented in Chapter 2 and the interaction based movement prediction presented in Chapter 3 to a foresighted, human-aware navigation policy for a service robot. It thereby presents an approach for the third key contribution of this thesis: a time-dependent navigation policy to compute a human-aware path for a service robot to the next destination at which the user will need assistance.
As stated in Chapter 1, the content of this chapter has already been published [97, 98].

4.1 Introduction

To enable a widespread use of service robots it is important to find behavior strategies that ensure a harmonic coexistence between those systems and their users. A solution to this is to mimic human social behavior [34]. Humans are more than dynamic obstacles and special constraints need to be fulfilled to enable efficient robot navigation that still abides by social rules, ensuring human comfort. A prerequisite for this is the ability to predict human movements, thereby foreseeing and avoiding situations in which the robot could violate social constraints with its navigation policy.
This chapter presents an approach to accomplish human-aware robot navigation in indoor

environments. The approach utilizes a long term prediction system to infer likely future navigation goals of the user, estimates the likelihood that they will need assistance at those locations, and computes a robot position close to the most likely next location where the robot will be needed. The prediction is based on an updated version of the framework presented in Chapter 3, however, it aims to predict not only the next, but the n next navigation goals of the user. Using these predicted positions the system then plans the path of the robot in a time-dependent cost map that takes into account the user's predicted movements as well as compliance to human comfort. Both the prediction and the cost map are periodically updated based on new robot observations to deal with uncertainties and prediction errors.

During the path planing the robot needs to abide by rules of human comfort, which are based on existing knowledge [99, 34] as well as the results of a specifically conducted interview and online survey.

Fig. 4.1 depicts an application example of the approach. As can be seen, the robot chose a path to the next navigation goal of the user at which the robot can offer assistance. The robot also tries to minimize discomfort of the human, by avoiding the direct route, which interferes with the human path.

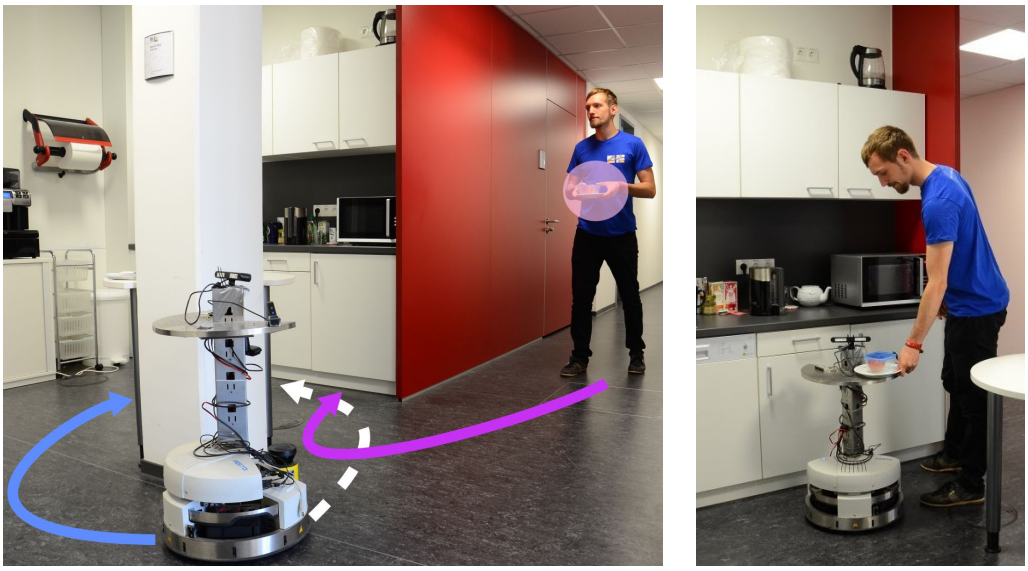
To summarize, the contributions presented in this chapter are the following:

1. An activity region based prediction system to infer the n next location where the user will need help based on previous object interactions and robot observations.

2. A human-aware navigation system based on long-term movement prediction, human comfort constraints, and path planning on a time-dependent cost-map.

3. A survey about human comfort to evaluate different navigation strategies.

4. An evaluation of the approach in comparison to state-of-the-art methods [100, 101, 73] in simulated environments with pre-defined metrics and in real-world and virtual reality experiments with direct human feedback.



(a) (b)

Figure 4.1: Motivating example of the approach. (a) The robot observes the moving user interacting with an object (pink circle) and predicts his most likely navigation goal and path (violet). Using the prediction, the robot is able to compute a short path to the navigation goal of the human (blue) while avoiding paths that would result in discomfort for the user (white). (b) Still, the robot arrives at the goal location before the user and early enough for direct assistance.

4.2 Related Work

A vast amount of research has been invested to determine human-aware robot navigation policies. Typical tasks for these include socially acceptable person following [100, 102], navigation through dense crowds [103, 104], and guiding people to a goal position [105, 106]. An overview is given by Kruse *et al.* [34]. The authors define three metrics to evaluate human-aware navigation policies: human comfort (absence of annoyance and stress for humans), naturalness (similarity between robot and human behavior), and sociability (adherence to cultural conventions). Common implementations of these metrics are proximity constraints w.r.t. the human [99] and anthropomorphic robot design [107] to increase the similarity between humans and robots, both visually and physically. The authors also highlighted that a reliable prediction about human movements is a prerequisite to achieve

good results with these metrics [99].

According to Foka *et al.* [108], prediction systems can be classified into short- and long-term prediction. A short-term system primarily forecasts the human motions for the next few time steps, while a long-term prediction system focuses on inferring navigation goals. Frameworks that aims at following a user commonly use short-term motion prediction, e.g., Pradhan *et al.* [109] proposed to use predictive fields to avoid moving obstacles and Ferrer *et al.* [73] developed a variant of the social force model [82] in which they additionally use the position and orientation of the human to predict their next movement. Kollmitz *et al.* [110] presented a method to predict the user's path to achieve a good local social navigation behavior of the robot. The authors proposed to model the sensitive area around the human using a Gaussian that decays with time to model the uncertainty in the prediction. The robot then takes into account the predicted occupied areas during planning to avoid interference. However, for applications that aim at generating foresighted robot behavior to reach the user's intended target locations, short-term motion prediction is not sufficient. Instead forecasting of the user's motion for a longer time horizon, i.e., long-term prediction, is necessary.

Long-term prediction systems often use a library of known paths [34] to predict the user's future motions based on observations. For example, Bayoumi *et al.* [101] developed a framework based on Q-learning to predict a user's navigation goal and determine the best robot actions. The learned policy is then applied to enable foresighted robot navigation. Usually, such approaches depend on a specific environment and typical human trajectories in it. The approach presented in this chapter, on the contrary, uses human transition probabilities between objects instead of key points on a given map so that the learned transitions are independent of a specific map [67, 41].

In contrast to all the methods discussed above, the presented approach combines long-term motion prediction with human-aware navigation, making explicit use of the prediction at every time step. This allows to use time-dependent path planning to avoid situations in which the robot would cause discomfort to the user.

Aside from navigation aspects there were various other improvements in social robotics over the last years. In a recent study, Tussyadiah and Park found that, besides familiarity with humans in form of anthropomorphic design and adherence to social norms, the perceived intelligence of the robot is of great importance for the user [111]. It can

be increased by correctly interpreting the intentions of a user and acting based on this information. This often corresponds to predicting either the user's next navigation goal or activity. For example, Ye *et al.* used a hidden Markov model to predict the category of the user's next activity (e.g., food, shopping, entertainment) and location where this activity is likely to occur [112]. Best and Fitch applied a Bayesian framework to estimate the navigation goal and future trajectory of an agent in a static environment [81]. Carlson *et al.* demonstrated how intention prediction can directly be used by robotic systems to help humans [113]. The authors designed an intelligent wheelchair that is able to predict the user's movement intention and helps to reach their navigation goal.
A promising development in this field is the utilization of smart home systems to monitor human-object interactions even if they are not directly observable through the robot. For example, Amri *et al.* used a multi-modal system to monitor activities of elderly people [114]. Muztoba *et al.* proposed special input devices with which a user can inform the robot about new commands at any location inside a smart home [115]. The authors experimented with gesture and speech detection as well as brain-machine interfaces. The system showed a lot of potential especially regarding the assistance of people with physical disabilities. Furthermore, Alam *et al.* demonstrated how a smart home system could be used in combination with a Markov model to accurately predict future activities of humans based on the observation history [116]. The presented approach would also benefit from the data collected in a smart home and could be easily deployed in such environments.

4.3 Constraints Derived From Studies About Human Comfort

An essential component of human-aware navigation is the identification and avoidance of robot actions that decrease human comfort. A detailed overview of research in this area is given by the surveys of Kruse *et al.* [34] and Rios-Martinez *et al.* [99] who analyzed that proximity rules are currently seen as most important for human comfort [117]. The approach presented in this chapter combines findings from these works with results from a survey which was conducted for this work, to define human comfort constraints as described in the following.
According to proxemic theory [118], humans have personal space regions around them in

that others, including robots, normally cannot intrude without causing discomfort. The size and form of these regions depend on the familiarity of the intruder, e.g., a friend is allowed to move closer to a person than a stranger. To model the allowed proximity of entities in an unfamiliar social context the *social zone (SZ)* is used, which is a circular interpersonal space region around the human with a radius of 1.2 meter [118]. As a service robot represents an example of an unfamiliar entity inside social context, we can model the SZ as a minimal distance that the robot must hold to humans. Furthermore, as noticed by Kitazawa *et al.* [119], objects inside a rectangular area with a length of 4 meters and a width of 1 meter in front of moving humans are considered to be potential obstacles. This area is called *information process space (IPS)*. Moving inside the IPS does thereby by definition disturb the path planning of humans and a robot should avoid this area to reduce interferences.

