A Survey on Human-aware Robot Navigation

Ronja Möller, Antonino Furnari, Sebastiano Battiato, Aki Härmä, Giovanni Maria Farinella

*Department of Mathematics and Informatics, University of Catania, V. Andrea Doria, Catania 95125, Italy
1Philips Research, High Tech Campus 34, 5656 Eindhoven, Netherlands
2Cognitive Robotics and Social Sensing Laboratory, ICAR-CNR, Palermo, Italy

Abstract

Intelligent systems are increasingly part of our everyday lives and have been integrated seamlessly to the point where it is difficult to imagine a world without them. Physical manifestations of those systems on the other hand, in the form of embodied agents or robots, have so far been used only for specific applications and are often limited to functional roles (e.g. in the industry, entertainment and military fields). Given the current growth and innovation in the research communities concerned with the topics of robot navigation, human-robot-interaction and human activity recognition, it seems like this might soon change. Robots are increasingly easy to obtain and use and the acceptance of them in general is growing. However, the design of a socially compliant robot that can function as a companion needs to take various areas of research into account. This paper is concerned with the navigation aspect of a socially-compliant robot and provides a survey of existing solutions for the relevant areas of research as well as an outlook on possible future directions.

Keywords: Robot Navigation, Human Robot Interaction, Active Vision, Activity Recognition

1. Introduction and Motivations

Robots have come a long way from initial attempts to create artificial humanoids that could walk and or talk to more functional designs that are truly autonomous, such as the work of Boston Dynamics® a spin-off from the Massachusetts Institute of Technology on one side, and near-lifelike copies of humans such as “Geminoid”® built by Hiroshi Ishiguro Laboratories, on the other side. While early robots served as outlooks for future technology or performed very simple tasks designed to impress the audience, more recent research has produced an increasingly diverse number of robots that can perform more and more complex tasks independently of user input. The inclusion of robots in industrial settings, beginning with a digitally programmable robot that lifted and stacked metal pieces by George Devol in 1954® has fundamentally changed virtually every industry (see, e.g. [1] for an excellent thought-provoking summary). Robots are now widely used in manufacturing, assembly, packing, transport, search and rescue, healthcare, surgery and laboratory research to name just a few.

The usage of robots in social contexts, on the other hand, is still in an earlier stage. A few companies have successfully marketed robots as pets, with the Tekno (a small robotic dog) being one of the more popular, and recently companion robots like Pepper®, Jibo® and Loomo® have been presented. Even with these successes, none of them have reached the popularity of intelligent non-embodied systems such as, for example, the Amazon Alexa.

In addition to being useful and comfortable to be around, a social robot naturally has to be affordable for the demographic it is meant for. While institutions such as hospitals or hotels can afford more sophisticated and expensive robots, the “standard” user - and researcher - will be constrained by a certain budget. As a result, the most commonly used sensors in research tend to be affordable options such as simple distance or proximity sensors (ultrasound, infrared and sometimes LiDAR - Light Detection and Ranging), odometry sensors, GPS (Global Positioning System) and images/video. Image sensors in particular are very popular across all kinds of applications and price ranges. This survey will focus on image-based research.

As soon as an embodied agent starts moving in the user’s space and interacting with them, a variety of social rules has to be obeyed, as discussed in the book by Knapp et al. on Non-Verbal Communication in Human Interaction [2]. The robot’s behaviour has to be carefully designed to be human-like enough to inspire trust and confidence without erring into the uncanny valley territory [3], risking user rejection. The way a robot moves is a particularly important aspect of user acceptance, with a variety of studies evaluating how the speed of movement,
observed distances and cultural biases influences the user’s perception of a robot [4], [5], [6]. Given the interdisciplinary nature of this task, solutions have to fuse the insights of several different areas and deal with the advantages and shortcomings of each. Robot Navigation, for example, is a well-established field of research with a variety of available metrics [7], benchmarks and datasets [8], [9], [10], [11]. Human-robot-interaction on the other hand is more subjective and evaluated under different criteria [12], making joint datasets or evaluations difficult. The necessary inclusion of human activity recognition (and potentially emotion recognition) - to enable the robot to move in reaction to the action its user is currently performing (and its mood) - further compounds the issue. Even with these difficulties, researchers are motivated to tackle this fascinating problem.

Artificial Intelligence on an embodied social companion has the potential to improve the quality of life, equalize disparities for disabled users and, as argued by Topol in his book on deep learning in medicine [13], make healthcare more “human” by solving menial tasks so the caregivers have more time for their patients. Social robots are also expected to have an important role in motivation and patient engagement, in the self-management of a chronic disease and adherence to therapies - to enable the robot to move in reaction to the action its user is currently performing (and its mood) - further compounds the issue. Even with these difficulties, researchers are motivated to tackle this fascinating problem.

While these ideas seem lofty and far-fetched, every embodied agent trying to accomplish those goals will have to move in a way that is socially compliant, hence solving this task will always be the first step. This survey will consider active vision, robot navigation, human-robot interaction and human activity recognition as well as their possible overlaps. Several very informative surveys cover robot navigation [14], [15], [16], [17] - some with a focus on visual input [18], [19], [20] - and the topics of human-robot interaction [21], [22], [23], [24] and human action recognition have been surveyed extensively as well [25], [26], [27], [28]. We will review the topic of robot navigation with a focus on recent trends and approaches that can be used in settings where humans will be encountered. Previous surveys on socially aware robot navigation focused purely on the social requirements [29], such as comfort, naturalness and cultural conventions, or the type of mapping that was used for navigation [30], such as metric mapping, semantic mapping and social mapping. Our aim with this work is to give a more general overview and highlight the interdisciplinary nature of the topic.

The preferred input sensor in most surveyed approaches will be RGB(D) (Red Green Blue (Depth)) cameras, but we will include publications that use different sensors if they have influenced later visual approaches in terms of strategies and evaluations. We will also cover popular concepts of human-robot interaction, concentrating on those relevant to everyday surface-level social “interaction” such as navigating around or with humans. However, the survey will not focus on actual social interaction such as talking, touching, and collaborating. For human-robot interaction in industrial settings, we would like to suggest [31] for a relatively recent overview. For a robot to interact naturally and economically with a human user it is also important to take into account what action is currently performed by the human(s). To this end, we will look into work on activity recognition that uses visual or depth sensors as input. We will not go into depth on alternatives like inertial sensors or RFID (Radio-frequency identification) tags.

We believe that in order for a socially-aware robot to behave “correctly”, it is important to combine insights from all four topics - namely active vision, robot navigation, human-robot interaction and human behaviour modeling - this survey that covers all these topics in conjunction, to inspire further research. We will start with an introduction to active vision as the paradigm underlying all research with embodied active agents, then we will cover the topic of visual robot navigation and its most popular strategies, measures and current state of the art. In the later sections, the focus will be on “the human factor” with introductions to the topics of Human-Robot-Interaction, the analysis and prediction of human movement and human behaviour. In each of these sections we will also highlight works that have successfully implemented socially aware navigation strategies that include the respective information (Activity Recognition, Trajectory Prediction etc.) The last section
2. The Active Vision paradigm

We will first give a brief introduction to the difference between the active and passive vision paradigms and how they relate to the tasks of classification, segmentation and navigation. The second section will go into detail about specific subsets of the active vision paradigm and how they can be applied to concrete problem settings.

2.1. Active vs. Passive Vision

Simply put, an Active Vision System can influence the input by controlling the camera, while a Passive Vision System cannot. This is a special case of Active Learning, which differs from Passive Learning in that the algorithm actively and intentionally chooses which samples to train on. In Passive Learning and by extension Passive Vision, the input data is chosen only by the researcher and the algorithm simply processes each given sample.

Instead of learning the correspondence between a fixed set of inputs and the desired outputs (for example classification or segmentation), an Active Vision System learns advantageous behaviour during training and or testing. In addition, through active learning, an agent can learn underlying dependencies or hierarchies that cannot be captured by learning from independent datapoints. A concrete example would be an agent that learns dependencies of objects and rooms, such as: fridge and microwave are in the kitchen, toilet and shower are in a different room. These dependencies and hierarchies can only be learned through either explicit modeling or active exploration of the sample space.

Differences in input data. The added complexity of an Active Vision System is also motivated by the fact that vision applications on robots, especially robots capable of movements, will always have to take into account that the input data differs a lot from standard classification datasets. While curated datasets such as ImageNet and MS COCO have undoubtedly greatly pushed the state of the art for image classification and segmentation, they contain samples that are well curated and mainly acquired from a third person perspective. This assumption does not hold for what a robot will see “in the wild” and an argument can be made that this way of learning also does not correspond to how humans learn. Indeed, humans learn not with large numbers of perfectly centered and lit samples but instead with fewer samples that are looked at from different directions. Moreover, humans are embodied agents, which allows them to continuously interact with the world they learn from. An example for this difference would be an algorithm that learns about an object by being fed thousands of image examples versus a robot navigating around objects of the specific type, see Figure 2. Not only will the input data look different in terms of orientation, occlusion, lighting etc. but it contains the added information about the actions the robot performed to arrive at the new input. In contrast, the samples on the right are all centered, unobstructed, from similar perspectives and independent with no relationship between the samples.

Starting from the outlined differences, two ways of handling the problems come to mind. One could either accept that the input data will be incomplete/degraded and learn strategies to cope - for example temporal smoothing, consistency checks and common sense rules. The other approach would be to accept that the issue is intrinsic to having an embodied agent and instead learn behaviour that improves the input quality and therefore yields a better results. In other words, the freedom of movement can be seen as a feature, rather than a problem. While both options are viable for many problem settings, such as classification, detection and segmentation, the task of object navigation tends be solved with the second one.

Examples for Passive Vision. There are successful works on robot navigation that use a straightforward approach of matching input data to a desired output: Bojarski et al. from NVIDIA proposed an end-to-end solution for self driving cars by mapping raw RGB input from a single front-facing camera on a car to directions. In a similar vein, Giusti et al. created their own dataset (with several viewpoints) and used it to map an image to the correct direction when navigating along a wilderness trail. A current iteration of this more “passive” learning approach was proposed by Chang et al.
al. [56], who used the implicit movement labels on real estate Youtube videos to train an agent.

Examples for Active Vision. In contrast, there are several arguments for movement during training and testing. A very early work from the sixties by Held and Hein [37] suggested that kittens, as a standin for intelligent agents who navigate using visual input, can only learn the correspondence between movement and visual input by physically performing this movement, not by merely observing changing environments. In their experiments, a set of kittens who where restricted in an apparatus to be moved passively, displayed worse movement capabilities compared to kittens who were allowed to actively move. Bajcsy et al. [58] make a convincing argument why any research on robot navigation necessarily has to include active perception, to quote their conclusion: “An agent is an active perceiver if it knows why it wishes to sense, and then chooses what to perceive, and determines how, when and where to achieve that perception”. An active solution for autonomous automotive navigation has been proposed by Kendall et al. [39], whereas an active approach for robot navigation was presented by Morad et al. [40].

The addition of a social aspect is another dimension that is very hard to learn in a passive way: it is challenging to definitively decide on a “correct” motion in relation to humans since they react differently to the presence of the robot and have their own specific behaviours as well [41], [4], [42].

The active paradigm will be present in most of the embodied robot research in this paper. Although the survey will show how many passive algorithms can be “re-purposed” to fit this viewpoint, the fundamental difference between active vision and the standard “passive” approach is still relevant.

