Predicting Human Navigation Goals based on Bayesian Inference and Activity Regions*

Lilli Bruckschen*, Kira Bungert, Nils Dengler, Maren Bennewitz

Humanoid Robots Lab, University of Bonn, Germany

Abstract

Anticipation of human movements is of great importance for service robots, as it is necessary to avoid interferences and predict areas where human-robot collaboration may be needed. In indoor scenarios, human movements often depend on objects with which they interacted before. For example, if a human interacts with a cup the probability that a table or coffee machine might be the next navigation goal is high. Typically, objects are grouped together in regions depending on the related activities so that environments consist of a set of activity regions. For example, a workspace region may contain a PC, a chair, and a table with many smaller objects on top of it. In this article, we present an approach to predict the navigation goal of a moving human in indoor environments. We hereby combine prior knowledge about typical human transitions between activity regions with robot observations about the human's current pose and the last object interaction to predict the navigation goal using Bayesian inference. In the experimental evaluation in several simulated environments we demonstrate that our approach leads to a significantly more accurate prediction of the navigation goal in comparison to previous work. Furthermore, we show in a real-world experiment how such human motion anticipation can be used to realize foresighted navigation with an assistance robot, i.e. how predicted human movements can be used to increase the time efficiency of the robot's navigation policy by early anticipating the user's navigation

*All authors are with University of Bonn, Germany. This work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) BE 4420/2-1 (FOR 2535 Anticipating Human Behavior).

*Corresponding author

Email address: brucksch@cs.uni-bonn.de (Lilli Bruckschen)

URL: www.hrl.uni-bonn.de (Lilli Bruckschen)

goal and moving towards it.

Keywords: Anticipating Human Behavior, Robot Path Planning, Human-Centered Systems

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 [1] or predict where assistance may be needed. Previous approaches have tackled this problem by learning typical human trajectories in known environments [2, 3] or reacting dynamically to humans in close proximity [4]. However, in many cases a lot can be learned about human movements by looking at the last objects they have interacted with. For example, if we know that a human has interacted with objects inside a kitchen it is likely that they will next move towards a dining area. In indoor environments, objects are typically grouped together in activity regions, i.e., regions containing objects related to certain activities. We therefore propose 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. The contributions of our work are the following:

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

To identify activity regions, we first conducted an online survey to identify how humans perceive activity regions. We were primarily interested in the question which objects would be grouped together and at what distances humans would stop perceiving close objects as associated groups. Based on these results, we designed a rule-based classifier 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 our previous work [5], we learn transition probabilities between different activity regions given the objects present in these regions. We combine this prior knowledge with online observations of the human's pose in a Bayesian inference framework to predict which activity region is likely to be the next navigation goal.

Our approach makes use of an RGB-D camera system from which we estimate the user's pose and human-object interactions [5]. We further use a semantic map of the environment, that includes object positions and activity regions as well as knowledge about typical human-object interaction sequences as prior knowledge. Using these information our approach automatically infers the next activity region in which the human will interact with an object.

As we show in the evaluation, our approach achieves a higher prediction accuracy than a trajectory based reinforcement learning method [6] while simplifying both the representation as well as the needed training data. We also demonstrate in a real-world experiment how our framework can be used for foresighted robot navigation.

Fig. 1 shows a motivating example of our approach in which a human interacts with a cup. As can be seen by the different colors our system identifies four different activity regions and determines the most likely one as the region consisting of the table and the water bottle.

This article is an extended version of our previous publication [7]. In particular, we improved the prediction framework by utilizing Bayesian inference, generalized the interaction detection by moving from individual objects to activity regions, and contributed additional experiments. In our experimental evaluation, we show that the extensions provide a significant improvement of the prediction accuracy compared to our previous work.

2. Related Work

As noted by Kruse *et al.* [1] the ability to predict human movements is vital for any robot that operates in the same environment. Therefore a lot of research has been

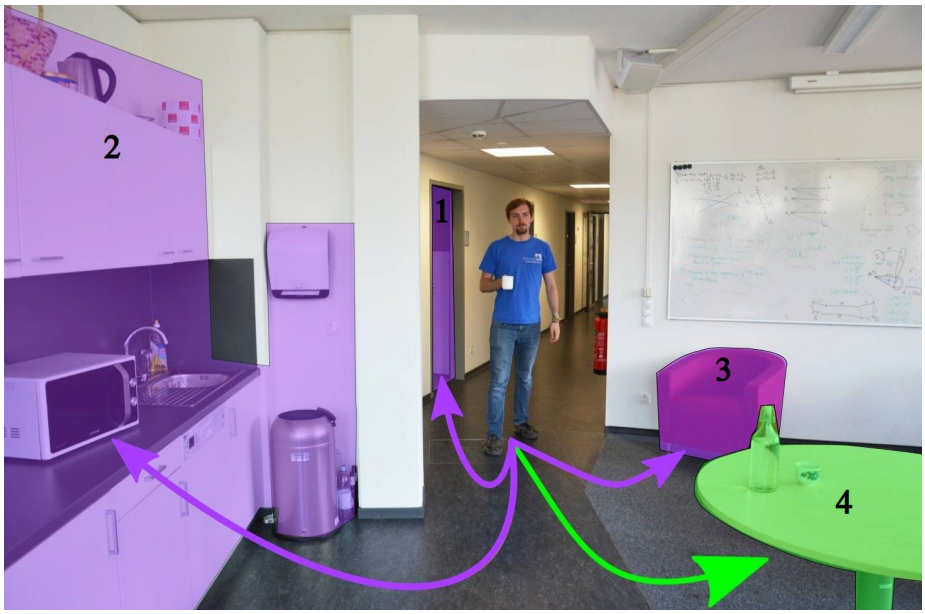


Figure 1: The aim of our framework is to infer the navigation goal of a moving human. This figure was segmented by hand and shows a sketch about the underlying idea of our framework. The user arrived from an office area where he interacted with a cup which he now carries. This was detected with an RGB-D camera using the approach from [5]. Four possible activity regions (1,2,3,4) 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 (4), consisting of the table, bottle and cup, is the most likely one, while the violet regions are also possible with lower probabilities.

put into frameworks that are used for the prediction of human navigation goals [2, 3] or navigation through dense crowds [8, 9].