To find further constraints an interview survey was conducted for this work with 8 student participants from the University of Bonn. Each interview lasted around 15 minutes. Participants were asked about their feelings towards service robots, for which tasks they could envision to use robots, how they would design them, and whether there was any behavior that they saw as desirable or undesirable. All participants favored unobtrusive robot behavior. Interestingly, this was even more important than efficient working of the robot for most participants. Further, most participants disliked behavior in which the robot would follow them, move silently or unpredictably, or would come too close to them.

Based on those findings, a follow-up online survey[1] with 261 participants was conducted, distributed via Clickworker [88]. The survey was restricted to German participants to stay consistent with the interview survey and the later real world experiments. No other formal requirements for participants was given and, according to Clickworker, the participants represented a cross section of the population [89]. In the survey participants were asked how they feel about robots following them, which distances to robots they prefer (based on pictures) and how they rate unobtrusive robot behavior against efficient working of robots. The results show that 61% of the participants did not want the robot to follow them and 57% would accept a reduced work performance if the robot would then behave more

[1] The survey was originally conducted for the above mentioned IROS publication [97], its questions are available on the website of the Humanoid Robots Lab https://www.hrl.uni-bonn.de/publications/comfort_survey.pdf

unobtrusively supporting the trend of our interview study. Regarding the robot distance, 77% of the participants had no problem if the robot was moving at approximately 3 meters in front of them (as long as the robot did not enter their IPS) while only 58% would say the same if the robot was moving at the same distance behind them. However, this number increased to 70% if the robot was at least 5 meters behind the human.

Based on the results the following constraints were created: the robot should not enter the SZ and IPS, it should minimize close following of the human and prioritize these constraints over its work efficiency. The formal representation of these constraints and resulting path planning is presented in the next section.

4.4 Prediction of Future Object Interactions

As shown in Chapter 3 we can infer the next interaction based on observation of the last interaction and the human pose. This section shows how this approach can be further improved to also predict possible interactions that lie further in the future. This prediction can then be used to identify interaction, and thereby also locations, at which the human may need the assistance of the robot. Let M be a grid map of the environment. Like in Chapter 3 activity regions and Bayesian inference are used to infer future human-object interactions. As this approach is an extension of the work presented in Chapter 3 we will mostly use the same terminology and definitions as given in Sec. 3.4 and only define new concepts. The approach uses a pre-learned regional interaction model I_R as prior knowledge and the current human state $S := (\mathcal{X}_h, \theta_h, v_h)$ as observation, with $\mathcal{X}_h$ as 2D position, θ_h as orientation and v_h as speed of the human.

Let $bel(R_i^n) = P(R_i^n|S)$ be the belief that the user's n-th future interaction will be with activity region R_i given the observed user state S. The probability of an activity region interaction with $n > 1$ can then be recursively inferred using I_R as

$$bel(R_i^n) = \sum_{R_j^{n-1}} P(R_i^n | R_j^{n-1}, S) \cdot P(R_j^{n-1} | S) \tag{4.1}$$

$$= \sum_{R_j^{n-1}} P(R_i^n | R_j^{n-1}, S) \cdot bel(R_j^{n-1}) \tag{4.2}$$

$$= \sum_{R_j^{n-1}} P(R_i^n | R_j^{n-1}) \cdot bel(R_j^{n-1}) \tag{4.3}$$

$$= \sum_{R_j^{n-1}} I_R(R_i | R_j) \cdot bel(R_j^{n-1}) \tag{4.4}$$

using the law of total probability (Eq. (4.1)), recursion (Eq. (4.2)), the fact that R_j^n is independent of the current human state S given R_j^{n-1}, and the definition of the interaction model (Eq. (4.4)).

The interaction at $n = 1$ can be inferred with the current observation about the human state and the prior knowledge:

$$bel(R_i^1) = \frac{P(S|R_i^1)P(R_i^1)}{\sum_{R_j \in \mathcal{R}} P(S|R_j^1)P(R_j^1)} \tag{4.5}$$

$$= \eta \cdot P(S|R_i^1)P(R_i^1) \tag{4.6}$$

$$= \eta \cdot P(S|R_i^1)I_R(R_i|R^0) \tag{4.7}$$

using Bayes' rule (Eq. (4.5)), a normalization constant (Eq. (4.6)), and the definition of the interaction model (Eq. (4.7)), with R^0 as the last observed activity region with which the human interacted. If R^0 is unknown we obtain $I(R_i|R^0)$ using marginalization over all possible previous objects.

The likelihood $P(S|R_i^1)$ considers the user's orientation and distance to the center of the activity region $\mathcal{X}_{R_i}$. Let $\mathcal{P}_{\mathcal{X}_h \to \mathcal{X}_{R_i}}$ be the A* path from the position of the human $\mathcal{X}_h$ to the position of the activity region R_i. Let further $\Delta a(\theta, \theta_{opt})$ be the difference between the human's orientation θ and the orientation θ_{opt} they would have if they moved to the next position on the A* path $\mathcal{P}_{\mathcal{X}_h \to \mathcal{X}_{R_i}}$. Let $\Delta t(\mathcal{X}_h, \mathcal{X}_{R_i}, v)$ be the time the human would take

from their current position to $\mathcal{X}_{R_i}$ on $\mathcal{P}_{\mathcal{X}_h \to \mathcal{X}_{R_i}}$ with respect to their observed velocity ν. To decrease the likelihood of activity regions from which the user moves away and to model the fact that they cannot turn around spontaneously, the approach considers the distance the user would travel until the next observation update if $\Delta a(\theta, \theta_{opt}) > 180°$. In other words, if the user moves away from an activity regions, it is assumed that they cannot turn around before the next observation update is scheduled and this is taken into account when computing Δt, which is defined as follows:

$$\Delta t(\mathcal{X}_h, \mathcal{X}_{R_i}, \nu) = \begin{cases} \frac{dist(\mathcal{X}_h, \mathcal{X}_{R_i})}{\nu}, \text{ if } \Delta a(\theta, \theta_{opt}) < 180° \\ \frac{dist(\mathcal{X}_h, \mathcal{X}_{R_i}) + (\nu \cdot f_{update}^{-1})}{\nu}, \text{ else} \end{cases} \tag{4.8}$$

with $dist(\mathcal{X}_h, \mathcal{X}_{R_i})$ as the A* distance between the position of the human $\mathcal{X}_h$ and the position of the possible goal activity region $\mathcal{X}_{R_i}$ and f_{update} $[\frac{1}{s}]$ as the update frequency of the prediction.

The smaller the Δa and Δt, the higher the likelihood, therefore, the observation likelihood is defined as

$$P(S|R_i^1) = \Delta a(\theta, \theta_{opt})^{-1} \cdot \Delta t(\mathcal{X}_h, \mathcal{X}_{R_i}, \nu)^{-1}. \tag{4.9}$$

Combining all of the above, the believe in case $n = 1$ is given as

$$bel(R_i^1) = \eta \cdot \Delta t(\mathcal{X}_h, \mathcal{X}_{R_j}, \nu)^{-1} \cdot \Delta a(\theta, \theta_{opt})^{-1} \cdot I_R(R_i|R^0) \tag{4.10}$$

Thus, we have all components to compute the belief about the user's n-th future interaction.

4.4.1 Calculating the Maximum Utility Position for the Robot

By combining the prior knowledge and the prediction, the robot is able to estimate at which locations the user will likely need its help in the future. However, the robot must still decide where it should place itself between these positions. This is a non-trivial task as the prediction is probabilistic and equally likely goal objects may be far apart. It is also not a priori obvious how far the robot should look into the future to make effective predictions. However we can estimate this by evaluating the belief over different values of n, which denotes the n-th future interaction or navigation goal, in regards to the standard deviation of the probability of the different goals. A low standard deviation corresponds to

a scenario in which no strong prediction can be made as the goals are more or less equally likely. Therefore, we are interested in values of n with a relatively high average standard deviation. The results of such an evaluation of the standard deviation for different n in simulated environments (described in detail in Sec. 4.6) are depicted in Fig. 4.2. As can be seen, the average standard deviation stagnates at a low level after $n = 4$. Therefore, only the next four goal activity regions of the human are taken into account to decide on the best robot position.