2.2. Active Perception and Active Learning

Active Perception refers to a setting in which the robot/sensor can move during testing to gain more useful information and thereby improve the performance. For a summary of active perception we recommend the original introduction by Aloi-monos et al. [43] from the eighties and the updated review by the same authors [38], as well as the more recent work by Chen et al. [44]. The applications of this mindset are also not necessarily limited to robotic navigation tasks, but can also extend to exploratory behaviour and image inference or hallucination - see the work of Grauman et al. [45], [46] - and object category detection - see Ammirato et al. [9] who used a novel dataset to simulate active vision.

In contrast to Active Perception, Active Learning [47] in a robotic context is more concerned with exploiting “movement” (as in, the ability to actively influence the input) during training, enabling the agent to intelligently choose what datapoints to exploit next. Applications include video segmentation, as proposed in Fathi et al. [48], image classification on intelligently chosen data subsets as in Sener et al. [49], and medical image analysis (see Mahapatra et al’s [50] work on Crohn disease).

Another interesting and very recent work by Chaplot et al. [51] combines embodied robot navigation with active vision to finetune a pretrained network for a segmentation task.

3. Visual Robot Navigation

The navigation of mobile robots has been of great academic interest for over 30 years and the field has continued to change with advancements in hardware and algorithms [18]. Depending on the design and purpose of the robot, different strategies for locomotion have been proposed [52]. For a comprehensive summary of the sensors used before 2000, we recommend the surveys by Lobo et al. [52] and Borenstein et al. [53].

The main driving industries for autonomous sensing in recent times have been the military and the automotive sector and the latter is very open about extensively using cameras [54]. This trend combined with the increased demand for complex and “human-like” sensing for assistive robots and the decreased price of commercial cameras has made visual robot navigation very appealing.

Given the historic progression of the field, there are different ways to break down the overall topic and these divisions keep changing with the preferred sensors and applications. In this survey, we will keep the distinction between map-based and mapless navigation from [18] and [19] with an emphasis on mapless navigation since our focus is on navigation in relation to humans in a variety of possibly unstructured or unknown environments. We will not be separating approaches into indoor-vs outdoor navigation like the mentioned surveys as the most socially relevant interactions tend to happen indoors. Instead, we will present some navigation tasks that are currently popular in the academic community and how they relate to human-aware navigation.

Although several of the presented problem settings do not contain humans or treat them as static obstacles only, they are convenient to study general navigation abilities, which are fundamental for social robots. Moreover, many strategies from classic navigation have been adapted to include a social dimension while using the same underlying algorithms.

3.1. Map-based vs. Mapless Navigation

When designing a robot and the accompanying algorithms, one of the earliest design decisions to be made is the type of environment the robot should operate in and whether or not it should create an internal representation of it. Benefits of internal representations (in the form of maps) are that the robot can navigate faster to areas it has already seen and therefore operate in a more economic way. An example where this type of behaviour is desired would be an industry robot that only operates in a certain unchanging setting and is always able to compute the shortest path to a goal in the known environment. In less controlled, less structured or non-static environments such as public spaces or homes with moving pedestrians, the creation of a fixed map might be impossible (due to complexity or
Socially aware Robot Navigation using Topological maps. An application of topological maps for socially aware robot navigation was proposed by Johnson et al. [68], who learn general social behaviour as well as biases (such as moving on the right). A survey on semantic mapping was compiled by Kostavelis et al. [69], with an emphasis on human-robot interaction.

In contrast to the map-based navigation strategies, mapless or map-free algorithms never explicitly build an internal map of the environment. Early examples such as [70] employ visual input and template matching, while [71] use stereo vision for navigation. Mapless navigation has historically been difficult to evaluate and received less attention than approaches based on the popular SLAM paradigm. Interestingly, with the increased interest in Reinforcement Learning for robot navigation and its combination with Deep Learning, mapless navigation has become a topic of interest again, albeit indirectly - see for example [72]. The MDP (Markov Decision Process) model, that Reinforcement Learning is based on, does not demand an explicit model or map of the environment and therefore many current navigation algorithms are technically mapless, although they are not explicitly labelled as such.

Recent works have also broadened the definition of maps and environment representations. While the environment can be stored in a metric grid, it is hard to reliably create and maintain such a map [77]. More flexible options are the topological maps mentioned before or the compressed internal representations of Neural Networks [73]. For a comprehensive review of Representation Learning approaches we recommend the work by Bengio et al. [74]. A map can also be created purely based on the specific task the robot is trying to solve, as in the work by Zhao et al. [75] that focuses on a healthcare context. This representation hypothesis has led to the usage of auxiliary tasks related to multi-task learning (see [76] for a survey on multi-task learning) that are known to improve another main Deep Learning task in the realm of Reinforcement Learning and Navigation [77]. It could be argued that this is a return to “map-based” algorithms with a broader understanding of maps or a middle ground between mapless and map-based.

3.2. Navigation Tasks

In this section, we will look at specific tasks that are currently used to evaluate robot navigation algorithms operating in three-dimensional environments. Some of the tasks are derived from specific applications (e.g. a search and rescue robot needs to perform exploration [78]), while others are derived from the availability and quality of sensors, such as the tasks in [7] which require GPS and RGB(D) input. Some of them are more relevant for the academic community than for the deployment of actual physical robots, where human comfort and features and traditional grid maps, the approach has more recently been used with visual input [63], topological maps [64] and even Deep Learning algorithms [65]. We recommend the survey by [65] for a detailed introduction to visual SLAM. A combination of Deep Learning, Reinforcement Learning, topological maps and SLAM was recently presented by Chaplot et al. in [67].

Figure 3: An example of an ideal metric map (left) that could be created by a SLAM algorithm vs. the same environment encoded in a topological map (right).

change). As a result, it might be more useful to instead learn general behaviours without an internally maintained map. An example for this type of behaviour would be a robot that learns to navigate in unknown scenes by minimizing collisions. Earlier work focused on metric or grid-like internal map representations (see for example the highly influential work by Moravec et al. [55], where a sonar is used to generate an occupancy grid). The robot can navigate among non-occupied space on the map. Further research has investigated other map representations. For example, the authors of [55] and [57], who introduced the concept of topological maps consisting of interconnected landmarks in a graph-like structure that encode traversability in the edges.

Figure 3 illustrates different map representations for the same environment. The edges between the vertices of the topological map denote accessibility/traversability. The rooms themselves can be encoded in the topological map in various ways, for example using visual similarity or activities that are performed frequently.

Metric maps are intuitive and provide a good basis for metric optimization problems like shortest path computations, but they are difficult to build, very inefficient for larger spaces and their usability degrades the less accurate the robot position is. Topological maps on the other hand can be more efficient, especially for simple environments, they don’t require the robot’s position to be very exact and they support more high-level planning approaches. The drawbacks of topological maps are ambiguity, their high complexity for articulated or cluttered environments and that they cannot be used for metric path optimization, but only for symbolic approaches. This division does not have to be strict either, as seen in the work by Thrun et al. [55]. Although the mentioned works focus on distance sensors and precede vision-based approaches, more current approaches that use cameras still use those paradigms. For a comprehensive overview on map-based navigation that is somewhat sensor-agnostic we recommend the two outstanding reviews by Meyer Filliat [59] [60] that include camera and non-camera sensors but also discuss approaches independently of the actual sensors.

Navigation systems can also rely on SLAM (Simultaneous Localization and Mapping) to tackle the real-world problem of simultaneously creating an internal map and localizing the robot inside it while the robot explores the environment [61]. While SLAM can be implemented with low-cost distance sensors (see [62] for a comparison of sensors for SLAM), handcrafted
material wear and tear might play a more important role. The overall goal of all of these tasks is to have a common language for specific problems that can be used for benchmarking in the research stage. To allow for a fair comparison, these benchmarking tasks tend to be evaluated in a simulated environment, which has its own advantages and drawbacks. These tasks are intentionally designed to be simple and straightforward in terms of the problem description, allowing for easy evaluation and comparison. Of course, successfully completing these tasks is just one step in the development of a fully autonomous robot that can assist humans in social, industrial or healthcare contexts. The deployment in social situations and in healthcare environments in particular will require more complex behaviour, but breaking down behaviour into subtasks such as the ones mentioned in this section is still useful if not necessary. For example a doctor could ask an assistive robot to fetch supplies or patients as part of its overall assistive behaviour.

3.2.1. PointGoal Navigation

In response to the traction that robot navigation had gained, Anderson et al. [7] proposed common problem definitions and language to deal with the influx of many different algorithms and evaluation protocols that were for the most part incompatible and incomparable. They propose three main navigation tasks, of which the first, PointGoal navigation, is currently common in publications [79] [80] [81]. The other two, ObjectGoal and AreaGoal are slight variations of the first one. The PointGoal navigation task is mastered successfully if the robot navigates to a certain position (point) in the 3D environment from any starting position and orientation (see Figure 4). In the the easy case, where start and stop position are in the same room (row 1), the solution would be trivial - just choose the action that decreases the GPS distance. In a more complex case (the two lower rows), the agent also has to learn how to navigate around obstacles - in this case a wall - to arrive at the destination. While the general goal of decreasing the GPS distance remains valid (2a, 2b), the agent learns to explore in order to deal with obstacles, even if the GPS distance is increased for a short amount of time (2d), otherwise it will fail (2c).

The standard PointGoal navigation task assumes that the robot gets GPS information encoding the distance or displacement to the goal as well as other sensors on top if desired.

This task definition was also used for the Habitat-challenge[9], an international challenge that aimed at benchmarking visual robot navigation in 2018. Unfortunately, this specific definition has two obvious drawbacks:

- The assumption of a precise displacement sensor is unrealistic in everyday settings, where the robot will have to deal with sensor noise and actuation noise, especially considering the problems indoor environments pose for GPS sensors;
- The availability of GPS coordinates for the goal is unrealistic to begin with in most real settings.

Independently of the drawbacks, the task does serve as a “proof of concept” for navigation algorithms if the chosen environments are complex enough in terms of obstacles. Any algorithms that can successfully navigate along the path of the shortest geodesic distance (the shortest path to the goal around obstacles) and does not just blindly decrease the geometric distance (the GPS displacement) has to have learned certain semantic properties of the scene. Although the classic PointGoal navigation task remains a common evaluation baseline it has become obvious that it is hard to translate into the real world. Researchers have started to look into the influence of noise (see the work by Rosano et al. [82] and Chaplot et al. [83]) and task definitions that do not assume the constant availability of the perfect GPS displacement[9].

Very similar iterations of this task are ObjectGoal navigation, where the agent is asked to navigate to an object instead of a point, and AreaGoal navigation, where the agent is asked to navigate to an area.

3.2.2. Exploration

In contrast to the previous task, where the exploration of unknown areas of the environment was treated as a byproduct of solving the specific task, some work treats exploration itself as the primary goal. An early approach by [78] starts from a task that is similar to PointGoal navigation (with given GPS coordinates for the agent and the goal) and then proposes

![Figure 4: Visualisation of the PointGoal Task in increasing levels of difficulty (from the first row to the last two rows) and with different solution strategies (columns). The agent starts at the green dot and has to move to the blue dot. The black outline denotes the floor plan of a simple environment.](https://aihabitat.org/challenge/2019/)

![Figure 5: Diagram showing different solution strategies for the PointGoal Task.](https://aihabitat.org/challenge/2020/)
a customizable exploration strategy that creates an internal topological representation while navigating to the goal.