An overview and taxonomy about recent prediction frameworks is given by Rudenko *et al.* [10]. The authors categorized approaches in regards to their modeling of the future human 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 our presented approach falls into the

planning-based category, as we infer likely long-term navigation goals of the human based on their previous behavior.

In previous work, we implemented a pattern-based prediction approach for moving humans in indoor environments. We hereby used typical trajectories to learn a foresighted navigation policy for a service robot via Q-learning [2] and to find a user quickly if they cannot be located in the proximity of the robot [6]. In contrast to our current work, objects have not been explicitly used during the learning or in the prediction, however, they are inherently related to human navigation goals. While this 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, our new framework utilizes the same training data for multiple environments.

Vasquez *et al.* [11] proposed 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 our approach, this system works only for short-term motion prediction. Ziebart *et al.* [3] 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. Ziebart *et al.* 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 our work, we can also 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 [12, 13] or learning of social models to predict the behavior of humans in lively places [4, 14]. However, those approaches have been developed for short-term prediction of human motions and trajectory adaptation 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 crowded spaces [8]. 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 [15]. 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.* [16]. The authors proposed a method to extract the navigation goal of a user based on their walking activities for a robotic walking aid. The authors 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 [17].

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.* [18]. 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 [19]. 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 [20]. In contrast to our activity regions the authors used the interaction frequency as classification criteria while we use proximity and composition based classification based on previously collected human feedback.

While object-based prediction, to the best of our knowledge, has not been applied in existing motion prediction systems, this concept has often been used in the context of higher-level action prediction [21, 22]. 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.* [23]. 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.

3. Identifying Activity Regions

We aim at predicting the navigation goal of a moving human based on knowledge about typical object interactions. One observation is that humans often interact with multiple objects in close proximity before starting to move to the next navigation goal. As an example, consider office work, where a human typically simultaneously interacts with a chair, table, and laptop before moving to some other place. In our previous work [7], we modeled this as three different object interactions, first an interaction with a chair, then with the table, and finally with the laptop. This resulted in multiple new navigation goal predictions in a very short time. To achieve a better generalization, we now propose to group objects together into so-called *activity regions* and describe in this section how to identify such regions. First, we consider objects that "overlap", i.e., objects for which depth values are approximately equal such as a table and objects on top of it as belonging to the same activity region. However, often objects that do not overlap are also used in combination for an activity, e.g., a chair and a table.

To define activity regions that make intuitively sense to humans we performed an online survey on how humans perceive activity regions and how specific activity regions are composed. In the remainder of this section we discuss both, the design of the survey as well as the results.

3.1. Design of an Online Survey to Identify Activity Regions

We created the survey[1] using Qualtrics [24] and published it on Clickworker [25]. The survey was online for one week and during this time 125 users participated. We included an attention test, which was passed by 106 german participants from a cross section of the population [26]. Responses from users that did not pass the attention test were excluded.

We included three different types of questions inside the survey. We first showed the participants pictures of office environments with given, color-coded groupings for different objects and asked them which of the options felt most natural to them (see Fig. 2 for an example), to gain insight of possible subconscious classification rules. We used 15 questions of this type and randomized their order as well as the order of possible answers.

For the second part, we asked the participants which objects they consider to be close to given objects. The participants had to indicate the likelihood of ten objects to belong together on an even Likert scale, which had six options ranging from very unlikely in close proximity to very likely in close proximity. We used seven questions of this type including the one which serves as attention test. Questions and possible answers were again randomized.

In the last part of the survey, we were interested which, if any, objects humans associate with two example activities: *food processing* and *office work*. Participants could choose between 15 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. We chose these example tasks as they represent activities where the robot would possibly be able to provide assistance.

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

(a) (b)

Figure 2: Example question of our survey. The participants were asked which one of the two object groupings appeared more natural to them, with the hope of gaining insight of possible subconscious classification rules. We recorded 106 answers to this question. 72 participants voted for option (a) while 34 voted for option (b). This corresponds to a p-value smaller than 0.01 with the One Sample Chi-Square Test for this particular question and does therefore imply a significant preference for option (a).

3.2. Results of the Survey

The results of the survey[2] match our expectations as we can see 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 we asked 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. 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 user. Fig. 3 shows an example of a question regarding possible groupings of close tables.

We also found a clear classification of objects into the two example activity classes, office work and food processing, as can be seen in Fig. 4 and Fig. 5 respectively. This supports our theory that grouping objects into regions related to activities matches typ-

[2]We published the complete results of the survey on our website `https://www.hrl.uni-bonn.de/publications/activity_region_survey_results`

(a) (b)

Figure 3: Example table grouping question from our survey. Participants were asked if they found grouping (a) or grouping (b) more natural. We recorded 106 answers to this question. The results were exactly split as 53 participants voted for option (a) and the remaining 53 for option (b).

ical human behavior.

In summary, we found that our 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 others center, and that our participants could clearly identify objects that they would expect inside two example activity regions. Using these results, we built a proximity based rule system to automatically identify activity regions on a semantic map, as we describe in the next section.