When required, the robot should provide assistance to the human as soon as possible, i.e. at the lowest value of n at which assistance my be required. Therefore, the robot should prioritize promising goal positions at a low n value over such with a higher n value. To model this the approach uses a probabilistic weight function $w(n)$ for the probability that the human will need the robot's help at the given value of n. This function is defined as the number of significantly likely goal activity regions at which the human would need assistance divided by the number of all significantly likely goal activity regions, both for the given value of n. Here, we define a goal activity region as significantly likely if its goal probability is above the sum of the average goal probability and their standard deviation for the given value of n. In other words, a high value of $w(n)$ corresponds to a high probability that the human will need assistance at the given value of n, whereas a low value corresponds to a low likelihood of this event. Using this function as well as the knowledge at which activity regions the human needs the robot's help, we can compute the best robot position between the possible goal locations weighted by the probability that the human will arrive at them at time step n, using $w(n)$ and the corresponding belief distribution. Let *max* be the maximum value for n (in our case 4), $dist(\mathcal{X}, \mathcal{X}_{R_j})$ be the A* distance between $\mathcal{X}$ and the position $\mathcal{X}_{R_j}$ of the activity region R_j and h be a function that returns 1 for activity regions where the robot should provide assistance and 0 otherwise, based on the prior knowledge. The utility $U^n(\mathcal{X})$ of position $\mathcal{X}$ considering the likely n next activity region interactions is therefore computed as follows:

$$U^n(\mathcal{X}) = w(n) \cdot \sum_{R_j} \frac{bel(R_j^n) \cdot h(R_j)}{dist(\mathcal{X}, \mathcal{X}_{R_j})} + (1\text{-}w(n)) \cdot U^{n+1}(\mathcal{X}) \tag{4.11}$$

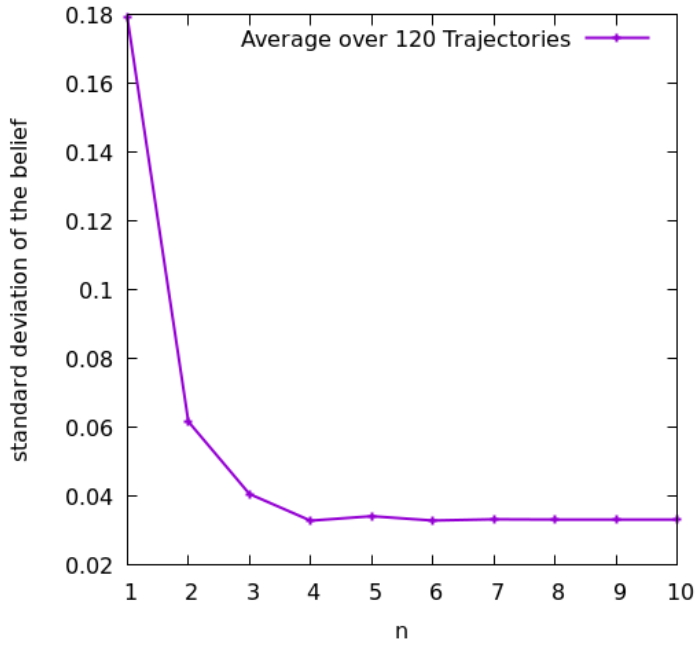


Figure 4.2: Standard deviation of the belief of possible goal locations (corresponding to the goal activity regions) for different values of n. As can be seen, there are no major differences for prediction results with n values greater than 4.

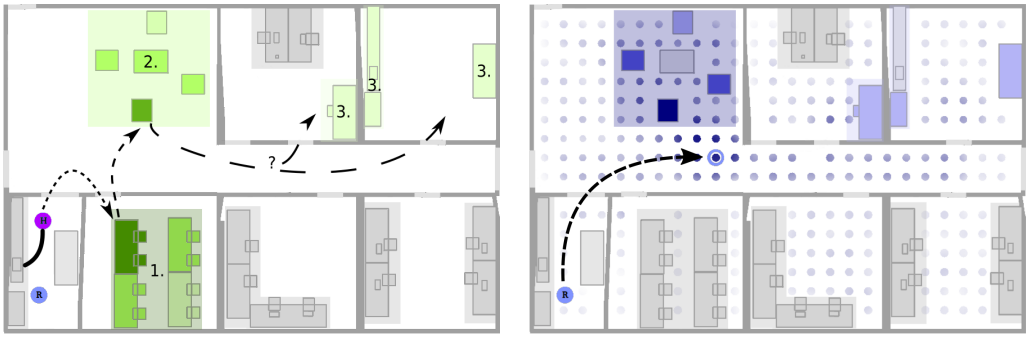
for $0 < n < max$ and else

$$U^{max}(\mathcal{X}) = w(n) \cdot \sum_{R_j} \frac{bel(R_j^{max}) \cdot h(R_j)}{dist(\mathcal{X}, \mathcal{X}_{R_j})} \tag{4.12}$$

Fig. 4.3 shows an example of a computed utility map in a simulated environment.
Once the position $\mathcal{X}_{max}^{n}$ with the highest utility for the n next activity region interactions is determined, the robot uses time-depended path planning to compute a non disturbing path to it. The details of this process are described in the next section.

4.5 Human-Aware Time-Dependent Path Planning

Given the previously discussed prediction of the maximum utility position of the robot in regards to the next destination at which the user will need assistance we next need to find a



(a) (b)

Figure 4.3: (a) Predicted next activity region interactions of the user. The robot (blue) observes the human (violet) interacting with a coffee machine and moving away from it (black line). Using the prediction framework, the robot computes the likely next activity regions with which the human will interact. Likely goals are depicted in green with their n value (corresponding to the n-th activity region interaction), the darker the color the higher the likelihood. The next two activity regions can be predicted with a certain accuracy, the third however is highly unsure as 3 different activity regions are likely. Using the predicted interactions at the different time steps, the system is also able to predict the trajectory of the human (dotted black line). (b) Utility map based on the predictions computed in (a). Activity regions at which the robot can assist the human are blue, as are possible positions. The darker the color the higher the goal belief and utility, respectively. Once the utility map is computed, the robot computes a path to the position with the highest utility (blue circle).

path from the current position of the robot to it. Thereby, the path should minimize the robot's arrival time and comply with the user's comfort. To accomplish this the presented approach applies a two step solution. First, it creates a time-dependent cost-map based on the predicted user positions at future time steps and social constraints, determined as described in Sec. 4.3. Once this map is computed it solves the time-dependent shortest path problem [120] on the given cost map and starts following the returned path. Both steps, as well as the computation of the best robot position, are regularly executed to update the prediction and recalculate the best robot navigation action based on new observations. Fig. 4.4 demonstrates the functioning of the system for an example scenario. In the following section, we discuss both steps in detail.

(a) (b) (c)

Figure 4.4: Example of the presented approach. The human (dark violet) starts their movement after interacting with a sofa. The robot (light blue) observes the interaction and starts to plan its movement with our approach. The computed likelihood of the objects to be the user's navigation goal are shown in green, the darker the color the higher the likelihood. Cost values of possible robot positions are shown in blue, the darker the color the lower the cost. The position with the lowest cost value is shown as blue cross. (a) Based on the initial observation the robot predicts that the human will head to a nearby office and moves accordingly. However, the robot then observes that the human is moving in another direction, concluding that its initial prediction is false. (b) The robot updates the prediction accordingly and computes a new navigation goal and path to it. The robot cannot pass by the user in the corridor as it does not want to enter their SZ or IPS (light violet). As other paths are not available it chooses to follow the user at 5 meter distance. (c) Once the user reaches a wider corridor, the robot is able to pass by the user and reach a position close to their true navigation goal.

4.5.1 Time-Dependent Path Planning

Given the computed utility, the robot needs to determine a path from its current position to the user's predicted navigation goal. The path should minimize the robot's arrival time while complying with the comfort of the user. To calculate the path, the approach uses a cost grid and time-dependent shortest path planning. Given the observed user state S, it assigns costs to each cell X that is not occupied with a static obstacle, by considering the constraints derived from the studies about human comfort introduced in Sec. 4.3. Thereby, it needs to predict the user's path given the belief about their navigation goal. Let us assume that the current observation $S = S^0$ was done at time t_0 and that R_g is the user's

most likely navigation goal according to the observation and the prediction as described in Eq. (4.10). To predict the user's path, the approach assumes that they follow the A* path from their current position $\mathcal{X}_h$ to R_g on the grid map with their current velocity until the next observation takes place. Let us further assume that the user reaches R_g at time step t_f according to the current velocity.

Based on the results of Sec. 4.3, three human comfort constraints emerge: minimizing time inside the SZ, IPS, and the region up to 5 meters behind the human. Further, the survey showed that if they must decide humans prefer non-disturbing robot paths over efficient ones for the robot. It was also observable that humans dislike situations in which the robot follows them. Therefore, the following modeling was chosen: The SZ and IPS are impassable regions for the robot, as by definition humans are disturbed if a robot enters these areas. Positions up to 5 meters behind the human have increased costs, the closer to the human the higher. As a result, the robot only enters this area if no alternative paths are available and only to reach more cost efficient positions, e.g., in front of the human.

Formally, this results in the following time-dependent cost function for the grid cells at time t:

$$C^t(\mathcal{X}) = \begin{cases} \infty \text{ if } \mathcal{X} \in SZ(\mathcal{S}^t) \\ \infty \text{ if } \mathcal{X} \in IPS(\mathcal{S}^t) \\ \frac{5}{dist(\mathcal{X}, \mathcal{X}_h)} \text{ if } \mathcal{X} \in B(\mathcal{S}^t) \\ 1 \text{ else} \end{cases} \tag{4.13}$$

with $SZ(\mathcal{S}^t)$ and $IPS(\mathcal{S}^t)$ as the SZ and IPS at time t, respectively, and $B(\mathcal{S}^t)$ as the backwards extended SZ or back area of the human, which corresponds to a rectangular region with a length of 5 meters, a width of 2.4 meters, and its center 2.5 meters behind the human, orthogonal to their movement direction at time t. This area represents the region in which humans tend to view a robot as a follower. Fig. 4.5 visualizes these areas.