While the mentioned work focuses on tactile sensors, later examples [84] use images and present a topological map-building algorithm that can be used for exploration. In applications where the exploration is not motivated by a long-term goal, different evaluation methods have been proposed that in turn have shaped different exploration strategies. One common strategy is frontier-based [85], where the robot learns to navigate to the boundaries between explored and unexplored space. Alternatives maximize coverage [86, 83] to cover areas and recently the authors of [46, 45] minimized uncertainty about the environment in a task-independent approach. In the uncertainty-based approach the agent is asked to predict unseen parts of the environment and the goal is to make this prediction match up with reality in as short a time as possible, which can be applied to robot exploration as well. A recent method by Ramakrishnan et al. [87], the winner of the 2020 Habitat PointNav Challenge, uses an exploration strategy where the robot anticipates a top down occupancy map.

3.3. Popular Paradigms

Historically, visual robot navigation has been achieved with classic image processing tools such as edge detectors and optical flow [18], but - like in the field of Image Processing in general - the approaches have changed and evolved into more complex algorithms. A popular choice (for SLAM solutions in particular [61]) is using Gaussian distributions and therefore variations of Kalman filters to track local features and later, non-Gaussian approaches such as particle filters [91].

Another well-established paradigm is Fuzzy Logic and Control, which is well suited to model imprecise and non-numerical information and decisions. It is a widely used in automation technology [92], recommendation systems [93] as well as medical inference systems [94] and Human-Robot Interaction [95].

The most recent change has come with the revival of Neural Networks and in particular Deep Convolutional Neural Networks [32], which now achieve dominating performance, especially in combination with Reinforcement Learning [96].

The paradigms mentioned here are not necessarily disjoint and can also be combined - as shown by [97] who use a fuzzy Reinforcement Learning approach for Robot Navigation, an approach that draws heavily on the fuzzy version of Q-Learning presented by Glorenc et al. [98].

3.3.1. Fuzzy logic and control

Fuzzy Logic and, as a result, Fuzzy Control, are based on the assumption that certain processes or decisions might be modeled best not with binary states (true or false) but with more imprecise descriptors such as “warm”, “cold”, “big” or “small”. The idea was formalised by Lotfi Zadeh [99], who introduced Fuzzy sets, building on earlier earlier work on many-valued or infinite-valued logic, such as the work by Lukasiewicz [100]. The Process for Fuzzy Control contains 3 main steps:

1. The input data is converted into a state where it is described in terms of memberships to certain “Fuzzy sets” (for example “young”, “old”, “middle-aged”), which can be understood as the controller’s version of how a human expert would describe the input.
2. With the converted - fuzzy - input data, several operations (logical operations such as “AND”, “OR”, “NOT”) are performed to calculate the fuzzy output. The conversion correspond to the “IF-THEN” rules a human operators would use to justify decisions.
3. The generated output is then de-fuzzified (often with a weighting scheme) to produce numerical output.

Compared to the more complex solutions, such as Deep Neural Networks, Fuzzy Control schemes have the advantage that they are completely “readable”, since both the definitions of the fuzzy sets and the logical operations on them can easily be transformed into verbal instructions. Fuzzy Control has been employed in the context of robot navigation, see [101] for an early overview as well as the more recent one by Hong et al. [102]. Given its reactive nature, Fuzzy Control is useful for collision avoidance: The works by Omran et al [103] and Pandey et al. [104] combine collision avoidance and goal-oriented navigation into a single fuzzy controller, Parhi et al. [105] not only include obstacle avoidance in cluttered environments but also other robots.

Socially aware Navigation using Fuzzy Control. A solution from the area of Fuzzy Logic was proposed by Palm et al. [106], wherein they use Fuzzy Control to predict the movement
Figure 5: Illustration of how a fuzzy controller could be used for socially aware hallway navigation. These prototypical encounters could easily be controlled with if-then rules. 1 - If the human comes form the left, do not yield. 2 - If the human comes from the right, yield. 3 - If the human is stationary, move around.

3.3.2. Reinforcement Learning

Machine Learning can be broadly separated into three major paradigms: 1. Supervised Learning, 2. Unsupervised Learning and 3. Reinforcement Learning [108]. The first branch is the “classic” use case where, based on an existing dataset, the task is to learn the correspondence between input samples and the correct output. Unsupervised Learning, in contrast, explores cases where no labels are available and therefore no input-output correspondence is known. It is more concerned with finding intrinsic structures in the data and inferring similarities and differences from those. Reinforcement Learning, the third branch, uses a trial-and-error approach to learn optimal behaviour of an “agent” in an “environment”. The aim is less to learn the one correct output or action for a certain input but instead overall optimal behaviour. This makes Reinforcement Learning particularly well-suited for modeling the behaviour of robots (the robot takes the place of the agent and the world it operates in corresponds to the environment) and for training them to perform a certain task. We recommend the survey by Arulkumaran et al. [96] for an overview on Reinforcement Learning for Robot Navigation. A very recent survey by Akalin et al. [109] explores the use of Reinforcement Learning in Social Robotics.

Socially aware Navigation using Reinforcement Learning. Recent works by Chen et al. [110] and Chen et al. [111] suggest that the complex problem of socially aware navigation could be tackled with a holistic end-to-end approach based on reinforcement learning. While the first work focuses on “what not to do” when navigating in proximity to humans, the second work is more concerned with navigation among crowds. Other solutions that employ reinforcement learning have been proposed by Kim et al. [112], Kretzschmar et al. [113] and Okal et al. [114].

3.3.3. Learning by Demonstrations

An intuitive approach to learning behaviour is to learn from a teacher or expert by replicating their actions, also called Learning by Demonstration. A comprehensive overview of Robot Learning from Demonstration was provided by Argall et al. [115], who looked at the design strategies, the data acquisition and the various learning and optimisation techniques that can be employed in Learning from Demonstration for robots. A more recent survey was compiled by Ravichandar et al. [116]. An example of expert demonstrations that cannot be mapped perfectly to the robot was presented by Skoglund et al. [117], who used the arm movements of a human expert to train a robot arm with a different morphology, thus highlighting the importance of generalisation and smart abstraction. In the context of Robot navigation, Learning by Demonstration has been used by Du et al. [118], who used a combination of group behaviour replication and human following to navigate sidewalks and Xiao et al. [119], who explored the adoption of navigation strategies via human expert demonstrations.

Socially aware Navigation using Learning by Demonstration. Since pedestrians can intuitively respond to human-like behaviour instead of “optimal” behaviour, Learning by Demonstration is naturally suited to the training of social robots. Li et al. [120] used roleplaying in simulation to train a robot to navigate in a socially compliant manner with Reinforcement Learning. The demonstrations were mined from real-world trajectory datasets. A shopkeeper robot was trained by Liu et al. [121], who used examples of human behaviour to train speech and navigation behaviour. Another Learning by Demonstration strategy for social navigation was suggested by Higuera et al. [122], using Inverse Reinforcement Learning and Optimal Rapidly-exploring Random Trees that can be trained with demonstrations. See table 1 for a comparison.

3.4. Simulation Based Learning

Training robots to navigate from scratch purely on real data is notoriously difficult, due to the costs involved in using a real robot, which could be damaged in the early stages of learning. This problem is only exacerbated by the inclusion of humans in the training procedure for interactive navigation. Indeed, it is not feasible or responsible to demand that human participants engage in procedures that might directly harm them in cases of
The Mohamed Bin Zayed International Robotics Challenge\footnote{https://www.mbzirc.com/} with a focus on (one or several) robots, the AlphaPilot drone challenge\footnote{https://ai2thor.allenai.org/} and the RoboThor challenge\footnote{https://github.com/StanfordVL/iGibson} where a robot is trained in a simulated version of a real room and has to navigate to an object in the real room at test time.

To combat this issue and in an attempt to guide research, steps have been taken to build simulators that are as universal as possible, with a variety of features that can be used by researchers to reliably benchmark their specific algorithms. As previously discussed, we will be focusing on visual input but for a more general view at simulation environments for robots, we recommend the surveys by Staranowicz et al.\footnote{https://aihabitat.org/} and Torres-Torriti et al.\footnote{https://www.lockheedmartin.com/en-us/news/events/ai-innovation-challenge.html}.

3.4.1. Simulation-to-real-transference

Unfortunately, even with impressive simulation frameworks, some issues remain. No matter how realistic the environment and the robot are modeled they still do not behave like they would in the real world and even small inconsistencies or biases in the simulation setup can skew the results of comparisons or competitions. A great example of this issue has been highlighted by the developers of the Habitat simulator themselves\footnote{https://www.mbzirc.com/}. They report that the agents that perform particularly well in the Habitat challenge (using the Habitat simulator), tend to exploit inconsistencies in the simulator, namely a small inaccuracy in the implementation of collisions. The team working on iThor has tackled the problem by creating a physical room with a corresponding counterpart in their simulation\footnote{https://ai2thor.allenai.org/}. Some solutions to combat issues like the ones above have been suggested. We will highlight some examples in the following:

**Close the gap by making the simulator very realistic:**
Simulators such as AI Habitat\footnote{https://ai2thor.allenai.org/} and Gibson\footnote{https://www.lockheedmartin.com/en-us/news/events/ai-innovation-challenge.html} use scans of real environments which are advertised to be photorealistic. Gibson (an interactive extension of iGibson\footnote{https://aihabitat.org/} and iThor\footnote{https://ai2thor.allenai.org/}) focus on modelling realistic interaction with dynamic objects in the scene. Habitat has also recently started to push the development of sensor and actuation noise models to minimize the difference between the real world and the simulation that results from noise. Another way to close the gap was proposed by Alhaija et al.\footnote{https://www.lockheedmartin.com/en-us/news/events/ai-innovation-challenge.html}, who augmented real data with simulated components, thereby guaranteeing variability and a certain level of realism.

**Learn a transfer function from simulated to real data:**
In cases where it is difficult to realistically model the environment the robot should operate in, a possible strategy might be to accept the differences and learn how to cope with them. The problem of translating images from a source domain to a different target domain has been tackled by Zhu et al. with their Cycle-GAN\footnote{https://www.mbzirc.com/}. Shrivastava et al.\footnote{https://www.lockheedmartin.com/en-us/news/events/ai-innovation-challenge.html} proposed the usage

\begin{table}[h]
\centering
\begin{tabular}{|l|l|l|l|l|}
\hline
Publication & problem definition as written in the paper & underlying algorithm & used dataset & used metrics & test with real humans \tabularnewline
\hline
120 & Learning socially with human companions & Reinforcement Learning & simulation & minimal condition that encoder: reached goal, collisions, lost companion & yes \tabularnewline
\hline
121 & Reproducing observed social interactions & Clustering, Bayesian classifiers & proprietary dataset & correctness of wording, speech and motion, appropriateness of responses and in social role & yes \tabularnewline
\hline
122 & Learning (social) navigation behaviors & Inverse Reinforcement Learning & proprietary dataset & time to goal, path length, distance to people and others & yes \tabularnewline
\hline
\end{tabular}
\caption{Comparison of the Algorithms that Use Learning by Demonstration for Socially aware Navigation.}
\end{table}
of GANs to train a transfer function that converts a simulated image into one that is indistinguishable from a real photo\cite{133}. An example from the domain of robot navigation is the work by Li et al.\cite{134}, who utilize a domain-invariant content space and learn the correspondences between the content space and the source domain as well as the one between the content space and the target domain. The survey by Csurka\cite{135} provides a comprehensive overview about visual domain adaptation techniques with a section dedicated exclusively to synthetic-to-real-transference.