4. Prediction of Navigation Goals

As explained above, we consider the problem of predicting the navigation goal of a moving human in an indoor environment. Our prediction is based on observations of the human's location and pose as well as on prior knowledge about a map of the environment and typical human transitions between objects. To obtain the prior knowledge, we propose to learn a semantic map of the environment [27] and afterwards group the objects into activity regions, as discussed in Section 3.

Furthermore, we use pre-recorded videos of humans acting in indoor environment to train a prediction model for transitions of human-object interactions. In other words, we learn a model to predict how likely it is that a human who interacted with object A will next interact with object B and call this the *interaction model*. The training videos

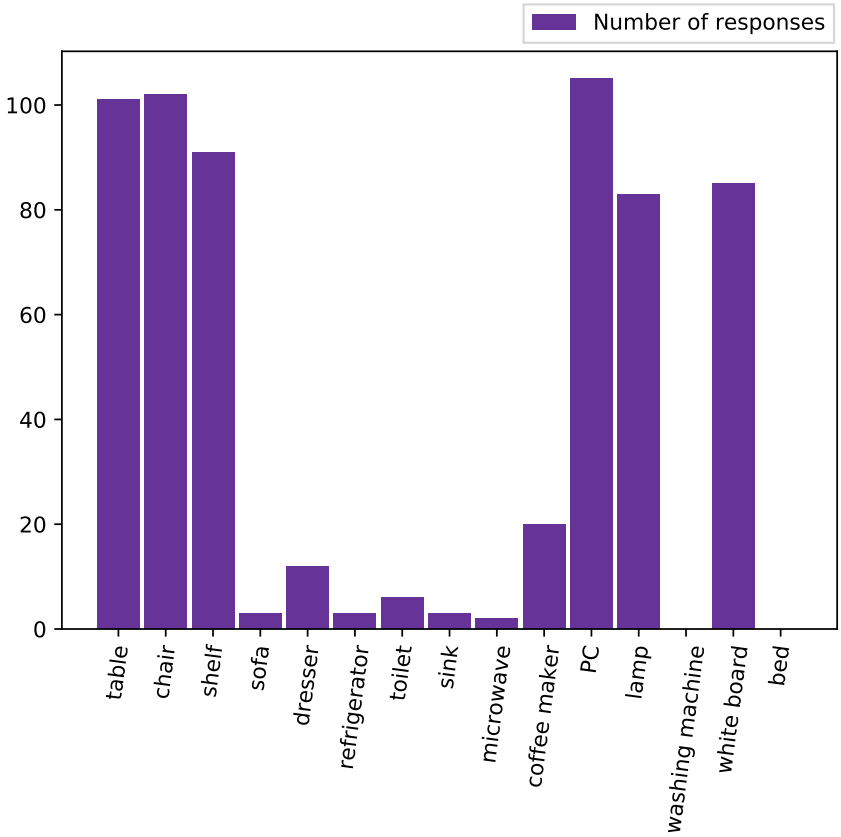


Figure 4: Responses of the 106 survey participants regarding the question which objects they expect in an office environment. As can be seen, there is a clear expectation towards the objects: table, chair, shelf, PC, lamp, and white board.

for which we hold the associated rights are published with the Bonn Activity Maps dataset [28].

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

As application scenario we use a Robotino mobile platform [29], equipped with a RealSense D435i RGB-D camera [30] and a laser scanner. However in principle the system can be used on any RGB-D camera system, even without a robot.

In the following, we explain all components of our prediction framework in detail.

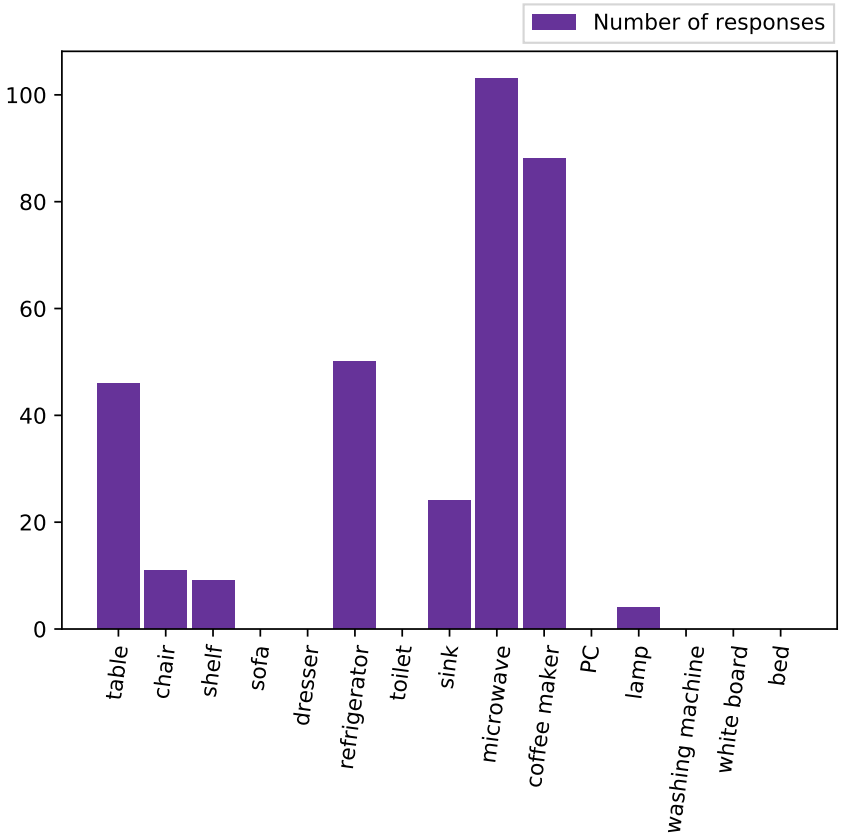
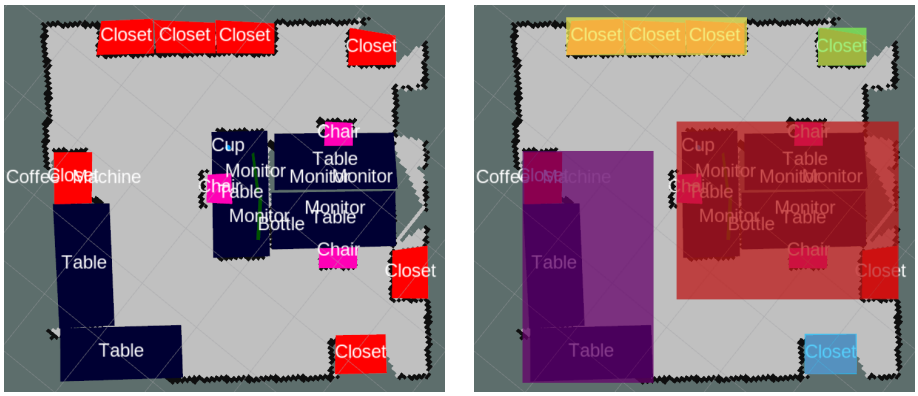