Given the time-dependent cost function C^t, the time-dependent shortest path problem from the position of the robot $\mathcal{X}_r$ to the maximum utility position U^{max} can be solved using A* following the approach of Zhao *et al.* [120], as shown in Alg. 2:

Once the path is computed the robot starts following it and after a fixed time interval performs a new prediction about the user's navigation goal based on a new observation. Afterwards, the best path for the robot is recalculated using an updated cost map. The

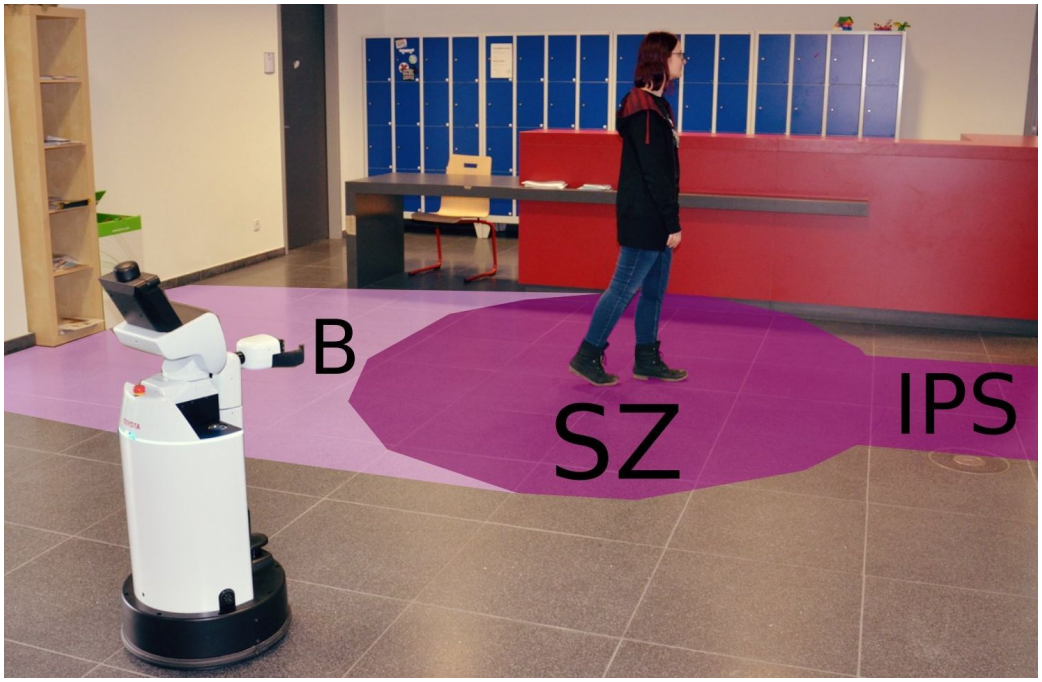


Figure 4.5: Visualization of the regions around the human that the robot should not enter. The social zone (SZ), with a radius of 1.2 meters, and the information process space (IPS), with a length of 4 meters and a width of 1 meter, are impassable regions depicted in dark violet. The area behind the human (B), with a length of 5 meters and a width of 2.4 meters depicted in light violet, can be entered by the robot but has increased costs since a robot in this region would likely be perceived as a follower.

process is repeated until the user has reached their destination.
A high level example of this process is depicted in Fig. 4.6.

4.6 Experimental Evaluation

The approach was evaluated with a quantitative evaluation in simulation with 140 trajectories, a real-world experiment in a lab with 11 participants and in a virtual reality experiment with 20 participants. For the quantitative evaluation, *human comfort* (HC) and the *difference in arrival time* between the robot and the user (ΔT) at a destination at which the user needed assistance were used as metrics. Positive ΔT values indicate that

Algorithm 2 : Pseudocode for the time-dependent shortest path problem A* algorithm, based on the approach from Zhao *et al.* [120].

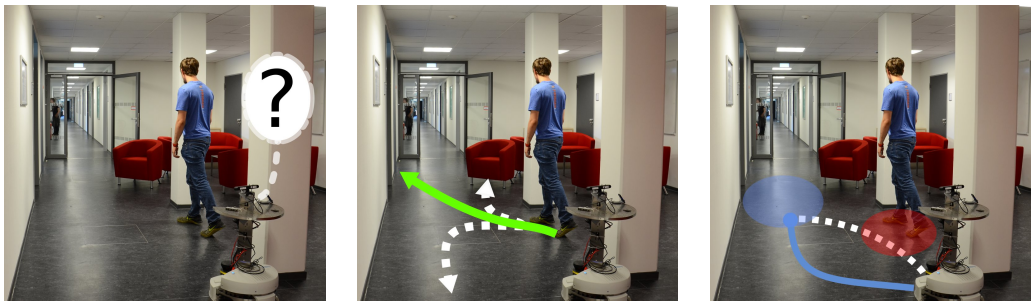
Input : grid map M, nearest cell to the robot at time t_0: *s*, maximum utility position and goal position of the robot: d

Output : g(d) and the s-d path found (i.e. the reverse of d, p(d), p(p(d)), ..., s)

1 status(s) := labeled, g(s) := t_0, status(v) := unlabeled for all v ≠ s, g(v) := ∞ for all v ≠ s, p(v) := v for all v.
2 Let v be a labeled cell of M with the smallest g(v) + h(v,g(v)) and h(v,g(v))=$\text{dist}_2^{g(v)}$(v,d) as the euclidean distance between the cells v and d at time g(v). In the case that there are multiple candidates, choose one with the smallest g(v).
3 **if** $v = d$ **then**
4 | GOTO 14
5 **end**
6 **for** *all non occupied neighbor cells w of v* **do**
7 | **if** *status(w) is unlabeled* **then**
8 | | status(w) := labeled, g(w) := g(v) + $C^{g(v)}$(w), p(w) := v
9 | **else if** *status(w) is labeled AND* $g(w) > g(v) + C^{g(v)}(w)$ **then**
10 | | g(w) := g(v) + $C^{g(v)}$(w), p(w)=v
11 | **end**
12 **end**
13 status(v) := finished. GOTO 2
14 Return results.

the robot arrives at the goal before the user. The higher this value the earlier the robot will be at the true navigation goal of the human. To quantify HC, the ratio of the number of robot positions outside of the SZ or IPS and the robot's overall number of positions as *social distance compliance* (SDC) as well as the average *human-robot distance* (HRD) were measured. Based on the surveys, an optimal human-aware robot path has an SDC of 1.0 and a high average distance. During the real-world experiment, participants were asked to rate their comfort with the robot navigation behavior on a 5-point Likert scale rating from very uncomfortable (1) to very comfortable (5).

The interaction model used in this evaluation was prelearned based on 195 human-object interactions [91]. 17 different object classes were used which were assigned to four activity classes to reduce complexity and allow a general classification of whether help



(a) (b) (c)

Figure 4.6: Example of the navigation policy. (a) The robot observes the moving user. In order to be of assistance it needs to predict were the human is heading and compute a good position for itself. (b) Using our prediction system, the robot computes possible paths for the human (white) and predicts the most likely one (green). (c) Based on this prediction, the robot plans a time-dependent shortest path (blue) free of interference (red) to the position with the highest utility, which is close to the user's true navigation goal if the prediction is correct.

can be provided at an object: resting *(bed, wardrobe, sofa)*, food processing (*bottles, cups, microwaves, workbenches, refrigerators, coffee machines*), office work (*chairs, tables, laptops, whiteboards, cupboards*), and hygiene (*toilets, washbasins, bathtubs*). The grouping was done based on functionality of the objects. To infer at which objects the robot should provide assistance eight interviews with students from the University of Bonn about preferable service robot behavior were performed. In this context, the participants were asked at which activities they would like the robot to provide assistance. The results showed that the robot should provide assistance for the user during resting, e.g., by providing fetch tasks for the user, and food processing activities. Therefore, for the purpose of this evaluation it is assumed that the user will need assistance only during resting and food processing activities.

In all experiments, the results were compared with three existing approaches: The social force approach by Ferrer *et al.* [73], which applies a short-term prediction system, the reinforcement learning approach by Bayoumi *et al.* [101], which uses long-term prediction but no time-dependent path planning, and the follower approach by Tee *et al.* [100], which does not use a prediction system.

4.6.1 Quantitative Evaluation

For the quantitative evaluation, five different simulated office and home environments with sizes between $100\,m^2$ and $150\,m^2$, a grid resolution of 0.25 meter and up to 110 different objects from 15 different classes were created using the V-REP editor [94]. A set of 140 test trajectories were randomly sampled over all environments and based on a training set of 128 previously recorded human object interactions. The same set of test trajectories was used for all evaluations.