**Pretraining on simulated data, finetuning on real data:**

Since one of the problems of working with synthetic data is that the algorithm will eventually overfit on the synthetic domain, it is intuitive to stop the training process before it is completely finished and then finetune the network on real data. Examples of this strategy can be found in segmentation and recognition problems: Richter et al.\cite{136} use synthetic data from the video game Grand Theft Auto (a popular game series in the academic community to model high fidelity scenes in the automotive context) to improve performance on the Cambridge-driving Labeled Video Database. Hunde et al.\cite{137} use simulated indoor environments to improve segmentation results on the NYUv2 and SUN-RGBD sets and\cite{172} use AI2Thor to pretrain a robot that is later finetuned to operate in a real environment.

**Randomization:**

A popular strategy to ensure that a policy transfers well (or at least better) to reality is to vary the properties of the simulation in order to train a policy that is robust to inaccuracies in the simulation. Peng et al.\cite{138} demonstrated this by varying the dynamics of the simulation of a robot arm that is tasked with pushing an object - by varying their simulation properties, they trained a policy that was able to execute the task in a real setup as well. A similar task, namely object localisation and robot grasping was tackled by Tobin et al.\cite{139} by varying the textures in the simulated environment. James et al.\cite{140} use a “goal corridor” approach in conjunction with a simulation to train a robot to perform tidying actions (picking up an object and placing it somewhere else). They achieve invariance to illumination, clutter and occlusion with domain randomisation in the simulator. It is interesting to note that both Tobin et al. and James et al. do not aim for realism in their randomisation efforts at all. In fact, the visual input of the randomisation strategies is highly unrealistic but it still transfers well to reality.

3.5. Datasets and Simulation frameworks for Robotic Navigation

In this section we will present a selection of datasets and simulation frameworks that are commonly used to benchmark visual robot-navigation algorithms. We will not go into detail on datasets that do not allow actual exploration, such as the work\cite{141}, where the agent can only move along a predetermined path or datasets and frameworks specifically meant for non-grounded robots and outdoor environments (such as AirSim and Flightmare, developed mainly for drone navigation\cite{17}), but we will list simulators that can be used to train grounded and non-grounded agents, see\cite{142} who used Gazebo for drone acrobatics.

3.5.1. Datasets based on the real world

The advantage of using photos or videos of the real world are obviously their high fidelity and the resulting small simulation-to-real-gap to a potential real world setting. Unfortunately, as discussed, real world data acquisition is costly and concessions in terms of continuous movement have to be made as even a small environment will result in a very big dataset with potentially redundant samples if very fine grained motion is allowed. An alternative are 3D scans (or 3D reconstructions based on photos and videos) that can be traversed freely. Another advantage of these 3D models is the potential to modify the model by adding or removing elements and - depending on the simulator - the option of different lighting/camera/movement configurations. Indeed, the robot should ideally be able to handle all these different conditions and a dataset that only consists of single images might not be able to account for this, or at least only with costly augmentation. Something to keep in mind with 3D scans is that they are not yet perfect when low-cost devices are used to acquire them. While for example the Matterport3D\cite{8} and Gibson\cite{129} datasets are widely used, many publications mention that only a subset of the available scans is actually chosen for training and testing because reconstruction artefacts can lead to unrealistic behaviour or the simulated robot simply getting stuck. The publication that introduced the Habitat simulator\cite{11} has an entire paragraph on this issue in the appendix - the authors manually evaluated every scene in the Gibson dataset and did not use those with significant reconstruction artifacts or overall bad quality. Lastly, both 3D scans and photos do not come with segmentation information so in case this is a task that should be trained for, the dataset will have to be annotated manually.

3.5.2. Artificially generated environments

In contrast to the previously mentioned datasets and frameworks, completely artificial environments are built from scratch by a human designer/programmer\cite{123}\cite{143}. This makes their creation more laborious but allows to avoid the reconstruction artefacts and inaccuracies of 3D scans. Another positive is that since every object in the scene is known, it is a lot easier to generate semantic maps for segmentation tasks. One very impressive example of this type of dataset is Replica\cite{143}, which is completely handcrafted but very detailed and near photorealistic, including reflections and material properties. Fortunately researchers are not the only demographic that is interested in high-fidelity simulated environments and computer games have advanced to the point that they can be used to reliably train pretrain algorithms (see the previously mentioned work by Richter et al.\cite{136} that utilizes the game Grand Theft Auto as a simulator for human movement in a traffic setting). However,
even the most intricate handmade environment will still differ from the reality, just in a different way that environments based on photos or videos of the real world do.

**A selection of Datasets and Simulation Frameworks**

We will differentiate between datasets in the “classic” sense, that consist of discrete data points with or without an underlying structure, and simulation frameworks that allow continuous movement by rendering the sensor (camera) input for a certain position and orientation on demand. For the purpose of highlighting this difference we will use the term “framework” whenever an underlying 3D environment is used to generate input data for an agent, although the authors themselves might refer to their work as “dataset”. In contrast, the term “dataset” will refer to a collection of images. Some names will come up in both, since it is possible to generate discrete data points from any continuous environment. Figure 6 provides a first impression of the quality of images that can be generated with simulators, showcasing examples from AI Habitat, Gibson and Replica. Tables 2 and 3 summarize the main properties of the datasets and frameworks respectively. An interesting “in-between” approach has been proposed by Rosano et al. [82], who used a 3D model of a room in habitat for continuous training, as well as samples from the real room that were registered with the 3D model to provide realistic input.

All the datasets and simulators are free, but some require an application process. We will not be listing the SUNCG [144] dataset because, even though it has had an impact on the development of robot navigation algorithms, it is no longer available due to a lawsuit of Planner 5D against Princeton and Facebook.

**Classic datasets**

**Matterport3D.** The Matterport3D [8] dataset is a RGB-D dataset that covers 90 building-sized scenes. The aim of the dataset is to provide seamless view transitions and to this end the dataset includes 10,800 panoramic images generated from nearly 200,000 RGB-D images. It is annotated with segmentation information (2D and 3D) and camera poses, so it can be used for active vision applications as well. According to the authors, the dataset is meant to further research in the topics of keypoint matching, prediction of surface normals, semantic segmentation and region classification. Although the “original” version of this dataset is comprised of individual samples with an underlying structure, the 3D scans that are used to sample the images can be used for simulation frameworks as well.

**Active Vision dataset.** The Active Vision dataset [9] like the Matterport 3D dataset, is aimed at navigation research in indoor environments and is based on the real world as well. The dataset covers 9 (later extended to 15) unique scenes and contains 20,000+ RGB-D (later extended to 30,000) images with 50,000+ (later 70,000) bounding box object annotations. In their original publication [9], the authors demonstrate how the dataset can be used to benchmark algorithms for object instance detection, as well as active vision and navigation.

**Simulation frameworks**

**UnrealCV.** UnrealCV [145] uses the Unreal game engine and provides researchers with the option to build new environments and interact with them. It can be used in conjunction with deep learning frameworks and to generate ground truth data from a virtual scene. The project website [24] provides documentation and tutorials, including the integration into OpenAI Gym, a framework for Reinforcement Learning.

**MORSE.** MORSE [146] is a simulator for academic robotics that focuses on 3D simulation of a variety of environments (small, big, indoor, outdoor). It can simulate one or several autonomous agents and can be accessed with the console or python scripts. It provides a variety of sensors such as cameras, range sensors, GPS and several motion setups. MORSE is based on the Blender engine for visualisation and the Bullet library for physics.

**Gazebo.** The Gazebo framework [25] provides a multitude of features such as test environments, physics simulations, noise models and ready-made virtual representations of real robots (it also supports communication with ROS). On the associated Github entry [26], the authors (and the community) offer a selection of the previously mentioned features - this variety in combination with functional features such as cloud support, cross-platform support, command line tools and remote training support makes Gazebo suitable for researchers that are interested in the physical and practical behaviour of robots. As a result, this framework is often used for “low-level” physics-related research and not so much for “high-level” conceptual research.

| Dataset                  | Data format               | Photorealistic | Interaction with the environment | Size                                                                  |
|--------------------------|---------------------------|----------------|---------------------------------|-----------------------------------------------------------------------|
| Matterport3D             | RGB-D images              | yes            | no                              | 10,800 panoramic views of 90 multi-floor houses containing 2000 room regions |
| Active Vision dataset    | RGB-D images with object bounding boxes | yes            | no                              | 30,000+ RGB-D images, 15 scenes                                        |

Table 2: Table of the classic datasets for robot navigation (samples with underlying structure, no simulation interface).
**Interactive Gibson.** The Gibson framework, also called iGibson, to highlight its interactive nature, focuses on interactive navigation. In contrast to previous navigation frameworks, this framework allows the robot to physically interact with the scene and with objects in it. Examples of interactions are: opening drawers or fridges, picking up small items such as fruit and moving slightly bigger objects such as chairs. In terms of size, the framework contains 572 buildings and 1400 floors, as well as several virtual representations of popular robot models (iGibson also supports multi-agent settings).

The framework also provides 3D annotations of the object classes chairs, desks, doors, sofas, and tables. The authors of [129], [124] introduce the framework in combination with the novel task of navigating in cluttered environments. They present several novel metrics to evaluate the combination of navigation and interaction and apply these metrics to several navigation baselines. This interactive focus sets this framework apart from the others - although some of them have recently started to add interaction options, this framework was especially designed for interaction tasks.

**Replica.** Replica [143] is a framework consisting of 18 handmade photo-realistic rooms. The rooms have high-resolution textures with reflection information and class annotations. It was designed for the training of egocentric agents that relies on realistic environments with as few artifacts and simulation biases as possible. Given the information contained in the scenes, Replica can also be used to train algorithms for semantic segmentation in 2D and 3D. The framework comes with a lightweight C++ renderer interface for training. However, given the affiliation of the authors it is not surprising that even at its introduction, Replica was already designed to be used in Habitat [11].

**3RScan and RIO10.** 3RScan [7] is a framework for 3D object instance-localization, as in estimating the pose of one or several objects in the same scene that changes over time (for example a living room where books are carried from point A to point B). The authors of [147] consider this task to be particularly relevant for real world applications - for example when a robot is asked to fetch a specific object that has changed location. The 3RScan framework contains 1482 RGB-D scans of 478 indoor scenes in varying states. For every scan, some objects in the scene have changed position. The scenes are annotated with object class segmentations, as well as the ground truth object position correspondences across the scans. The authors demonstrate a first possible solution to this matching problem with a convolutional 3D correspondence network. Although this framework is primarily meant to explore re-localisation, the high-quality 3D scenes can also be used for active vision and navigation. In the year after 3RScan was introduced, the authors posed an even harder problem with the corresponding RIO10 framework [148]. The purpose of the second framework is to evaluate algorithms that are able to perform re-localisation tasks in changing indoor scenes. In addition to changing objects in the scene, this framework also offers realistic illumination variations. These frameworks explore use cases that are very hard to solve but essential for the deployment of embodied agents in the real world.

**Habitat.** Habitat [11] is a simulation framework developed and maintained by Facebook, with the aim to push the development of embodied navigation. It can be used to evaluate algorithms for active perception, robot navigation and limited interaction scenarios. The framework consists of a simulator and an API that handles the communication between python and the simulator. The repositories support existing 3D scans from Matterport3D, Replica and Gibson as well as custom environments. In addition, the authors provide baseline implementations of reinforcement learning algorithms such as Actor-Critic [108] and PPO [149] as well as DDPPO [150].

**AI2 Thor.** The House Of inteRactions (THOR) [10] is a framework for visual AI research - namely embodied agents. It contains 30 varied photo-realistic indoor scenes and supports interaction with objects in the scene. AI2 Thor is intended as a starting point for behaviour tasks such as deep reinforcement learning, scene interaction, imitation learning and more classic computer vision tasks such as object detection and segmentation. An extension of AI2 Thor is RoboThor [125], a subset of the original dataset with corresponding real world rooms. The idea is to train robots in simulation and verify their performance in the real world. Another subset is iThor, a set of 120 rooms with interactive objects. The focus on the sim-to-real transfer and the corresponding challenge sets THOR apart from other options.