Figure 5: Responses of the 106 survey participants regarding the question which objects they expect in a food processing environment. As can be seen, there is a clear expectation towards the objects: table, refrigerator, microwave, and coffee maker.

4.1. Semantic Environment Representation

We represent the environment as a static inflated occupancy map [31], with an additional semantic layer to encode objects and activity regions. The occupancy mapping can be realized with a common SLAM approach [32]. Object information was added to the map through semantic mapping by using RGB-D masks from CNN object detectors and projecting them to the 2D plane [27]. To infer activity regions for a map we use a rule-based system. First, we assign each object its own activity region. If two regions overlap, i.e., have overlapping object bounding boxes with similar depth



(a) (b)

Figure 6: Example of the generation of activity regions. (a) a semantic map which contains 23 objects, (b) grouping of all objects which centers are less than 2 meters apart from each other into the same activity region. This results in 5 different activity regions, shown as bounding boxes around the objects.

values of points inside them (the average depth values differ by at most 5%), we merge them. The new region then consists of all objects of the previous regions. This process is repeated until no more regions can be merged. Second, we use the results of the survey discussed in Sec. 3 by merging activity regions that are less than 2 meters apart from each others center. 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 $R = (\mathcal{X}_R, C_R)$. We define the position of an activity region $\mathcal{X}_R$ as the center of the bounding box around all objects inside the region. Fig. 6 shows an example of activity regions for a previously recorded semantic map. Note that the human is not part of the map nor is the map updated during the prediction steps.

4.2. Interaction Model

We define the interaction model $I(\tau_a|\tau_b)$ as the distribution that describes the probability that a human which previously interacted with an object of type τ_b will next interact with an object of type τ_a. To train the model, we collected data of users in indoor environments and recorded their object interaction sequences as described in

detail in our previous work [5]. Note that to generalize well between different environments, we only consider object-interaction sequences to learn the interaction model and do not consider the actual trajectories of the human. As we use I 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. We hereby use the normalized sum of all transition probabilities between objects of different types in the individual regions. In other words, if two chairs are present in region A and one sofa in region B, we will only count the transition probability of one chair towards the sofa. 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:

$$I_R(R_b|R_a) = \eta \cdot \sum_{x \in T_b} (\sum_{y \in T_a} I(x|y)) \tag{1}$$

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. Note that, as the positions of objects and activity regions is given as prior knowledge, we can directly infer 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.

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, we use RGB-D observations from the robot to obtain information about the human. We first detect the human and their pose using a pose estimation system (OpenPose [33]). Once the human is detected, we use the robot's laser data to determine the distance to the robot. Using the known position of the robot, we can now infer the position $\mathcal{X}_h$ of the human. For the orientation of the human θ_h we use pose data to

track the face and shoulders. We then infer the general direction the human is orientated to, as shown in [5]. The complete state of the human at the current time step is defined as $S := (\mathcal{X}_h, \theta_h)$. Note that for simplicity we currently consider scenarios in which only one human is present, however our approach 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 increasing the prediction accuracy of our approach.

4.4. Bayesian Inference

We use 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 last interacted object, to predict the human's next navigation goal. 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:

$$P(R_i|S) = \frac{P(S|R_i)P(R_i)}{\sum_{R_j \in \mathcal{R}} P(S|R_j)P(R_j)} \tag{2}$$

Using η as a normalizing parameter Eq. (2) can therefore be simplified to:

$$P(R_i|S) = \eta \cdot P(S|R_i)I_R(R_i|R_L) \tag{3}$$

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

$$I_R(R_i) = \sum_{R_j \in \mathcal{R}} I_R(R_i|R_j) \tag{4}$$

This modifies Eq. (2) in the case of no observed interaction to:

$$P(R_i|S) = \eta \cdot P(S|R_i) \cdot I_R(R_i). \tag{5}$$

In both Eq. (2) and Eq. (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. We therefore 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, we compute the 2D orientation difference $\Delta a(\theta_h, \theta_{opt})$ 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 $L(\mathcal{P}_{h \to R_i})$ be the length of the path $\mathcal{P}_{h \to R_i}$. With an added value of 1 to avoid a situation in which we would divide by zero, the observation likelihood $P(S|R_j)$ can then be defined as:

$$P(S|R_j) = (L(\mathcal{P}_{h \to R_i}) + 1)^{-1} \cdot (\Delta a(\theta, \theta_{opt}) + 1)^{-1}. \tag{6}$$