Tab. 4.1 shows the evaluation results. As can be seen, the presented approach achieves the highest HRD and ΔT. It thereby significantly outperforms all other approaches based on a paired t-test with $\alpha = 0.05$. Further, it shows a near optimal SDC with 0.99 similar to the social force approach [73], again it significantly outperforms the reinforcement learning approach [101], as well as the non-predictive approach [100]. Social distance violations were mostly encountered when the user moved unexpectedly while the robot was trying to pass them. Typically, the robot would pass by the user as early as possible and wait at key points of the map to update its prediction. In contrast to that, the reinforcement learning approach [101] focuses on reaching the most likely goal position as fast as possible. Early predictions tend to be false, which resulted in unnecessary movement of the robot and some passes through the SZ and IPS. When using the social force approach [73], the robot followed the user outside of their SZ. However, while the robot was able to predict the user's short-term movements and anticipate and avoid situation in which it would enter the SZ, it was not able to reach the goal activity region before the user nor hold a high distance to the user. The short HRD from less than five meters and the fact that the robot could not predict long-term goals resulted in the behavior of closely following the user. The non-predictive approach [100] performed similarly. However, as no prediction system was used the robot was not able to anticipate changes in the user's movement pattern and many intrusions inside the SZ happened.

4.6.2 Real-World Experiments

For the qualitative evaluation, a real-world user experiment with 11 student participants from different departments of the University of Bonn was designed. The task of the participants was to follow a specific trajectory, while a robot would predict and navigate to

	Avg. SDC	Std. SDC	Avg. HRD	Std. HRD	Avg. ΔT	Std. ΔT
Presented approach	**0.99**	0.01	**10.5m**	2.5m	**44.2s**	14.2s
Social force approach [73]	**0.99**	0.01	2.5m	0.4m	-4.5s	1.7s
Reinforcement learning approach [101]	0.87	0.04	4.2m	0.8m	9.5s	2.8s
Non-predictive approach [100]	0.50	0.05	2.0m	0.4m	-9.0s	1.4s

Table 4.1: Results of the quantitative evaluation with 140 trajectories in five different simulated environments. As can be seen, the presented approach achieves the highest average human-robot-distance (HRD) difference in arrival time (ΔT), while simultaneously achieving a near optimal social distance compliance (SDC). According to the paired t-test with $\alpha = 0.05$ the presented approach outperforms all other approaches in regards to HRD and ΔT, further it outperforms both the reinforcement learning as well as the non-predictive approach in regards to SDC.

their movement goal using our approach as well as the three different methods introduced above. After the experiment, the participants had to rate their feeling of comfort for each of these approaches on a Likert scale with five values, ranging from 1 to 5: *very uncomfortable* (VU), *uncomfortable* (U), *neutral* (N), *comfortable* (C), *very comfortable* (VC). Tab. 4.2 depicts the results of the qualitative evaluation. The same trend as in the preceding surveys can be observed. Participants felt uncomfortable to very uncomfortable if a robot passed through their SZ and or IPS (reinforcement learning, non-predictive) and comfortable to very comfortable if a robot would not enter these areas (presented approach, social force). Participants also particularly disliked if the robot would just follow them (non-predictive). The results further showed that participants felt more comfortable if they thought that the robot had a policy to actively avoid them (presented approach, social force). As these results demonstrate, the navigation strategy produced by the presented approach achieves the highest rating and was on average seen as comfortable.

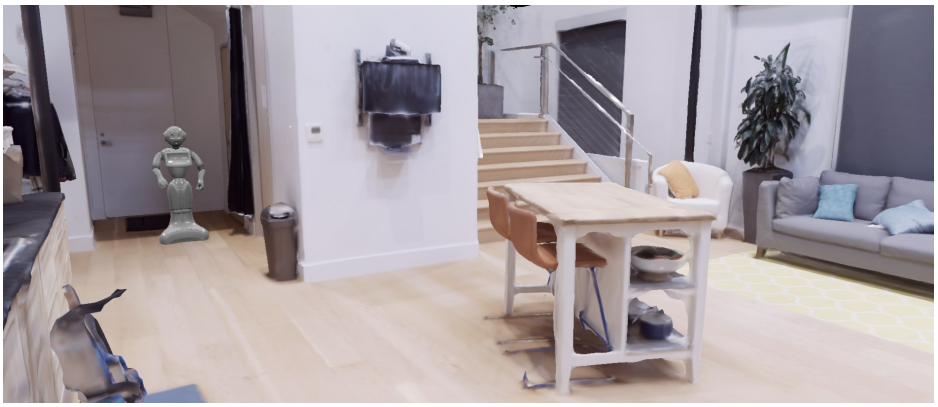
4.6.3 Evaluation in a Virtual Reality Setting

To further evaluate how humans rate the robot behavior generated by the presented approach in a home environment, a virtual reality experiment with 20 student participants

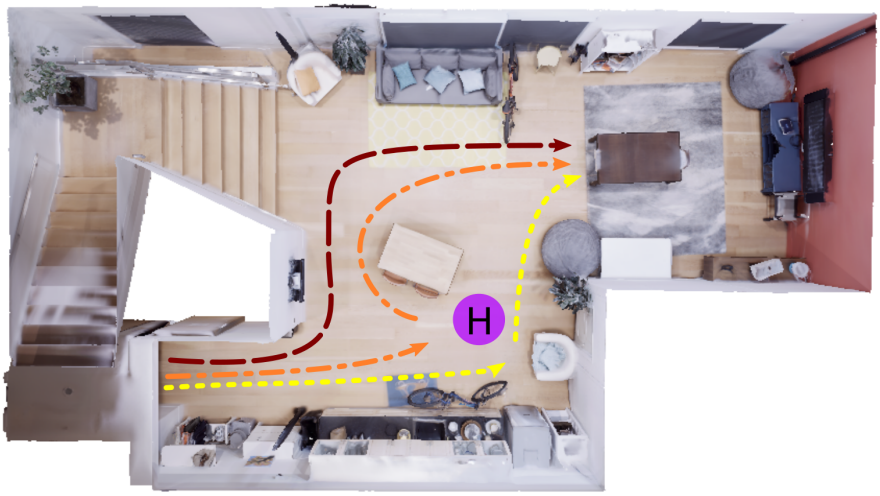
	VC (5)	C (4)	N (3)	U (2)	VU (1)	Avg. value	Std. value
Presented approach	27%	**63%**	0%	9%	0%	**4.1**	0.8
Social force approach [73]	18%	27%	**36%**	18%	0%	3.5	1.0
Reinforcement learning approach [101]	9%	9%	18%	**45%**	18%	2.5	1.6
Non-predictive approach [100]	0%	0%	9%	36%	**54%**	1.5	0.7

Table 4.2: Results of the qualitative evaluation with 11 participants. Participants rated the robot's behavior on a Likert scale with five values, ranging from 1 to 5: very uncomfortable (VU), uncomfortable (U), neutral (N), comfortable (C), very comfortable (VC). As can be seen, our approach achieves the highest rating and was on average seen as comfortable (4.1). A more detailed discussion of the results are given in the text.

from the University of Bonn was conducted, using a HTC Vive. For this experiment an environment from the Facebook replica dataset [121] and a model of the Pepper robot were used. The position of the participants was in front of a table. They were instructed to observe the behavior of the robot while staying at their position. The robot was given the information that the user did not need help at their current destination and would next interact with a table, where assistance was likely required. Each participant was shown three different robot trajectories in the VR environment, one after the other, as shown in Fig. 4.7. First, the presented approach (dark red line) was applied, here the robot is aware of the user's intentions and moves foresightedly. Accordingly, the robot predicts that the user does not need help at their current position and moves close to the likely next point where the human is expected to need help. With the approach presented in [67], the robot does not know where the user might need its help and checks if they require assistance at the predicted immediately next navigation goal by moving to this position. As the robot is not called to help the user, it realizes that the human does not need assistance at this location and moves to the predicted next navigation goal based on the observed object interaction (orange line). The last trajectory (yellow line) results from the follower approach [100]. Here, the robot moves to the position of the user. However, in contrast to the previous approaches, the robot does not move in advance to the user's predicted next



(a)



(b)

Figure 4.7: (a) view of the user in the virtual reality environment to evaluate different robot navigation behaviors. (b) bird's eye view of the complete environment. The non-moving participant (violet circle) observed three different trajectories, corresponding to the robot behavior resulting from the presented approach (dark red trajectory), the prediction approach [67] (orange trajectory), and the follower approach [100] (yellow trajectory).

navigation goal but stays close to the human. The reinforcement learning approach [101] was not used in this experiment as no training trajectories for this map were available.

	pleasant	neutral	unpleasant
Presented approach	**0.85**	0.15	0.00
Predictive approach [67]	0.05	0.25	**0.70**
Follower approach [100]	0.10	**0.50**	0.40

Table 4.3: Results of the virtual reality experiment evaluating the robotic navigation behavior of three different approaches. Each navigation behavior could be rated as *pleasant*, *neutral*, or *unpleasant*. In general, participants rated the presented approach as pleasant, while the approach [67], which only predicts the next navigation goal, was mostly rated as unpleasant for situations in which the robot's help is not needed. The follower approach [100] was rated as neutral but too close to the user and, therefore, unpleasant.

After all scenarios, the participants were asked to rate the different robot navigation behaviors based on their satisfaction with them. The results are depicted in Tab. 4.3. As can be seen, 85% of the participants rated the presented approach as pleasant, while 15% rated it as neutral. The behavior of the robot was described as clear and predictable, participants were also pleased with the robot's distance to them. The second approach [67], was rated as unpleasant by 70%, while 25% rated it as neutral, and only 5% as pleasant. Participants were unsure about the robot's intentions, as the robot first moved towards them and then moved away from them. Some were even annoyed by this behavior. The follower approach was rated as unpleasant by 40%, as neutral by 50%, and as pleasant by 10%. The biggest critique of the users was that the robot came too close to them. However, most of the participants also noted that the robot acted the way they expected it to act. These results support the assumption that users feel more comfortable if the robot avoids unnecessary interference or coming too close to them. The results also show that users highly appreciate predictability in the robot's behavior, as the most unpleasant approach was mainly criticized for its unpredictable navigation behavior and the third, while unpleasantly close to the user, praised for its simple navigation strategy.