---

3RScan: https://waldjohannau.github.io/3rscan/
RIO10: https://waldjohannau.github.io/RIO10/
Habitat: https://aihabitat.org/
AI2 Thor: https://ai2thor.allenai.org/
RoboThor: https://ai2thor.allenai.org/robothor/challenge/

---

Image 6: Examples of images generated by Habitat using the demo scenes that come with the simulator, the Replica dataset and the Gibson dataset. The scenes are photorealistic and can be used to create rendered images from every possible viewpoint.
Table 3: Table of the simulation frameworks for robot navigation.

| Dataset            | data format | photorealistic | interaction with the environment | size            |
|--------------------|-------------|----------------|----------------------------------|-----------------|
| IteratedC         | simulation framework | yes     | no | several available AI environments, custom environments |
| MORSE             | simulation framework | no     | no | several available environments |
| Gazebo            | simulation framework | no     | only between objects in the scene | several available environments, contains creation GUI |
| Interactive Gibson| simulation framework | yes     | yes | 372 buildings, 1400 floors |
| Replox            | simulation framework | yes     | no in native SDK, yes in AI Habitat | 18 rooms |
| 3RScan            | simulation framework | yes     | no | 1492 RGB-D scans of 478 environments |
| AI Habitat        | simulation framework | yes     | yes (some physics support) | supports Replica, Matterport, Gibson and others |
| AI2 Thor          | simulation framework | yes     | yes | 120 rooms (89 apartments for RoboThor) |
| Nvidia Isaac Sim 2020 | simulation framework | yes     | yes | supports custom environments |
| Unity3D mlagents  | simulation framework | yes     | yes | supports custom environments |

NVIDIA Isaac NavSim. NVIDIA Isaac NavSim is a simulator developed and maintained by NVIDIA that aims at utilizing their GPU functionalities for the simulated training of robots. It is based on Unity3D and provides a variety of test environments as well as tools for automated data generation. NVIDIA Isaac NavSim uses the Isaac SDK that was created specifically to interface between the algorithmic and simulation side (Unity3D, tensorflow) and the used NVIDIA hardware.

The desired audience for this simulator is researchers and the industry sector. Possible use cases are: warehouse interactions (stacking cases, moving objects) and the direct deployment on physical robots. A similar NVIDIA product is the recent NVIDIA Drive Constellation system for the development of algorithms for self-driving vehicles.

Unity3D mlagents. Unity3D mlagents is a plugin for the Unity3D game engine that is aimed at game developers who want to include artificial intelligence into their games and at researchers who want to build custom environments (Figure 7). It provides the Unity3D functionality for designing and building environments that can be augmented with 3D assets and an interface for setting up Reinforcement Learning agents. It supports different input modalities such as range sensors and images. It is designed primarily for game developers, so the focus is on the generation of the environment, which has made mlagents an entry point into the topic.

4. Human Robot Interaction

Ever since robots have started to act and move, researchers have been interested in studying their interaction with humans, both in social contexts or in work settings where robots and humans have to coexist or even cooperate in the same space. Human Robot Interaction (HRI) as a separate field is relatively young, one of the earliest conferences being the 1992 IEEE International Symposium on Robot and Human Interactive Communication (RoMan). The field is historically multidisciplinary: researchers from robotics, psychology, natural language processing, cognitive sciences and even philosophy work together and the results are accordingly diverse. Given the increased automation and use of technology in everyday activities, HRI is and will be an important aspect of daily life. In this survey, we will give a brief introduction of the topic but will not be focusing on the ethical implications (see [151] for a relatively recent survey) or purely psychological work.

4.1. Social Interaction

The field of social robot interaction studies how humans and robots interact in a natural, “human-like” manner. Researchers from this field theorize that, depending on the application, humans prefer to interact with robots that learn and display certain social skills [152]. The long term goal is to create robots that can understand human behaviour, including verbal and nonverbal cues, social conventions and emotional signals and - in order to be intuitively understood by humans - use these tools to communicate information in turn. This dimension of intelligence is considered to be a crucial aspect in designing robots that are a useful and comfortable partner.

This way of looking at intelligence runs counter to many paradigms in Machine Learning and Artificial Intelligence, that define intelligence by measurable benchmarks: winning a game of chess (and recently, a game of go [153]) against human players, classifying images [32], navigating efficiently [11] or answering factual questions [88]. Indeed, social interaction abilities are difficult to evaluate and often subjective. For more comprehensive reading on the topic, we suggest the work by Dautenhahn et al. [154], [155] and Breazeal [156], which provide conceptual overviews. Recent works have proposed meth-
ods for intention-forecasting in human-robot interaction \cite{157} and a reinforcement learning strategy to learn specific social behaviour such as handshakes \cite{158}. On the more holistic side, Triebel et al. \cite{159} presented a robot that is able to assist, inform and guide pedestrians in an airport environment. Their system used a combination of detection and tracking, behaviour and social hierarchy analysis and social mapping to generate socially compliant paths for the robot.

4.2. Interaction in an Industrial Context

Any work environment that has humans interacting with robots has to evaluate the potential for both significant increase in productivity as well as danger \cite{160}. With increased automation, especially in assembly lines, robots are working in closer contact and on more complex tasks with human workers.\footnote{42} Early work by Lenz et al. \cite{161} presented a concept for a collaborative workspace, whereas Albu-Schäffer et al. suggested designing a robot specifically for collaborative work \cite{162}. For further reading we point to the recent survey by Tsarouchi et al. \cite{163}, who review research and progress on HRI in the manufacturing/production environment and the recent work by Villani et al. \cite{31}.

4.3. Interaction in the Context of Home Assistance and Healthcare

Robots are of big interest in the domain of healthcare and could potentially improve a patient’s convalescence in the hospital and at home \cite{164} (see Figure 8). They can be used for observation tasks, assist both doctors and patients and fill care gaps in a time when the average age in developed countries is rising and the healthcare sector is already struggling to find enough qualified staff \cite{165}. A robotic companion or “nurse” can collect information useful for doctors, healthcare workers and the patients themselves \cite{166}. The work by Erickson et al. \cite{167} tackles both of the tasks of dressing and bathing with an armed robot that can support the patient in real time. Other works proposed robot support in the kitchen \cite{168} and robot support for aging citizens in general \cite{169}.

COVID-19. During the COVID-19 pandemic, robots have been immensely helpful because they can support frontline healthcare workers and reduce the transmission risk. They can also help with auxiliary support tasks such as cleaning and logistics, as well as reconnaissance tasks (for example compliance monitoring during quarantine). The survey by Shen et al. \cite{170} provides an overview of robotic systems that have been used during the pandemic, from both an industrial and an academic point of view. Given the infectiousness of the virus, it is of vital importance to either use protective gear or avoid contact altogether. Robots can help with both problems. They can monitor patients, provide certain diagnostic features such as taking a temperature, conveying messages and providing instructions. As a result, less healthcare workers were exposed to the virus and less PPE (Personal Protective Equipment) was necessary.\footnote{43} Yang et al. \cite{171} propose and test a telerobotic system, that combined medical expertise with the adherence to “social distancing” guidelines. In their review, Seidita et al. \cite{172} examined published works and projects in the areas of clinical care, logistics and reconnaissance.

On the logistics side, robots can deliver protective material such as PPE, medication or potentially dangerous waste. All of these problems can be tackled with classic goal-oriented navigation strategies. Robots can also disinfect contaminated areas to relieve cleaning personnel and reduce infection.\footnote{44} Similarly to standard cleaning robots, those robots can employ coverage-based exploration algorithms.

On the topic of reconnaissance, robots have been used to monitor patients during quarantine and alleviated isolation in cases where in-person visits would have posed a risk \cite{172}.

The Covid-19 epidemic of 2020 has also emphasized the role of preventive healthcare \cite{167} especially in people who are at a high risk of developing a serious condition and the potential for robots that monitor compliant behaviour, for example social distancing \cite{173}.

\footnote{42}https://secondhands.eu/
\footnote{43}https://edition.cnn.com/2020/11/11/tech/robots-india-covid-spc-intl/index.html
\footnote{44}https://www.lgnewsroom.com/2020/12/lg-announces-autonomous-robot-with-disinfecting-uv-light-for-various-b2b-applications/
\footnote{45}https://www.phs-uv.com/en/
diseases, which is often considered the largest healthcare problem in the world. Poor lifestyle behaviour is a common factor for the development of various common chronic diseases such as Type II diabetes, and chronic heart disease, and behavior change has been show to be the most effective prevention of these conditions. A companion robot can have an important role in motivating the patient to exercise [174] [175]. In home healthcare patients are typically either acutely ill but do not require hospitalization, in rehabilitation after a critical episode, or chronically ill. A robotic assistant may play many roles in these conditions [166].

In the case of patients with impairments in terms of movement, cognitive abilities, perception, communication or strength, a robot could help by assisting, clarifying, translating, reminding etc.. Unfortunately, advancements in terms of hardware and software are not yet at the point where robots can be seamlessly integrated into a patient’s hospital stay or convalescence at home, but the first promising steps have been taken. The overview by [176] provides a good entry point into the topic of healthcare, identifying tasks, involved parties, challenges and opportunities in this relatively young field. Something to keep in mind in this area is that, even if robots can provably benefit the patient, they still have to be accepted [177] and properly used [178] by the patients. The works of Kawamura et al. [179] and [180] and Fischinger et al. [181] have explored the potential of robots as companions or assistants for the elderly, while the authors of [182] and [183] propose solutions for disabled patients in the form of mounted wheelchair arms and prosthetics and orthotics, respectively.

An interesting subset of research on robots in the context of healthcare is concerned with autistic tendencies. Research has shown that many children with autism spectrum disorders (ASD) respond well to robots and find them easier to communicate with than other humans [184]. A large body of research has covered potential reasons for this effect and how to exploit it to assist children [185], [186], [187]. We especially recommend the works by Scassellati et al. [184] and Cabibihan et al. [188] in this context.

4.4. Socially aware Navigation for natural Human Robot Interaction

For a robot to naturally navigate among and interact with humans, it has to not only not endanger humans but also obey a variety of explicit and implicit rules. Some of these rules are very specific, such as a minimal distance to every human to avoid collisions and guarantee a sense of comfort, some are less well-defined such as social conventions (who goes first at an impasse, what side of a path is used) and “natural” movement. Humans prefer to interact with robots that abide by familiar rules of movement and whose behaviour can be predicted intuitively. Kruse et al. [29] summarize that the requirements that social navigation strategies have to fulfill are 1. comfort, as in the humans should not be stressed or inconvenienced 2. naturalness, as in the robot should behave similar to humans and 3. sociability, as in the robot should follow cultural conventions.

Depending on the underlying task of the robot such as navigating to a certain point, moving with one or several humans or engaging, different solutions have been proposed. A general caveat of all research in this area is that the evaluation of algorithms for Socially aware Navigation is very difficult given its subjective nature. The problem definitions and the resulting datasets vary a lot and given the potential danger, one strategy that is frequently employed is to train on simulated episodes and perform a qualitative evaluation afterwards with real humans. In that case the results will naturally depend a lot on the used simulation framework and the test subjects.