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

$$P(R_i|S) = \eta \cdot (L(\mathcal{P}_{h \to R_i}) + 1)^{-1} \cdot (\Delta a(\theta, \theta_{opt}) + 1)^{-1} I_R(R_i|R_L) \tag{7}$$

if the last object interaction was observed and otherwise as

$$P(R_i|S) = \eta \cdot (L(\mathcal{P}_{h \to R_i}) + 1)^{-1} \cdot (\Delta a(\theta, \theta_{opt}) + 1)^{-1} I_R(R_i) \tag{8}$$

The belief about the navigation goal is updated in a constant interval as long as the human is visible and moving. For our implementation we chose an interval of 5 seconds.

5. Experimental Evaluation

We evaluated our framework based on the accuracy of the prediction in regards to the human's true navigation goal. We performed a quantitative and qualitative evaluation as well as a comparison to previous approaches. Furthermore, we demonstrate the applicability of our system to foresighted robot navigation in a real-world experiment.

5.1. Data Collection

To guarantee comparability of different approaches and eliminate noise in the observations, we performed the majority of our experiments in simulation. Using the V-REP editor [34], we created 10 different office and home environments with sizes between $100\,m^2$ and $150\,m^2$, modeled after real-world examples. 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 our previous work [5], the detector we use is able to register interactions with the 510 different objects and animals from the Open Image dataset [35] 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 our training set for our interaction model we restricted the number of objects for the evaluation. We trained the object interaction model with a set of 161 recorded human-object transitions [5] from which we then determined the regional interaction model. Furthermore, we collected a dataset of 64 typical human transitions between two objects, based on recorded movements within our real test environments as well as a survey about typical indoor movements inside our simulated home environments. As mentioned in Sec. 4, the videos for which we hold the associated rights are published with the Bonn Activity Maps dataset [28].

We used the data to randomly sample 300 transitions between objects. Based on these transitions, we computed 300 different trajectories distributed over all environments using A*. The simulated trajectories were then used as test data for the quantitative evaluation, the recorded real-world transitions between objects were used for the qualitative evaluation.

5.2. Quantitative Evaluation

For the quantitative evaluation, we tested our approach on 300 test trajectories in 10 different office and home environments, as described in Sec. 5.1. Each trajectory had on average 64 different possible goal objects, grouped in 19 different activity regions. Fig. 7 shows an example of one of our office maps with a test trajectory and possible

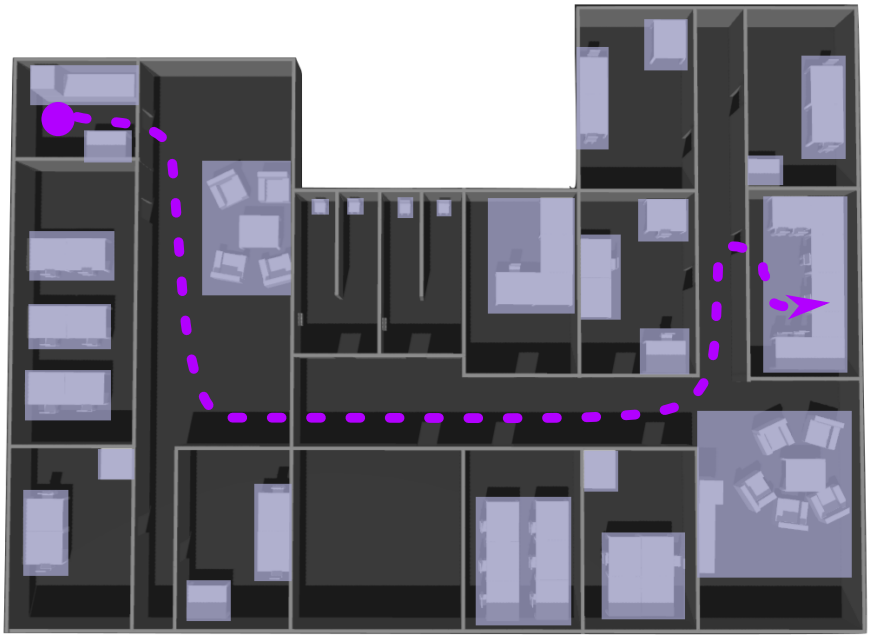


Figure 7: Sketch of an example simulated office environment from the qualitative evaluation. Objects are shown in gray with activity regions in light gray on top of them. An example trajectory from the starting position of the user (violet circle) near a coffee machine to an office is shown as dotted 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.

navigation goals. For the first evaluation, we assume that the human is always observable during their movement and that the robot has perfect observations, i.e. the probability for false positive or negative observations is 0. Once every second our system makes a prediction about the human's most likely navigation goal and compares this prediction to the human's true navigation goal. We use this general prediction accuracy as standard metric in the evaluation and, furthermore, evaluate a variant called *top* 5% accuracy. Here, we count a prediction as correct if the true navigation goal is among the returned top 5% of most likely navigation goals. We chose this metric to show that even if our 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, we used the A* distance from the center of the prediction navigation goal to the true navigation goal of the user 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.