4.7 Discussion

Based on the results of the evaluation the presented approach is generally seen favorably by users, especially in contrast to the compared non human-aware reinforcment learning

[101] and follower [100] approaches. Furthermore, by taking into account at which objects the user will need help and, perhaps more importantly, at which objects they will not, the robot can significantly decrease both its overall movement as well as the time it needs to arrive at destinations at which the user needs assistance. As the system depends on the prediction, a loss of observation data or false positive observations present a threat to the overall performance of the whole approach. However, this can be countered by providing additional means of observation, e.g. utilizing the capabilities of a smart home environment [115, 114, 116] or an RFID system for non visual interaction detection [42]. An interesting observation during the evaluation was that participants disliked it if they were not able to localize the robot. This could present a problem for the approach in environments with multiple independent paths to a possible goal destination, as the presented approach tries to minimize unnecessary contact with the user. This problem could again be solved by incorporating the robot into a smart home system, thereby allowing the system to localize it if desired by the user, or by using periodically acoustic signals. All in all the results support the initial hypothesis that by combining long term movement predictions with a human-aware navigation framework we can both increase the comfort of the user as well as minimize the travel distance of the robot and the time it takes until it can provide assistance to the user. Furthermore, the main weak points of the system can be countered by incorporating it into a smart home system, which can be realistically expected to be available in an household that also uses a service robot.

4.8 Conclusion

In this chapter, we discussed a novel approach to human-aware navigation for assistant robots in indoor environments. This approach is based on long-term prediction of human motions, utilizing prior knowledge about human object transitions, in combination with time-dependent path planning under consideration of human comfort constraints. It further uses a utility map to weight likely future goal positions at which the human may need help. Using this map, it finds the robot position which simultaneously minimizes the time the robot needs to provide assistance if needed, as well as unnecessary movement and disturbances of other activities of the user.
As demonstrated in a quantitative evaluation, the approach achieves a high average distance

between the robot and the human and avoids interpersonal space breaches, which are seen as uncomfortable by humans. At the same time, the framework leads to efficient navigation behavior, so that the robot is often able to arrive before the user at positions at which assistance may be required. Furthermore, the qualitative evaluation with a real robot in a lab enviroment as well a virtual reality experiment shows that humans rate the robot behavior as comfortable. Using direct human feedback as metric, the approach also outperforms existing methods based on social forces [73], reinforcement learning [101], and non-predictive following [100].

CHAPTER 5

Summary

In this thesis we discussed an approach to implement foresighted human-aware robot navigation based on predicted object interaction. For this we considered three key contributions:

1. An approach to robustly detect and extract human-object interaction in video streams based on spatio-temporal and pose information over subsequent frames.

2. A Bayesian inference framework based on transition probabilities between objects to predict the navigation goal of a moving human.

3. A time-dependent navigation policy to compute a human-aware path for a service robot to the next destination at which the user will need assistance.

Each of those contributions was discussed in a separate chapter. In the first of those we looked at a human-object interaction detector for video streams, representing the first key contribution of this work. The detector uses an R-CNN to detect objects, a pose estimator to detect humans and their poses and an RBG-D camera to infer depth values. Based on the collected information, it finds human-object interactions on the current frame by looking for objects which the user simultaneously faces and touches. Each found interaction is then verified by observing whether an interaction with the same object also

occurs on the subsequent frames. Based on the duration of the interaction, i.e. the inferred amount of seconds from the observed number of frames with the interaction and the known number of frames per second, a likelihood value is computed. Finally, interactions found in succeeding frames for a minimal duration of 3 seconds with a likelihood value over a predefined threshold are returned or otherwise discarded. As shown by an experimental evaluation, the approach is able to robustly detect human-object interactions.

This ability opened the path to the next chapter, in which we discussed the second key contribution. The overall goal of this chapter was to find a way to predict the navigation goal of a moving human in an indoor environment by utilizing knowledge about human-object interaction sequences. The presented approach uses knowledge about typical human-object interaction sequences, learned from training data, and groups close-by objects into activity regions to achieve a better generalization. The system combines the learned object transition probabilities with observations about the human's movement and pose to infer the navigation goal using Bayesian inference. As demonstrated in various experiments, the framework reliably predicts the navigation goal of a moving human. By using transitions between the activity regions in contrast to single object transitions, it achieves a significant increase in the prediction accuracy. Furthermore, the system outperforms a trajectory-based prediction approach that relies on previously learned trajectories between fixed destinations.

Similarly to the interaction detector the prediction system served as foundation for the subsequent chapter. In this case the chapter is concerned with the third key contribution regarding human-aware navigation based on long-term movement predictions. The approach presented in this chapter utilizes prior knowledge about human object transitions, in combination with time-dependent path planning under consideration of human comfort constraints. It further uses a utility map to weight likely future goal positions at which the human may need help. Using this map, it finds the robot position which simultaneously minimizes the time the robot needs to provide assistance if needed, as well as unnecessary movement and disturbances of other activities of the user. As demonstrated in a quantitative evaluation, the approach achieves a high average distance between the robot and the human and avoids interpersonal space breaches, which are seen as uncomfortable by humans. At the same time, it leads to efficient navigation behavior, so that the robot is often able to arrive before the user at positions at which assistance may be required.

Furthermore, the qualitative evaluation with a real robot in a lab enviroment as well as a virtual reality experiment shows that humans rate the robot behavior as comfortable. Using direct human feedback as metric, the approach also outperforms existing methods based on social forces [73], reinforcement learning [101], and non-predictive following [100]. This navigation approach also fulfills our overall goal of finding a foresighted human-aware robot navigation approach based on predicted object interaction. The results highlight that this navigation approach can outperform state-of-the-art literature approaches both in the path efficiency of the robot as well as in the comfort of the human. The main downside of it is its dependency on each of its individual parts, e.g. if the object detection does not work correctly, the precision of the prediction will suffer, which in turn would lead to sub-optimal navigation. However, the results of the various evaluations also showcase that even in such scenarios the approaches still perform adequate. In this specific scenario the performance would linearly worsen with the amount of false positive or negative object detections. This weakness can be turned into a strength, as it is easily possible to improve individual parts of the system without changing others, thanks to its modular nature.
All in all the approach showed the usefulness of long-term planning and human-aware movement and highlights that path efficiency and human comfort can go hand in hand. It also showcases the knowledge that can be extracted from object interaction sequences and with it a new source of observation data for long-term indoor movement prediction.

5.1 Outlook

The system can be improved in various ways, especially as new technological possibilities are emerging. A very useful extension would be the integration of the approach into a smart home system. This would potentially eliminate the need of self gathered observation data, as these would be automatically provided by the smart home environment. As it is likely that a smart home system would be present in a household which also employs a robotic assistance this could be an easily achievable constraint. This system could also gather training data of object interaction sequences, possibly individually tailored to individual users. This would likely increase the quality of the used interaction models, as a lot of data would be available for the specific workspace of the robot. Thereby the complexity of the system would be decreased, as the detection component could be outsourced to the

smart home system. At the same time the quality of the used data would be increased with a better interaction model and more available sensors from the smart home system. The prediction framework could also benefit from such a system, as it allows a continuously updated map. Therefore, it would be easily possible to identify objects which were moved around and update the map in real time, even if the robot is not in the same room. Thus, the robot would never have to deal with outdated activity region positions. Regarding the navigation approach, more studies are needed to specify additional possible navigation constraints for the robot and in turn further increasing human comfort. Also the already concluded studies show that users dislike irrational robot behavior, which could e.g. occur if the robot wrongly predicts a navigation goal and therefore moves to a position at which the user does not expect it. Some kind of validation function would be useful here to check whether a wrong prediction could disturb the user and, if so, a way to compute an alternative non-disturbing path to the goal.
Another possible future work point would be the design of the robot, as it is likely that this is also a factor for the comfort of the user. As the design is also constrained by the functions, work environment and costs of the robot, it is therefore likely a difficult task to find a both visually pleasing and efficient robot design. However, with the current speed of the ongoing digital revolution such research is likely needed sooner rather than later.

List of Figures

List of Tables

Liste der Algorithmen

Bibliography

[1] Kira Bungert, Lilli Bruckschen, Kathrin Müller, and Maren Bennewitz. Robots in education: Influence on learning experience and design considerations. In *Proceedings of the European Conference on Education*. IAFOR, 2020.

[2] Joseph L Jones. Robots at the tipping point: the road to irobot roomba. *IEEE Robotics & Automation Magazine*, 13(1):76–78, 2006.

[3] Amit Kumar Pandey and Rodolphe Gelin. A mass-produced sociable humanoid robot: Pepper: The first machine of its kind. *IEEE Robotics & Automation Magazine*, 25(3):40–48, 2018.

[4] C Bronk Ramsey, Michael W Dee, Joanne M Rowland, Thomas FG Higham, Stephen A Harris, Fiona Brock, Anita Quiles, Eva M Wild, Ezra S Marcus, and Andrew J Shortland. Radiocarbon-based chronology for dynastic egypt. *science*, 328(5985):1554–1557, 2010.