Proxemics. A big part of guaranteeing behaviour that is perceived as “safe” is controlling how quickly and how close to the human the robot operates [4]. The area of research that deals with the amount of space humans like to keep between themselves and others is called Proxemics and is of particular relevance for robots that operate in social situations or in healthcare contexts. Of course this area depends on the individual as well as the social and cultural context, but it is an aspect of human-robot-cohabitation that has to be kept in mind to ensure that the robot can execute its task smoothly, without inconveniencing the human, see Figure 9. We will present a select few works that explore proxemics for robots, the overall field is much broader and leans heavily into social and cultural studies.

The work by Henkel et al. [5] evaluates different distance strategies by how they affect the human’s perception of the robot’s likeability, intelligence and submissiveness, Kim et al. [6] looked at the influence of an underlying task (cooperation vs competition) on the perception. Takayama et al. explored the causes for the difference in perception (such as previous experience with robots or pets) [189], Obaid et al. [190] conducted posture experiments and Mead et al. [41] focused on the design implications of social conventions for human-robot interactions. While the previous works tended to focus on one-on-one interaction with the robot, Mavrogiannis et al. [42] ex-
explored the effects on a crowd of people. Pacchierotti et al. [191] presented a proxemics-based control strategy for hallway navigation and Koay et al. [192] explored proxemic preferences for handover tasks. Papadakis et al. [193] model proxemic zones depending on human-human interactions to inform their path-planning strategy.

5. Analyzing and Modeling Human Behaviour

For a robot to naturally interact with humans it is crucial to equip them with the ability to not only detect them but also to infer what the human is doing. Different activities might indicate that movement is imminent, that the human does not want to be disturbed or - depending on the task of the robot - that an intervention is necessary. Activity recognition has been evaluated in the context of Human Robot Interaction [194] as well as the contexts of socially aware path planning and navigation and healthcare - see Farie et al. [195] and their follow-up [196]. In the rest of this section we will highlight more general Solutions and mention instances where activity recognition is used to inform robot behaviour in the later subsection on Socially aware Navigation using Activity Recognition.

The problem of Human Activity Recognition (HAR) is widely studied [197], due to its varied applications such as surveillance [198], healthcare [199] and - most relevant for this survey - human-robot interaction [200]. Video-based activity recognition is very appealing since data acquisition is relatively cheap and unlabeled data is easy to come by online. Labelling the data, on the other hand, tends to be expensive. Given its complex nature, HAR tends to combine the newest advances in classic computer vision algorithms such as background modeling, motion estimation (optical flow) and stereo vision with more complex machine learning advances such as human pose detection, object-detector neural networks, relation graphs and recurrent neural networks (since actions have a temporal dimension). For a detailed overview on the topic, we recommend Moeslund et al. [201] for the early 2000s and Zhang et al. [202] for more recent works.

Human movement has a very hierarchical nature that has to be taken into account when evaluating it for certain goals. At the lowest level, humans perform motion primitives [202] such as moving a hand, turning the torso etc. and these primitives are executed sequentially or in parallel to perform a certain action (in some publications a difference is made between actions and activities, we will use them interchangeably in this survey). A set of actions - again in sequence or in parallel - generates behaviour. A concrete example would be: moving the hand to grab a tomato, coring the complete tomato, cooking a bowl of soup. As per the example, it is not enough to just evaluate the parts of the human that are in motion or still, one also has to take into account what objects are interacted with, either directly (opening a drawer) or indirectly (running after a ball in a game of soccer). Further complexities are introduced with multi-person activities [203] and multi-modal activity recognition (e.g. audio-visual activity recognition [204]). We will give a brief introduction into core concepts and topics that are of interest to current activity recognition research. Human behaviour understanding is mostly performed on videos shot from a third person perspective, where most or all parts of the human body are in frame, such as the Kinetics [205] dataset, the ActivityNet [206] dataset and the Charades [207] dataset. Other datasets have also focused on first person vision, such as EPIC-KITCHENS [208] and the Georgia Tech Egocentric Activities [209] dataset. A robot can use onboard sensors and wearable sensors attached to the human as well as other third-position sensors.

5.1. Pose Estimation

The first step in many Human Behaviour understanding pipelines is to estimate where the human is in the frame and how their limbs are positioned, see Figure 10. This has the dual benefits of not only extracting the most relevant information for further evaluation, but also significantly compress the information in the frame, which makes further processing easier. Context information such as backgrounds/scenes and or objects can be extracted separately. Although Pose Estimation is well-explored to the point that robust cross-platform solutions such as OpenPose [47] exist, it remains a well investigated task. Indeed, recent works focus on further refining the exact 2D or 3D position of human joints in image (e.g. OpenPosed [48], https://github.com/mmu-Perceptual-Computing-Lab/openpose) and video data (e.g. OpenPose [49]). Depending on the desired goal, it might also be beneficial to focus only on selected limbs such as the hand, see [204], who used Fuzzy Clustering and Takagi Sugeno (TS) modeling to recognize and imitate grasping behaviours. More recent examples use texture [223], common sense rules and external information such as objects [225] and scene information to finetune further. The recent detectron2 [49] library, developed and maintained by Facebook, includes several pose estimation modules as well - it was designed and is used as a starting point for more complex evaluations such as action recognition.

Figure 10: Main goals of a 2D Human Pose Estimation approach. Starting from the RGB image as input (left - photo by Joanna Nix Walkup on unsplash), algorithms can segment relevant body parts (center) and infer the 2D skeleton of the human considering a given set of joints (right).
Since robots (and other surveillance applications) will have varying viewpoints and might not always see the observed human in the “optimal” way. Indeed, it is of great interest for researchers to develop algorithms that can handle incomplete and occluded input.

5.2. Action Recognition

Action recognition from RGB video or images is a very popular research topic that has been tackled in a variety of ways. Just like in other areas of computer vision, early solutions employed low-level vision features such as motion, contours and keypoints [227] while more recent work tends to favor Deep Learning architectures [228].

Due to the complex spatiotemporal nature of human action (and its frequent overlap), researchers rarely attempt to solve the overall action recognition problem and instead break down the problem into specific subtasks such as temporal localisation [229], multiview action recognition [230] or co-occurrence [231], [232]. Alternatively, some researchers focus on speed [233], robustness [234] or particular feature encodings [235], [236]. As mentioned in the previous section, pose estimation is a standard preprocessing step and several researchers have presented algorithms that can infer the type of action based on 2D or 3D joint positions [237], [238], [239], [240], [241], [242]. In addition to skeleton data, human action is also determined by the type of objects that are interacted with [243], [245]. [246]. For more detailed overviews we recommend the surveys by Zhang et al. for “classic” solutions before 2011 [26] and the more current follow ups by Herath et al. [27] and Zhang et al. [28].

5.3. Action Forecasting

Another important ability for robots to understand human behaviour is the prediction of future actions, be it from past trajectories or videos, still frames or audio. Being able to anticipate the next action of a human has a wide range of use cases, it is relevant for applications such as automotive safety [247] as well as observation tasks for healthcare, surveillance or social robots [248], [249]. It is also a very conceptually relevant topic, since any algorithm that can successfully predict the next motion necessarily needs an understanding of how human action is structured and how it might intervene with secondary actions, other agents or objects. Another useful task in this context is trajectory forecasting, which we will touch on later - in some cases the distinction is hard to make, since walking trajectories can be considered actions, like in the early work by Kitani et al. [250]. We will list some successful solutions to the different iterations of this problem.

Huang et al. [251] look at human interaction, whereas Gao et al. [252] train a Reinforced Encoder-Decoder (RED) network to predict possible future sequences. In contrast to the majority of works that focuses on the immediate future actions happening in a few seconds, Farha et al. [253] present an algorithm suited for long term prediction.

A specific subtask of action forecasting is first-person action prediction, which relies on egocentric input. Figure 11 gives an idea of the possible applications. While the majority of current algorithms and evaluations are geared towards third person action recognition and anticipation, the increasing popularity of head-mounted smart devices such as Google Glass50 and HoloLens51 combined with the demand for personalised entertainment, support in industry and healthcare applications, could place more emphasis on first-person research in the future. Recent work by Rhinehart et al. [254] uses inverse reinforcement learning, while the approach by Furnari et al. [255] utilizes LSTMs to predict future actions in a domestic setting.

5.4. Socially aware Navigation using Activity Recognition

The inclusion of Activity information can be implemented in very basic ways such as slightly improved proxemics-based collision avoidance but it can also include fine-grained activity classes that demand specific behaviour from the robot (for example in a healthcare context [195]). The survey by Martinez et al. [256] gives an overview of human-aware robot navigation that touches on Proxemics, social conventions and formal definitions for Activity and Affordance Spaces. They suggest that, in addition to avoiding collisions with humans performing certain activities, socially aware robots also have to take into account both Activity Spaces (where an activity is performed) as well as Affordance Spaces (which denote potential Activity spaces, for example a sign than can be looked at), see Figure 12.

An example for a comparatively low-level approach was given by Banisetti et al. [257], who focused on ways the interpersonal distance can be used to infer certain activities and in turn how this distance can be used by the robot to replicate natural behaviour. Robot path planning can also use the information about the performed activity to simply determine what areas should not be traversed by the robot. These areas can be estimated by combining the position of the human and the action that is performed [258] (see also [256] for the definition of Activity and Affordance Spaces) as well as the shape of the interpersonal zone that corresponds to certain actions [259]. The action running, for example, corresponds to

50https://www.google.com/glass/start/
51https://www.microsoft.com/en-us/hololens
an elongated forward-facing interpersonal zone that should not be blocked by the robot, see also Figure 9. Charalampous et al. 260 on the other hand use a more complex approach where human activity is determined by a deep learning module. The information about the performed activity is combined with rule-based proxemics to continuously plan and re-plan the best collision-free path for the robot. The authors also took care to evaluate the algorithm in a human-free setting. An approach focused on group activities was proposed by Okal et al. 261. See Table 4 for an overview.

Activity recognition for robot navigation. Given the specific requirements of robotic sensing, Activity recognition has been tackled specifically in this context, with Gori et al. 262, who used RGB(D) data to encode visual, velocity and distance features and an SVM and Olatunji 263, who employed motion capture data and a Convolutional Neural Network to determine the performed activity. An application from the area of domestic robots that includes fine grained Activity Recognition can be found in the work of Piyathilaka et al. 264, who use 3D skeleton features to estimate what action a human is performing to accordingly adjust the behaviour of a robot. The example they give is that a cleaning robot would be able to use information about the current human activity to plan what area to clean next. Rezaazadegan et al. 265 suggested that standard Activity Recognition approaches have to be modified to account for background cues and camera motion before they work on a robot. To this end, they generate action region proposals and then extract shape and motion features to classify the activity with a Convolutional Neural Network. Faria et al. 195 used Spatio-temporal 3D skeleton features to infer not only the performed activity but also potentially dangerous situations such as falls for robot-assisted living.

5.5. Datasets

In this section we will cover some of the more popular datasets for pose estimation and action recognition. The datasets each have different scopes and unique properties. We will first give a brief overview of each dataset and later summarize the main characteristics in a table for quick comparison. We will group the datasets roughly by the main purpose (Human Pose Estimation and Action Recognition) and whether or not they are based on recordings of real humans.

| Publication | Problem definition as written in the paper | Underlying algorithm | Used dataset | Used metrics | Tests with real humans |
|-------------|-------------------------------------------|----------------------|--------------|-------------|-----------------------|
| 258 | Appropriate behavior for a mobile robot | Gaussian Mixture Model | proprietary dataset | accuracy of activity prediction | yes |
| 259 | Human-aware navigation | Deep Learning, Probabilistic reasoning, random tree path planners, adaptive band algorithm | simulation (RoboComp) | time and path length to target, human activity interruption | yes |
| 260 | Social navigation for escorting and following | Deep Learning, rule-based decision making and topological maps | proprietary dataset | fraction of safe trajectories | yes |
| 261 | Navigation in a human-populated environment | SLAM and Dynamic Bayesian mixture model | proprietary dataset | Precision and Recall (Activity Recognition) | yes |
| 262 | Activity recognition for navigation in human sounds | Histograms with velocity features | simulator | Accuracy (Activity Recognition) | no |

Table 4: Comparison of the Algorithms that use Activity Recognition for Socially aware Navigation. It is apparent that there is no consistent task definition, dataset or evaluation metric.