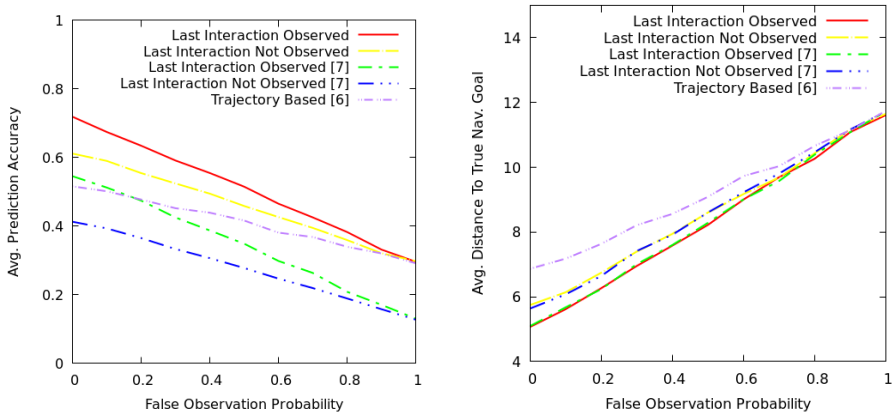
We evaluated our region-based approach both with and without a known initial object interaction, as specified in Eq. (7) and Eq. (8) respectively. For comparison, we also performed the experiments with our previous approach [7], which does not use activity regions and uses objects instead of activity regions as possible navigation goals, as well as with a trajectory-based prediction approach [6], which does not use information about object interactions and does not consider the orientation in the prediction. Activity regions are navigation goals for our approach and the trajectory-based approach [6], while single objects are possible navigation goals for our previous object-based approach [7]. Both our activity region and object-based approach use the same interaction model trained on 161 recorded human-object transitions for all maps. The trajectory based approach needed a further training for each of the 10 maps using 30 additional sampled trajectories for each map. The results of the quantitative evaluation are shown in Table 1.

As can be seen, the use of activity regions improves the average prediction accuracy by 0.15 if the last interaction was observed, in comparison with our previous work without them [7]. Similar results are achieved if the last interaction was not observed, here we achieve an improvement of 0.20. 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 our previous 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

	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 [7]	0.53	0.68	5.09
Last interaction **not** observed [7]	0.46	0.66	5.63
Trajectory-based approach [6]	0.36	0.57	6.86

Table 1: Results of the quantitative evaluation. The first two rows represent the results of our approach with activity regions. The third and forth row show the results of our previous work [7] with single objects instead of activity regions and the last row shows the results of the trajectory-based apprach [6].

of both objects. The average distance between the predicted and true navigation goal supports this theory. The resulting distances for our current and previous approach are very close together. We interpret this as a sign that false predictions in our previous 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 [36]. 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 [6], we achieve an improvement of 0.32 for the average accuracy and an improvement of 0.14 for the top 5% metric with our new method considering the activity regions. Furthermore we were able to reduce the average distance between



(a) (b)

Figure 8: (a) change of the average prediction accuracy, (b) average distance between the predicted and true navigation goal with false observations. As can be seen, the accuracy and distance decrease linearly for all evaluated approaches.

the predicted and true navigation goal by 1.8 meters.

We further performed an evaluation with noisy observations. We added a false observation probability to model how likely it is that the robot observes a random user pose during an update step, instead of the correct observation. If the human would normally not be visible for the robot at the chosen position the observation information are discarded and only the prior knowledge is used. A false observation probability of 1 depicts a scenario in which all observation information are randomly chosen, i.e. the robot thinks that the user has a randomly chosen visible position on the map with a random orientation or no information about the user's orientation and pose. The results of this evaluation are shown in Fig. 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 our approach is able to yield an average prediction accuracy of more than 0.6 up to a 40% probability of false observations.

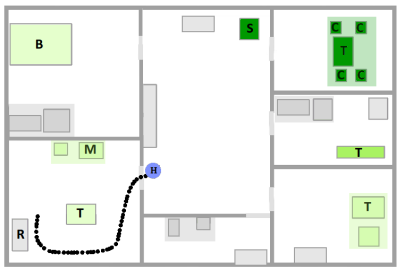
5.3. Qualitative Evaluation

To evaluate the importance of individual observations at different times during the movement of the human, we analyzed how the returned probability of the true navigation goal changes over time. Fig. 9 shows 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 our approach. The importance of observation decreases the closer the human gets towards their navigation goal. Once the last 60% of the trajectory are reached the predicted goal region 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.

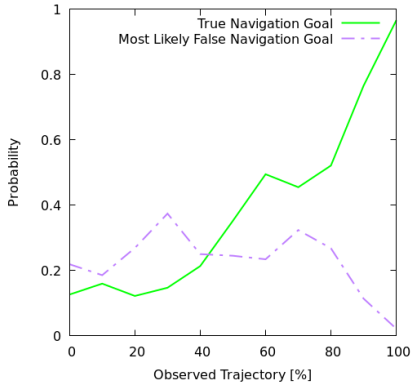
Fig. 10 shows an example of a typical human trajectory and the evolution of belief over time.

5.4. Application

A framework for predicting human movements is essential for foresighted robot navigation [1]. To highlight this use case, we combined the presented system with a positioning approach. We tested our framework on a Robotino mobile platform [29] equipped with an RGB-D camera and a laser scanner in a university environment. The robot uses a 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 simultaneously avoiding interferences. The prediction is updated every 5 seconds. To compute the optimal position of the robot, we use the average distance between all possible navigation goals weighted by their probability while avoiding positions in a $1.5m$ radius around the human. We use a function $C(\mathcal{X}, S)$ to compute the costs of each possible robot position $\mathcal{X}$ on M based on the current observation of the



(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 9: Example of a recorded trajectory from our dataset in a simulated environment. In this case, the human first interacts with a refrigerator and then moves to a chair-table activity region. (a) Trajectory observed so far at the point where the prediction is correct for the first time. (b) Corresponding evolution of the belief about navigation goal.