[5] Ian Shaw. *The Oxford history of ancient Egypt*. Oxford University Press, 2003.

[6] Gaston Maspero. *Manual of Egyptian Archaeology and Guide to the Study of Antiquities in Egypt*. Grevel, 1895.

[7] Alan David Melville and Edward John Kenney. *Metamorphoses*. Oxford Paperbacks, 1998.

[8] Deborah Levine Gera. *Ancient Greek ideas on speech, language, and civilization*. Oxford University Press, USA, 2003.

[9] Andy Orchard. *Cassell dictionary of Norse myth and legend*. Cassell, 1998.

[10] Moshe Idel. Golem: Jewish magical and mystical traditions on the artificial anthropoid. 1990.

[11] Ronnie Littlejohn and Jeffrey Dippmann. *Riding the Wind with Liezi: New Perspectives on the Daoist Classic*. SUNY Press, 2011.

[12] FM Martín Sánchez, F Rodríguez Millán, J Bayarri Salvador, J Redorta Palou, F Escovar Rodríguez, S Fernández Esquena, and H Mavrich Villavicencio. History of robotics: from archytas of tarentum until da vinci robot.(part i). *Actas urologicas espanolas*, 31(2):69–76, 2007.

[13] Gaby Wood. *Living dolls: a magical history of the quest for mechanical life*. Faber & Faber, 2002.

[14] Shimon Y Nof. *Handbook of industrial robotics*. John Wiley & Sons, 1999.

[15] Paul Mickle. 1961: A peep into the automated future. *The Trentonian. Accessed August*, 11, 2011.

[16] Martin Hilbert. Information quantity. *Encyclopedia of big data*, pages 1–4, 2017.

[17] David Reinsel, John Gantz, and John Rydning. The Digitization of the World – From Edge to Core. *International Data Corporation (IDC) White Paper*, 2018.

[18] Irena Bojanova. The Digital Revolution: What's on the Horizon? *IT Professional*, 16(1):8–12, 2014.

[19] Min Xu, Jeanne M David, Suk Hi Kim, et al. The fourth industrial revolution: opportunities and challenges. *International journal of financial research*, 9(2):90–95, 2018.

[20] Guoping Li, Yun Hou, and Aizhi Wu. Fourth industrial revolution: technological drivers, impacts and coping methods. *Chinese Geographical Science*, 27(4):626–637, 2017.

[21] Klaus Schwab. *The Fourth Industrial Revolution*. Currency, 2017.

[22] Georges Akhras. Smart materials and smart systems for the future. *Canadian Military Journal*, 1(3):25–31, 2000.

[23] Mike Elgan. How iphone changed the world. *Cult of Mac*, 2, 2011.

[24] Michael Caccavale. The impact of the digital revolution on the smart home industry. *Forbes. Fecha de consulta:(Mayo, 2019). Recuperado de: https://www. forbes. com/sites/forbesagencycouncil/2018/09/24/the-impact-ofthe-digital-revolution-on-the-smart-home-industry/# cb205093c76d*, 2018.

[25] Dave Evans. The Internet of Things How the Next Evolution of the Internet Is Changing Everything. *Cisco Internet Business Solutions Group (IBSG) White Paper*, 2011.

[26] Felix Wortmann and Kristina Flüchter. Internet of things. *Business & Information Systems Engineering*, 57(3):221–224, 2015.

[27] Kevin Ashton et al. That 'Internet of Things' Thing. *RFID journal*, 22(7):97–114, 2009.

[28] Apeksha Kaushik, Peter Havart-Simkin, Kay Sharpington, Matt Arnott, Eric Goodness, Alfonso Velosa, and Peter Middleton. Scenarios for the IoT Marketplace, 2019. *Gartner Research*, 2019.

[29] Internet World Stats. World Internet Users and 2020 Population Stats. https://www.internetworldstats.com/stats.htm. Accessed 2020-07-05.

[30] United Nations. 2019 Revision of World Population Prospects. https://population.un.org/wpp/Graphs/Probabilistic/POP/TOT/900. Accessed 2020-07-05.

[31] Paolo Magrassi. Why a Universal RFID Infrastructure Would Be a Good Thing. *Gartner Research*, 2002.

[32] Pamela Cohn, Alastair Green, Meredith Langstaff, and Melanie Roller. Commercial drones are here: The future of unmanned aerial systems. *McKinsey & Company*, 2017.

[33] Martin Hagele. Robots conquer the world [turning point]. *IEEE Robotics & Automation Magazine*, 23(1):120–118, 2016.

[34] Thibault Kruse, Amit Kumar Pandey, Rachid Alami, and Alexandra Kirsch. Human-aware robot navigation: A survey. *Robotics and Autonomous Systems*, 61(12), 2013.

[35] Parmy Olson. Softbank's robotics business prepares to scale up, 2018.

[36] H Richardson. Robots could help solve social care crisis, say academics. *Education & Family: BBC News*, 2017.

[37] Bots.co.uk. Pepper robot price. https://bots.co.uk/pepper-robot-price/, 2020. last visited 2020-07-13.

[38] Markets and Markets. Service Robotics Market by Environment - Global Forecast to 2025. https://www.marketsandmarkets.com/Market-Reports/service-robotics-market-681.html, 2020. last visited 2020-07-13.

[39] Patrick Holthaus, Catherine Menon, and Farshid Amirabdollahian. How a robot's social credibility affects safety performance. In *International Conference on Social Robotics*, pages 740–749. Springer, 2019.

[40] Cynthia Breazeal, Cory D Kidd, Andrea Lockerd Thomaz, Guy Hoffman, and Matt Berlin. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *2005 IEEE/RSJ international conference on intelligent robots and systems*, pages 708–713. IEEE, 2005.

[41] Lilli Bruckschen, Sabrina Amft, Julian Tanke, Jürgen Gall, and Maren Bennewitz. Detection of Generic Human-Object Interactions in Video Streams. In *Proc. of the International Conference on Social Robotics (ICSR)*, pages 108 – 118, 2019.

[42] Hanchuan Li, Can Ye, and Alanson P Sample. IDSense: A human object interaction detection system based on passive uhf rfid. In *Proc. of the ACM Conf. on Human Factors in Computing Systems*. ACM, 2015.

[43] Bangpeng Yao and Li Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2010.

[44] Chule Yang, Yijie Zeng, Yufeng Yue, Prarinya Siritanawan, Jun Zhang, and Danwei Wang. Knowledge-based role recognition by using human-object interaction and spatio-temporal analysis. In *Proc. of IEEE Int. Conf. on Robotics and Biomimetics (ROBIO)*, 2017.

[45] Alessandro Prest, Vittorio Ferrari, and Cordelia Schmid. Explicit modeling of human-object interactions in realistic videos. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(4), 2013.

[46] Anna Senina, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. Coherent multi-sentence video description with variable level of detail. *CoRR*, abs/1403.6173, 2014.

[47] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*, 2018.

[48] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013.

[49] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[50] Alessandro Prest, Cordelia Schmid, and Vittorio Ferrari. Weakly supervised

learning of interactions between humans and objects. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(3), 2012.

[51] Abhinav Gupta, Aniruddha Kembhavi, and Larry S Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10), 2009.

[52] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[53] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. *CoRR*, abs/1501.02530, 2015.

[54] Limin Wang, Yu Qiao, and Xiaoou Tang. Video action detection with relational dynamic-poselets. In *ECCV*, 2014.

[55] H. Wang, A. Kläser, C. Schmid, and C. L. Liu. Action recognition by dense trajectories. In *CVPR 2011*, 2011.

[56] Hasim Sak, Andrew W. Senior, and Françoise Beaufays. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *CoRR*, abs/1402.1128, 2014.

[57] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the 2013 IEEE International Conference on Computer Vision*, ICCV '13. IEEE Computer Society, 2013.

[58] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from RGB-D videos. *Intl. Journal of Robotics Research (IJRR)*, 32(8), 2013.

[59] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, and Kevin Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. *CoRR*, abs/1611.10012, 2016.

[60] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(12), 2018.

[61] Abhijat Biswas, Henny Admoni, and Aaron Steinfeld. Fast on-board 3d torso pose recovery and forecasting. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, 2019.

[62] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018.

[63] Weisstein. Gamma function. http://mathworld.wolfram.com/GammaFunction.html. Accessed 2019-02-24.

[64] Weisstein. Gamma distribution. http://mathworld.wolfram.com/GammaDistribution.html. Accessed 2019-02-17.

[65] Weisstein. Incomplete gamma function. http://mathworld.wolfram.com/IncompleteGammaFunction.html. Accessed 2019-02-24.

[66] A. Bayoumi, P. Karkowski, and M. Bennewitz. Learning foresighted people following under occlusions. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 6319 – 6324, 2017.

[67] Lilli Bruckschen, Nils Dengler, and Maren Bennewitz. Human motion prediction based on object interactions. In *Proc. of the Europ. Conf. on Mobile Robotics (ECMR)*, pages 1 – 6, 2019.

[68] Lilli Bruckschen, Kira Bungert, Nils Dengler, and Maren Bennewitz. Predicting human navigation goals based on bayesian inference and activity regions. *Robotics and Autonomous Systems*, forthcoming 2020.

[69] 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/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 3931 – 3936, 2009.