5.5.1. Datasets for Human Pose Estimation

Table 5 summarizes the main datasets for Human Pose Estimation.

Human3.6M Dataset. The Human3.6M Dataset 266 is currently the biggest dataset of human 3D poses collected with motion capture technology and video. It contains 3.6 million 3D human poses and the corresponding videos, as well as 3D meshes of the actors that performed the actions (11 professional actors - 6 male, 5 female). The dataset covers daily and mostly stationary indoor activities such as talking on the phone, smoking, discussing, taking photos etc. The authors introduce the dataset in 266 for the training and evaluation human pose estimation models and algorithms. In addition they also provide code for baseline pose prediction algorithms, visualisation and evaluation tools on their website. Three example poses can be seen in Figure 13.

HumanEva. This dataset is similar to the Human 3.6M dataset, but smaller in size. It is a combination of two datasets, namely HumanEva I and HumanEva II, and like Human3.6M it focuses on indoor environments. 267 The dataset consists of RGB/grayscale videos with synchronized joint positions acquired with motion capture technology. The 7 video sequences contain 4 subjects performing 6 actions such as walking and gesturing.

http://vision.imar.ro/human3.6m/description.php
http://humaneva.is.tue.mpg.de/
Total Capture Dataset. The Total Capture dataset contains around 1.9M frames of indoor multi-view video, IMU and motion capture sensor data. The recordings were made in a very controlled setting using a black motion capture suit and in an empty indoor volume, with 8 cameras, capturing synchronized video streams. The aim of the dataset is to benchmark 3D pose estimating using IMU and multiview cameras. The ground truth poses are provided by the motion capture. In contrast to previous datasets that focus mostly on everyday activities, this dataset also includes more challenging and one-person activities performed by the 5 subjects such as yoga, crawling and bending over.

SURREAL (Synthetic hUmans foR REAL tasks) Dataset. This dataset is meant to provide synthetic data, which can be used to augment real recordings and contains rendered images created from 3D sequences of motion capture data and varying backgrounds. Given how difficult it is to collect and label high quality large scale data for human pose estimation, the dataset can be used to combine random poses, lighting conditions, camera positions and backgrounds - making the most out of existing data (the poses are sampled from the previously mentioned Human3.6M dataset). The authors took several measures to ensure that the generated images are still as realistic as possible. They used the SPML model from [270] to create realistic human shapes. Their results demonstrate how the generated dataset improves 2D pose estimation and depth estimation on real images.

JTA Dataset. The JTA (Joint Track Auto) dataset was created for pose estimation and tracking of pedestrians in urban scenarios. It is rendered using the video game Grand Theft Auto V, a photo-realistic action-adventure game that takes place in an urban environment. The dataset contains 512 full-HD videos (split into training and testing 50/50), each 30 seconds long, recorded at 30 fps. Since the scene is rendered from a known 3D model, synchronized reliable 3d and 2d joint information is automatically available for every video. In addition to the dataset, the authors provide a tool to generate additional videos.

GTA-IM Dataset. Like the JTA dataset, the GTA Indoor Motion dataset (GTA-IM) has been created with the video game Grad Theft Auto V. In contrast to JTA, however, its main concern is with indoor human-scene interaction such walking through rooms with furniture, ascending stairs and walking through doorways. The HD RGB-D videos are synchronized with the 3D joint positions and camera pose annotations. The “actors” in the scenes, as well as the scenes themselves and the actions performed in them are diverse. The creators of this dataset use it to predict how a human will move given an observed trajectory and they find that their results outperform previous approaches both in terms of quantity and quality.

5.5.2. Datasets for Activity Recognition

We review a selection of the main datasets here. For a more complete list of datasets, we suggest the overview website by the University of Bonn which also contains result metrics from state-of-the-art-methods on the listed datasets. The table summarizes the main characteristics of the discussed datasets.

| Dataset       | main research purpose                                                                 | Data format                                                                 | Annotation format                      | size            | synthetic or real recordings |
|---------------|----------------------------------------------------------------------------------------|----------------------------------------------------------------------------|----------------------------------------|----------------|-----------------------------|
| Human3.6M     | pose estimation                                                                       | videos, joint positions, joint angles, meshes                              | activity class, ground truth human pose | 3.6 million 3D poses | recordings of real humans   |
| HumanEva(2009)| pose estimation                                                                       | videos, joint positions                                                    | activity class, ground truth human pose | 400,000 frames  | recordings of real humans   |
| Total Capture | pose estimation                                                                       | videos and joint positions                                                 | activity class, ground truth human pose | 5.9 million frames | recordings of real humans   |
| SURREAL       | pose estimation                                                                        | rgb video frames                                                           | ground truth human pose, segmentations | 5.9 million frames    | synthetic and recordings of real humans |
| JTA           | pose estimation                                                                        | rgb video frames                                                           | ground truth human pose                | 256 minutes      | synthetic                    |
| GTA-IM        | pose estimation and trajectory forecasting                                             | rgb video frames                                                           | ground truth human pose, segmentations | 1 million frames, 119 clips | synthetic                    |
| CMU Panoptic  | activity recognition and pose estimation                                              | video, 3D pose and face landmarks                                         | activity annotations, speaker id, ground truth human pose | 65 sequences (5.5 hours) | recordings of real humans   |

https://cvssp.org/projects/totalcapture/TotalCapture/ https://www.di.ens.fr/willow/research/surreal/data/ https://github.com/fabbrimatteo/JTA-Dataset

https://github.com/2heC/GTA-IM-Dataset https://actionrecognition.net/index.html https://deepmind.com/research/open-source/kinetics

Table 5: Datasets for human pose estimation
to download the respective videos. As a result, the actual size of the dataset varies with the chosen video quality and with the availability of each annotated video on youtube. The advantage of this dataset is its impressive size, whereas a drawback is the non-trivial downloading method.

**Carnegie Mellon University Multimodal Activity (CMU-MMAC) Database**. The CMU-MMAC database is a multimodal dataset focused on kitchen-related activities. It contains footage of activities such as making brownies, pizza, sandwich, salad and scrambled eggs. The authors used 6 RGB-D cameras with varying spatial and temporal resolutions, one of these cameras is worn by the actor and hence provides a first person view. Additional modalities include sound and motion capture. Although this dataset is relatively small, the varied input modalities differentiate it from other datasets.

**ActivityNet**. ActivityNet contains videos of varying human activities, ranging from simple to complex. It covers 203 activity classes with around 140 videos per class and often more than one action per video. The annotations contain global information about the videos, as well as start and stop information for certain activities. The videos are sourced from the internet and annotated using Amazon Mechanical Turk. All in all, the dataset consist of 849 hours of annotated footage. The website provides an introduction, instructions for usage and first results on the dataset for video classification, trimmed activity classification as well as activity detection. The annotations are provided in a hierarchical structure (for example playing sports - discus throw). Like the Kinetics dataset, ActivityNet is sourced from youtube videos that have to be downloaded by the user - the authors provide instructions and code to this end, as well as an evaluation server. Just like with the Kinetics dataset, downloading the dataset is non trivial, but it provides a representative sample size.

**Charades and Charades-Ego**. Charades and Charades-Ego are datasets of daily activities, recorded from a first and, in the case of Charades-Ego, also a third person perspective. For the creation of these datasets, the actors (through Amazon Mechanical Turk) were asked to film themselves performing a certain activity (similar to the game charades, hence the name). The results are varied datasets of 9848 videos with 66,500 temporal annotations for 157 action classes (Charades) and 7860 videos with 68,536 temporal annotations for 157 action classes (Charades-Ego). Since the videos were recorded by the actors themselves, these datasets are very diverse, but the setup of the data acquisition process (acting out something) makes them not as natural as others.

**Cornell Activity Datasets**. The CAD-60 and CAD-120 dataset consist of 60 and 120 RGB-D video sequences of four humans performing activities. The available modalities are RGB video, depths information and tracked skeletons. They were recorded with a Kinect sensor and cover activities (such as brushing teeth, cooking, relaxing), each containing sub-activities. The four different actors were recorded in different daily indoor environments, such as a kitchen, living room, office, etc. The inclusion of depth and skeleton data sets these datasets apart from the previous ones and makes them very interesting also for pose estimation, although they are much smaller.

**EPIC-KITCHENS**. EPIC-Kitchens-100 is the extension of the existing EPIC-KITCHENS-55 (for 55 hours), a dataset focusing on egocentric videos of daily activities that occur in the kitchen, such as cooking, making drinks, cleaning etc. (see Figure 14). The actions are non-scripted, performed in the actors’ homes and densely annotated via narration, temporal action segments and object bounding boxes. The extended dataset contains 100 hours of footage with 90,000 action segments, 20,000 unique narrations, 90 verb classes and 300 noun classes. It was published in conjunction with set of challenges to push the research in the area of egocentric vision. The variety, size and scope of this dataset make it very attractive for researchers who are specifically focused on egocentric vision. While other datasets contain some egocentric videos, this dataset is geared completely towards it.

**Extended Georgia Tech Egocentric Activity Gaze+ Dataset**. Georgia Tech offers several datasets for egocentric activity recognition, the most recent one is EGTEA Gaze+. This dataset consists of HD videos with the corresponding audio, gaze trajectories, hand mask segmentations and frame-level annotations for the activities that are performed. The 32 subjects in the dataset performed cooking activities in 86 sessions, amounting to 28 hours of footage. In total, the dataset contains...
10325 action instances and 15,176 hand masks on 13,847 video frames.

**ADL (Activities of Daily Living) Dataset.** This dataset of activities was recorded with a chest-mounted camera that recorded HD videos of 18 daily living activities performed by 20 actors. The actions were not described in detail to ensure a varied dataset and each recording is around 30 minutes, resulting in a total over 10 hours. The videos are annotated with the 18 action labels, as well as object bounding boxes (42 classes), object ids and human-object interactions. As a result, this dataset is suitable for egocentric activity recognition, forecasting and, to a lesser extent, object tracking.

**CMU Panoptics dataset.** The CMU Panoptics dataset focuses on multi-view recordings of multi-person activities to study the 3D motion of groups engaging in social interactions. It has been collected with the aptly named “Massively Multi-view System”, which contains 480 VGA camera views, 30+ HD views and 10 RGB-D sensors. In addition to the video streams, the system also provides the 3D body poses and 3D facial landmarks as well as transcripts of the performed activities including speaker ID for instances where the actors talked. This dataset occupies an interesting niche as there is no other recording setup that provides similarly detailed multi-view and 3D information.

**CNTU RGB+D 120.** The CNTU RGB+D 120 dataset is an extension of the CNTU RGB+D dataset, a large-scale dataset for RGB+D human action recognition. It contains 106 subjects performing actions for more than 114 thousand video samples and 8 million frames. Among the 120 action classes are drinking water, eating a meal, brushing teeth and hair etc. The authors introduced the dataset in CVPR 2016 (first iteration) and later released an extended version in TPAMI 2019 and use it to showcase their action recognition framework.