human $S := (\mathcal{X}_h, \theta_h)$ and prior knowledge:

$$C(\mathcal{X}, S) = \begin{cases} \infty \text{ if } \texttt{dist}(\mathcal{X}, \mathcal{X}_h) < 1.5m \\ \sum_{R_i \in \mathcal{R}} ((1 - P(R_i|S)) \cdot \texttt{dist}(\mathcal{X}, \mathcal{X}_h)) \text{ else} \end{cases} \tag{9}$$

With $\texttt{dist}(\mathcal{X}_a, \mathcal{X}_b)$ as the Euclidean 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. 11 illustrates an example experiment. Here, the robot i initially in a corridor where it observed a human-object interaction with a cup. The robot then updates the prediction about the navigation goal and computes a new position for itself based on Eq. (9) (see Fig. 11 (a)). During the movement, the robot regularly updates both, the prediction of the navigation goal as well as the cost of possible positions (see Fig. 11 (b)). In Fig. 11 (c), the human has entered the room containing the

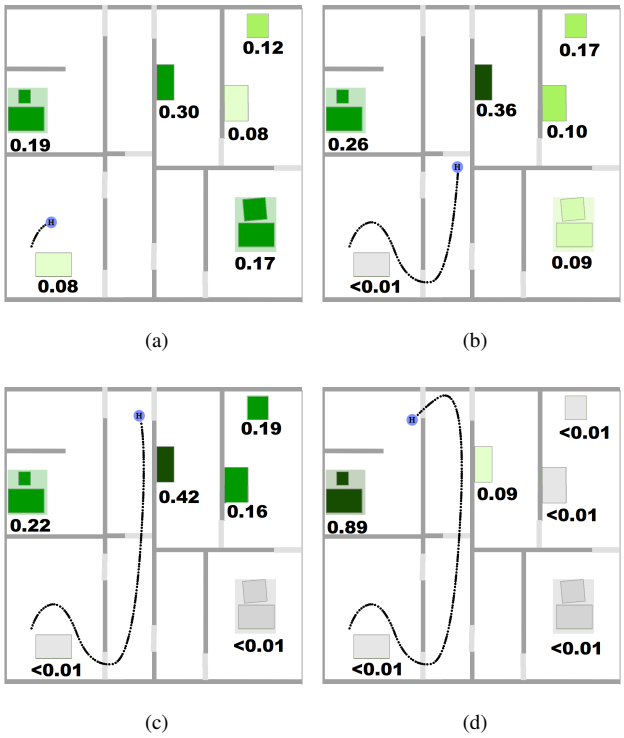


Figure 10: Evolution of the belief over time for an recorded human trajectory from our dataset in a simulated environment. As can be seen, the initial belief based on the interaction model (a) is continuously updated with new observations (b), (c), (d). Objects are shown in green and activity regions in coloured shades with their probabilities to be the navigation goal: the darker the green the higher the probability. Doors are colored in grey and walls in dark grey. The human is depicted as a blue dot with their trajectory as dashed line.

navigation goal and the robot correctly updates 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.

We further tested this application in our simulated environment, using the same environments and trajectories as in Sec. 5.2. We assumed that the human is always observable for the robot as well as perfect sensors, i.e., a false positive observation chance of 0. We configured the user to travel with an average velocity of $1\frac{m}{s}$ and the

(a) (b) (c)

Figure 11: Application example of our approach to 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 (orange circle) and user (blue 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 (orange 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.

robot to travel with $2\frac{m}{s}$. As metric we used the average arrival time difference between the user and the robot when they first entered an area $1.5m$ around the goal activity region. We found that the robot arrived on average $8.5s$ before the human.

6. Discussion

As shown in Sec. 5, our framework reliably predicts the navigation goal of a moving human. Our approach 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 our 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, our framework has no means to rule out the wrong navigation goal until the human passes by. We plan to solve such cases in the future by further enhancing the prior knowledge by learning better interaction models. Furthermore we also plan 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 our system could involve an improved detection of human-object interactions, e.g. by the utilization of smart home systems and their detection capabilities.

7. Conclusion

In this article, we presented an approach to predict the navigation goal of moving humans in indoor environments. We proposed to use knowledge about typical human-object interaction sequences and to group close-by objects into activity regions to achieve generalization. To get information about typical activity regions in indoor environments, we performed and evaluated an online survey with 125 participants.

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

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

Acknowledgments

We would like to thank Sabrina Amft, Sandra Höltervennhoff, Sophie Jenke, Padmaja Kulkarni, Jenny Mack, Saskia Rabich, Mosadeq Saljoki, and Matthew Smith for their help during our experiments.

References

[1] Kruse T, Pandey AK, Alami R, Kirsch A. Human-aware robot navigation: A survey. Robotics and Autonomous Systems 2013;61(12):1726 –43.

[2] Bayoumi A, Karkowski P, Bennewitz M. Learning foresighted people following under occlusions. In: Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS). 2017, p. 6319 –24.

[3] Ziebart BD, Ratliff N, Gallagher G, Mertz C, Peterson K, Bagnell JA, et al. Planning-based prediction for pedestrians. In: Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS). 2009, p. 3931 –6.

[4] Kretzschmar H, Spies M, Sprunk C, Burgard W. Socially compliant mobile robot navigation via inverse reinforcement learning. The International Journal of Robotics Research 2016;35(11):1289 – 1307.

[5] Bruckschen L, Amft S, Tanke J, Gall J, Bennewitz M. Detection of Generic Human-Object Interactions in Video Streams. In: Proc. of the International Conference on Social Robotics (ICSR). 2019, p. 108 –18.