[70] Henrik Kretzschmar, Markus Spies, Christoph Sprunk, and Wolfram Burgard. Socially compliant mobile robot navigation via inverse reinforcement learning. *The International Journal of Robotics Research*, 35(11), 2016.

[71] A. Bayoumi, P. Karkowski, and M. Bennewitz. Speeding up person finding using hidden Markov models. *Robotics and Autonomous Systems*, 115:40 – 48, 2019.

[72] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social LSTM: Human trajectory prediction in crowded spaces. In *Proc. of the IEEE Intl. Conf. on Computer Vision (ICCV)*, pages 961 – 971, 2016.

[73] Gonzalo Ferrer, Anaís Garrell Zulueta, Fernando Herrero Cotarelo, and Alberto Sanfeliu. Robot social-aware navigation framework to accompany people walking side-by-side. *Autonomous Robots*, 41(4):775 —- 793, 2017.

[74] Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M Kitani, Dariu M Gavrila, and Kai O Arras. Human motion trajectory prediction: A survey, 2019. arXiv preprint arXiv:1905.06113.

[75] Dizan Vasquez. Novel planning-based algorithms for human motion prediction. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, pages 3317 – 3322, 2016.

[76] Sujeong Kim, Stephen J Guy, Wenxi Liu, Rynson WH Lau, Ming C Lin, and Dinesh Manocha. Predicting pedestrian trajectories using velocity-space reasoning. In *Algorithmic Foundations of Robotics X*, pages 609 – 623. Springer, 2013.

[77] Stéphanie Lefèvre, Dizan Vasquez, and Christian Laugier. A survey on motion prediction and risk assessment for intelligent vehicles. *Robomech Journal*, 1(1):1 – 14, 2014.

[78] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In

Proc. of the Europ. Conf. on Computer Vision (ECCV), pages 549 – 565. Springer, 2016.

[79] Mark Pfeiffer, Giuseppe Paolo, Hannes Sommer, Juan Nieto, Rol Siegwart, and Cesar Cadena. A data-driven model for interaction-aware pedestrian motion prediction in object cluttered environments. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, pages 5921 – 5928, 2018.

[80] Jared Glover, Sebastian Thrun, and Judith T Matthews. Learning user models of mobility-related activities through instrumented walking aids. In *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*, volume 4, pages 3306 – 3312. IEEE, 2004.

[81] Graeme Best and Robert Fitch. Bayesian intention inference for trajectory prediction with an unknown goal destination. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5817 – 5823. IEEE, 2015.

[82] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282 – 4286, 1995.

[83] Andrey Rudenko, Luigi Palmieri, Achim J Lilienthal, and Kai O Arras. Human motion prediction under social grouping constraints. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3358 – 3364, 2018.

[84] Hakan Karaoğuz, Nils Bore, John Folkesson, and Patric Jensfelt. Human-centric partitioning of the environment. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 844 – 850. IEEE, 2017.

[85] Tian Lan, Tsung-Chuan Chen, and Silvio Savarese. A hierarchical representation for future action prediction. In *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, pages 689 – 704, 2014.

[86] Paul Duckworth, David C Hogg, and Anthony G Cohn. Unsupervised human activity analysis for intelligent mobile robots. *Artificial Intelligence*, 270:291 —-294, 2019.

[87] Qualtrics. Qualtrics software. https://www.qualtrics.com, 2020. last visited 2020-05-04.

[88] Clickworker GmbH. Clickworker software. https://www.clickworker.de/, 2020. last visited 2020-05-04.

[89] Clickworker GmbH. Clickworker user base. https://www.clickworker.com/about-us/clickworker-crowd/, 2020. last visited 2020-05-07.

[90] Tobias Zaenker, Francesco Verdoja, and Ville Kyrki. Hypermap mapping framework and its application to autonomous semantic exploration, 2019. arXiv preprint arXiv:1909.09526.

[91] Julian Tanke, Oh-Hun Kwon, Patrick Stotko, Radu Alexandru Rosu, Michael Weinmann, Hassan Errami, Sven Behnke, Maren Bennewitz, Reinhard Klein, Andreas Weber, et al. Bonn activity maps: Dataset description. *arXiv preprint arXiv:1912.06354*, 2019.

[92] Roland Siegwart, Illah Reza Nourbakhsh, and Davide Scaramuzza. *Introduction to autonomous mobile robots*. MIT press, 2011.

[93] Giorgio Grisetti, Cyrill Stachniss, Wolfram Burgard, et al. Improved techniques for grid mapping with Rao-Blackwellized particle filters. *IEEE Trans. on Robotics (TRO)*, 23(1):34 – 46, 2007.

[94] M. Freese E. Rohmer, S. P. N. Singh. V-Rep: A versatile and scalable robot simulation framework. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 1321 – 1326, 2013.

[95] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al.

Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github. com/openimages*, 2(3), 2017.

[96] Festo Didactic GmbH & Co. KG. Robotino manual. https://www.festo-didactic.com, 2020. last visited 2020-05-04.

[97] Lilli Bruckschen, Kira Bungert, Nils Dengler, and Maren Bennewitz. Human-aware robot navigation by long-term movement prediction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020.

[98] Lilli Bruckschen, Kira Bungert, Moritz Wolter, Stefan Krumpen ans Michael Weinmann, Reinhard Klein, and Maren Bennewitz. Where can i help? human-aware placement of service robots. In *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2020.

[99] Jorge Rios-Martinez, Anne Spalanzani, and Christian Laugier. From proxemics theory to socially-aware navigation: A survey. *International Journal of Social Robotics*, 7(2), 2015.

[100] Mark Tee Kit Tsun, Bee Lau, and Hudyjaya Siswoyo Jo. An improved indoor robot human-following navigation model using depth camera, active ir marker and proximity sensors fusion. *Robotics*, 7(1), 2018.

[101] AbdElMoniem Bayoumi and Maren Bennewitz. Learning optimal navigation actions for foresighted robot behavior during assistance tasks. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2016.

[102] M. Kuderer and W. Burgard. An approach to socially compliant leader following for mobile robots. In *Social Robotics*, volume 8755 of *Lecture Notes in Computer Science*. Springer International Publishing, 2014.

[103] Daniel Althoff, Dirk Wollherr, and Martin Buss. Safety assessment of trajectories for navigation in uncertain and dynamic environments. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2011.

[104] Peter Henry, Christian Vollmer, Brian Ferris, and Dieter Fox. Learning to navigate through crowded environments. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2010.

[105] David Feil-Seifer and Maja Matarić. People-aware navigation for goal-oriented behavior involving a human partner. In *Proc. of the IEEE Int. Conf. on Development and Learning (ICDL)*, volume 2, 2011.

[106] Anaís Garrell and Alberto Sanfeliu. Local optimization of cooperative robot movements for guiding and regrouping people in a guiding mission. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2010.

[107] Julia Fink. Anthropomorphism and human likeness in the design of robots and human-robot interaction. In *International Conference on Social Robotics*. Springer, 2012.

[108] Amalia F Foka and Panos E Trahanias. Predictive autonomous robot navigation. In *IEEE/RSJ international conference on intelligent robots and systems*, volume 1, 2002.

[109] Ninad Pradhan, Timothy Burg, Stan Birchfield, and Ugur Hasirci. Indoor navigation for mobile robots using predictive fields. In *2013 American Control Conference*. IEEE, 2013.

[110] M. Kollmitz, K. Hsiao, J. Gaa, and W. Burgard. Time dependent planning on a layered social cost map for human-aware robot navigation. In *Proc. of the Europ. Conf. on Mobile Robotics (ECMR)*, 2015.

[111] Iis P Tussyadiah and Sangwon Park. Consumer evaluation of hotel service robots. In *Information and communication technologies in tourism*. Springer, 2018.

[112] Jihang Ye, Zhe Zhu, and Hong Cheng. What's your next move: User activity prediction in location-based social networks. In *Proceedings of the SIAM International Conference on Data Mining*. SIAM, 2013.

[113] Tom Carlson and Yiannis Demiris. Human-wheelchair collaboration through prediction of intention and adaptive assistance. In *IEEE International Conference on Robotics and Automation*, 2008.

[114] Mohamed-Hédi Amri, Yasmina Becis, Didier Aubry, and Nacim Ramdani. Indoor human/robot localization using robust multi-modal data fusion. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2015.

[115] Md Muztoba, Rohit Voleti, Fatih Karabacak, Jaehyun Park, and Umit Y Ogras. Instinctive assistive indoor navigation using distributed intelligence. *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, 23(6), 2018.

[116] Muhammad Raisul Alam, Mamun Bin Ibne Reaz, and MA Mohd Ali. Speed: An inhabitant activity prediction algorithm for smart homes. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 42(4), 2011.

[117] Jonathan Mumm and Bilge Mutlu. Human-robot proxemics: Physical and psychological distancing in human-robot interaction. In *Proc. of ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2011.

[118] Edward T Hall et al. Proxemics. *Current anthropology*, 9, 1968.

[119] Kay Kitazawa and Taku Fujiyama. Pedestrian vision and collision avoidance behavior: Investigation of the information process space of pedestrians using an eye tracker. In *Pedestrian and evacuation dynamics 2008*. Springer, 2010.

[120] Liang Zhao, Tatsuya Ohshima, and Hiroshi Nagamochi. A* algorithm for the time-dependent shortest path problem. In *WAAC08: The 11th Japan-Korea Joint Workshop on Algorithms and Computation*, 2008.

[121] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, et al. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.