[6] Bayoumi A, Karkowski P, Bennewitz M. Speeding up person finding using hidden Markov models. Robotics and Autonomous Systems 2019;115:40 –8.

[7] Bruckschen L, Dengler N, Bennewitz M. Human motion prediction based on object interactions. In: Proc. of the Europ. Conf. on Mobile Robotics (ECMR). 2019, p. 1 – 6.

[8] Alahi A, Goel K, Ramanathan V, Robicquet A, Fei-Fei L, Savarese S. Social LSTM: Human trajectory prediction in crowded spaces. In: Proc. of the IEEE Intl. Conf. on Computer Vision (ICCV). 2016, p. 961 –71.

[9] Ferrer G, Zulueta AG, Cotarelo FH, Sanfeliu A. Robot social-aware navigation framework to accompany people walking side-by-side. Autonomous Robots 2017;41(4):775 – 793.

[10] Rudenko A, Palmieri L, Herman M, Kitani KM, Gavrila DM, Arras KO. Human motion trajectory prediction: A survey. The International Journal of Robotics Research 2020;39(8):895–935.

[11] Vasquez D. Novel planning-based algorithms for human motion prediction. In: Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA). 2016, p. 3317 –22.

[12] Kim S, Guy SJ, Liu W, Lau RW, Lin MC, Manocha D. Predicting pedestrian trajectories using velocity-space reasoning. In: Algorithmic Foundations of Robotics X. Springer; 2013, p. 609 –23.

[13] Lefèvre S, Vasquez D, Laugier C. A survey on motion prediction and risk assessment for intelligent vehicles. Robomech Journal 2014;1(1):1 – 14.

[14] Robicquet A, Sadeghian A, Alahi A, Savarese S. Learning social etiquette: Human trajectory understanding in crowded scenes. In: Proc. of the Europ. Conf. on Computer Vision (ECCV). Springer; 2016, p. 549 –65.

[15] Pfeiffer M, Paolo G, Sommer H, Nieto J, Siegwart R, Cadena C. 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). 2018, p. 5921 –8.

[16] Glover J, Thrun S, Matthews JT. Learning user models of mobility-related activities through instrumented walking aids. In: Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA); vol. 4. IEEE; 2004, p. 3306 –12.

[17] Best G, Fitch R. Bayesian intention inference for trajectory prediction with an unknown goal destination. In: Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS). IEEE; 2015, p. 5817 –23.

[18] Helbing D, Molnar P. Social force model for pedestrian dynamics. Physical review E 1995;51(5):4282 –6.

[19] Rudenko A, Palmieri L, Lilienthal AJ, Arras KO. Human motion prediction under social grouping constraints. In: Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS). 2018, p. 3358 –64.

[20] Karaoğuz H, Bore N, Folkesson J, Jensfelt P. Human-centric partitioning of the environment. In: Proc. of the International Symposium on Robot and Human Interactive Communication (RO-MAN). IEEE; 2017, p. 844 –50.

[21] Lan T, Chen TC, Savarese S. A hierarchical representation for future action prediction. In: Proc. of the Europ. Conf. on Computer Vision (ECCV). 2014, p. 689 – 704.

[22] Koppula HS, Gupta R, Saxena A. Learning human activities and object affordances from RGB-D videos. Intl Journal of Robotics Research (IJRR) 2013;32(8):951 – 970.

[23] Duckworth P, Hogg DC, Cohn AG. Unsupervised human activity analysis for intelligent mobile robots. Artificial Intelligence 2019;270:291 – 294.

[24] Qualtrics . Qualtrics software. `https://www.qualtrics.com`; 2020. Last visited 2020-05-04.

[25] Clickworker GmbH . Clickworker software. `https://www.clickworker.de/`; 2020. Last visited 2020-05-04.

[26] Clickworker GmbH . Clickworker user base. `https://www.clickworker.com/about-us/clickworker-crowd/`; 2020. Last visited 2020-05-07.

[27] Zaenker T, Verdoja F, Kyrki V. Hypermap mapping framework and its application to autonomous semantic exploration. 2019. ArXiv preprint arXiv:1909.09526.

[28] Tanke J, Kwon OH, Stotko P, Rosu RA, Weinmann M, Errami H, et al. Bonn activity maps: Dataset description. 2019. ArXiv preprint arXiv:1912.06354.

[29] Festo Didactic GmbH & Co. KG . Robotino manual. https://www.festo-didactic.com; 2020. Last visited 2020-05-04.

[30] Intel RealSense . Realsense d435i. https://www.intelrealsense.com/depth-camera-d435i/; 2020. Last visited 2020-13-09.

[31] Siegwart R, Nourbakhsh IR, Scaramuzza D. Introduction to autonomous mobile robots. MIT press; 2011.

[32] Grisetti G, Stachniss C, Burgard W, et al. Improved techniques for grid mapping with Rao-Blackwellized particle filters. IEEE Trans on Robotics (TRO) 2007;23(1):34 – 46.

[33] Cao Z, Hidalgo G, Simon T, Wei SE, Sheikh Y. OpenPose: Realtime multi-person 2d pose estimation using part affinity fields. 2018,.

[34] E. Rohmer S. P. N. Singh MF. V-Rep: A versatile and scalable robot simulation framework. In: Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS). 2013, p. 1321 –6.

[35] Krasin I, Duerig T, Alldrin N, Ferrari V, Abu-El-Haija S, Kuznetsova A, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from https://storage googleapis com/openimages/web/index html 2020;.

[36] Bruckschen L, Bungert K, Wolter M, ans Michael Weinmann SK, Klein R, Bennewitz M. Where can i help? human-aware placement of service robots. In: IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). IEEE; 2020,.