**JPL-Interaction dataset.** The JPL First-Person Interaction dataset is geared towards researchers that are interested in egocentric social activity. It consists of human activity videos taken from the viewpoint of one person in a multi-person group. The focus is on first-person interactive activities such as shaking hands, hugging and waving. The footage was captured by outfitting the observer (a teddy bear) with a gopro and having others interact. To remove the influence of humanoid walking patterns, the subject with the camera was placed on wheels and pushed - similarly to how a wheeled robot would move. It is one of the few datasets in this list that tackle the notoriously subjective and hard to evaluate area of social interaction.

| Dataset | research purpose | data format | annotation format | size |
|---------|-----------------|-------------|------------------|------|
| ADL     | activity and pose estimation | video | activity with start and stop | around 1000 hours, over 620000 videos |
| CMU-MMAC | activity and pose estimation | video | activity with start and stop | videos of 13 actors making 5 recipes |
| UCF-EGO | activity and pose estimation | video | activity with start and stop | around 940 hours, about 28000 videos |
| Cornell Activity | activity recognition | video with depth | activity segments, skeletons | 180 videos together |
| EPIC-Kitchens-100 | activity recognition | video | spoken annotations, marks and bounding boxes | 100 hours of footage |
| JPL-Interaction dataset | activity recognition | video with depth | activity class, hand masks and audio | 700 hours of footage |
| ADL     | activity recognition | video | activity class, objects with id and interaction | 28 hours of footage |
| CMU Panoptics | activity recognition | video | activity class, objects with id and interaction | 55 sequences (5.5 hours) |
| CNTU RGB+D 120 | activity recognition | video | activity class, objects with id and interaction | 114 videos, 120 classes, 106 subjects |
| JPL-Interaction dataset | activity recognition | video | activity class, objects with id and interaction | 84 videos |

Table 6: Datasets for activity recognition.

### 5.6. Trajectory forecasting

A lot of early work on social robot navigation among humans has focused on trajectory forecasting and achieved impressive results because even when the robot does not necessarily interact with humans, it is still perceived as socially capable as long as it does not “get in the way”. In addition, trajectory prediction is crucial for autonomous vehicles and other safety applications. Alahi et al. presented a forecasting model based on LSTMs, Kim et al. use Kalman filters and Maximum Likelihood estimations and Xiao et al. use a pretrained SVM to group activity classes, predict the trajectories based on those classes and test the result in a real kitchen. Solutions that use deep learning and reinforcement learning have been suggested by Perez et al. and Fahad et al. respectively. The work by Pfeiffer et al. takes cluttered environments into consideration and the approach by Chung et al. builds a spatial behavior cognition model to model how human trajectories influence each other. The recent work by Vemula et al. proposed an attention-based model that utilizes recurrent neural networks. For more in-depth reading we recommend the survey by Rudenko et al. who have also extended the idea of human trajectory prediction to the problem of occupancy prediction. Instead of predicting individual trajectories, they infer pedestrian motion from semantic information about the environment.

#### 5.6.1. Socially aware Navigation using Trajectory Analysis

Early work by Henry et al. uses simulated crowds to generate human-like motion behaviour while proposed a Trajectory planning algorithm based on Rapidly-exploring Random Trees that was tested in simulation. The work by
We will discuss a select few datasets that are used to benchmark trajectory prediction algorithms - our focus will be on the prediction of pedestrian movement, see Table 8. Another research sector that is very interested in trajectory prediction is the automotive sector. Examples for those types of datasets that contain cars, buses, cyclists and pedestrians are Argoverse\(^7\) in Lyft Level 5\(^8\) and TRAF\(^7\). The focus of this survey, as previously mentioned, is visual perception and as a result we will not go into details on datasets created with Lidar or other 3D sensors, such as the very big and busy shopping center dataset by Brsic et al.\(^4\).

We recommend this github repository\(^5\) for an overview that includes traffic and non-traffic setups as well as the TrajNet challenge/benchmark\(^6\), that combines a number of the mentioned datasets and offers tools for evaluations across datasets. Something to keep in mind with these datasets is that the authors rarely obtain the informed consent of the people in the videos and as a result they might not be available in the future (see for example the Oxford Town Centre dataset\(^7\), and that in most cases the perspective from which the data is collected is not a potential agent’s egocentric view (most datasets are captured from a very high vantage point).

**ETH dataset.** The ETH BIWI Walking Pedestrians (EWAP)\(^7\) dataset is comprised of videos showing pedestrians in unconstrained situations, shot from a high vantage point. It was recorded at 25 fps and manually annotated at 2.5 fps, totalling 650 human trajectories over 25 minutes. The dataset contains videos from two different vantage points and was used for the ICCV 2009 paper\(^5\) that presents a solution to the task of multi-person tracking in crowded environments.

**Crowds UCY dataset.** This dataset contains aerial view videos of everyday crowd scenarios. The videos show students, shoppers and pedestrians in general and are shot at 2.5 fps. The dataset shows the trajectories of roughly 200 pedestrians and was presented by Lerner et al. in 2007\(^5\). Although this dataset is not very new, it remains one of the more popular ones and is often used in conjunction with the ETH dataset.

**Train Station Dataset.** The Train Station dataset or Grand Central Station dataset was presented by Zhou et al. in 2018\(^5\) and contains, as the name suggests, recordings of pedestrians moving freely at the Grand Central Station in New York. The surveillance video was taken from above and shows over 33 minutes at 25 fps, at a resolution of 720x480. It contains key-point trajectories for each of the recorded pedestrians. Although it is small, the specific environment and the number of pedestrians make it interesting for research on areas with high foot traffic.

**Edinburgh Informatics Forum Pedestrian Database.** This dataset\(^6\) consists of “compressed” trajectories, observed by a

| Publication | Problem definition as written in the paper | Underlying algorithm | Used dataset | Used metrics | Test with real humans |
|-------------|------------------------------------------|----------------------|--------------|--------------|----------------------|
| 2023        | Human-aware navigation                   | MDP and Social Force Model | Edinburgh Informatics Forum Pedestrian Database | Success (position goal reached) | no |
| 2024        | Socially-aware navigation among people   | Social Force Model | Edinburgh Informatics Forum Pedestrian Database | Average number of accidents | no |
| 2025        | Modeling of pedestrian motion            | Social Force Model | proprietary simulation | Percentage of successful navigation | yes |
| 2026        | Socially-compliant navigation            | Social Force Model | collected from experiments | Minimal robot-pedestrian distance and travel time from start to goal | yes |
| 2027        | Collision avoidance                      | Social Force Model | simulation (Netlogo) | Average number of accidents | no |
| 2028        | Trajectory planning for a robot in dynamic human environments | Social Force Model | proprietary simulation | Percentage of successful navigation (no collision) | no |
| 2029        | Prediction of human motion in populated space | Social Force Model | Synthetic trajectories | Negative Log-Probability and modified Hausdorff Distance | no |

Table 7: Comparison of the Algorithms that use Trajectory Modeling for Socially aware Navigation.

O’Callaghan\(^294\), in contrast, does not aim to explicitly create human-like trajectories but rather to obey a set of social rules and the generated trajectories are human-like as a result. Luber et al.\(^295\) learn motion prototypes from extracted ground truth behaviour and demonstrate superiority over a purely proxemics-based baseline. The work by Ferrer et al.\(^296\) focuses more on modeling human-agent interaction with the popular social force model, introduced by Helbing et al.\(^297\), and human feedback. A similar direction was pursued by Shiomi et al.\(^298\) when avoiding robot-human collisions. Tai et al.\(^299\) used GANs for behaviour cloning and Riaz et al.\(^300\) proposed a collision avoidance system for automotive applications that is based on human behaviour. Rudenko et al.\(^301\) used a MDP (Markov Decision Process) algorithm with a Social Force model that emphasised the influence of group motion over individual motion. See Table 8 for a comparison.

5.6.2. Datasets for Trajectory Prediction

We will discuss a select few datasets that are used to benchmark trajectory prediction algorithms - our focus will be on the prediction of pedestrian movement, see Table 8. Another research sector that is very interested in trajectory prediction is the automotive sector. Examples for those types of datasets that
top down camera at the University of Edinburgh over a timespan of several months. The dataset does not contain video frames like the other datasets, instead the trajectories are described by the position of the bounding box marking the detection and a histogram representation of the detected pedestrian. Given the large timeframe, the authors [303] managed to extract more than 7 million pedestrian detections and as a result the dataset contains more than 90,000 trajectories. The viewpoint of the dataset is particularly interesting as it covers not only the free interaction between pedestrians but also the behaviour at building entrances.

**Stanford Drone Dataset.** The aim of this dataset [302] is to showcase the interaction of pedestrians, skateboarders, cars, buses, and golf carts in various environments on the Stanford Campus. The authors [302] recorded 8 different scenes from above to explore how social rules and common sense influences human trajectories. It contains 60 videos with annotations for the mentioned classes and is currently one of the bigger datasets that are available for trajectory forecasting.

**PETS 2009 Dataset.** This dataset [301] was presented by Ferryman et al. [310] in 2010 as part of a workshop on pedestrian detection, counting and trajectory estimation as well as crowd event detection. Given the different tasks that the dataset is concerned with, it is divided into subsets for specific problem settings. One subset is meant for general training, one is for benchmarking person count and density estimation, another is aimed at pedestrian tracking and the third is for evaluating event recognition and flow analysis algorithms. The dataset consists of RGB videos of students and pedestrians moving across the Whiteknights Campus, University of Reading, taken from an elevated view.

| Dataset                | research purpose                                                                 | annotations                                                                 | participants | size                                                                 | perspective of recording |
|------------------------|----------------------------------------------------------------------------------|----------------------------------------------------------------------------|--------------|----------------------------------------------------------------------|--------------------------|
| ETH Pedestrian Dataset  | multi-person tracking in crowded environments                                     | pedestrian IDs                                                            | pedestrians  | 600 trajectories over 25 minutes                                    | from above               |
| Crowd UCL Datasets      | crowded simulation                                                               | pedestrian IDs                                                            | pedestrians  | 200 pedestrian trajectories                                         | from above               |
| Clinic Campus Dataset   | crowded behaviour analysis                                                        | key-point trajectories                                                    | individuals  | 25 minutes of footage                                                | from above               |
| Edinburgh Informatics   | pedestrian behaviour in open space and around entries                           | compressed trajectories                                                   | pedestrians  | 90,000 trajectories                                                  | from above               |
| Forum Pedestrian Dataset| interaction of various traffic participants                                       | bounding boxes on the videos                                              | traffic participants (pedestrians, skateboarders, cars etc.) | 60 videos with 930,000 annotated frames and 20,000 detected traffic participants | from above               |
| PETS 2009 Dataset       | pedestrian detection, trajectory estimation, crowd event detection               | depending on the subset: detection of pedestrians, crowd density and person count | pedestrians  | 1 GB of 566 GB RGB video                                              | from above               |

Table 8: Datasets for trajectory forecasting.

6. Discussion

We will highlight observed trends or research gaps in this section. In many cases, the underlying algorithms or feature representations already exist, but they are not yet widely used to enrich algorithms for Socially Aware Robot Navigation.

**Formalised Evaluation Strategies.** Unlike goal-oriented Robot Navigation or Trajectory Prediction, Socially aware Robot Navigation is difficult to evaluate in terms of objective metrics. Not only does the robot have to perform a certain task, but the positive or negative effect it has on the humans sharing the space has to be taken into account as well. Often this is done with small scale human studies, where the users are asked to rate the behaviour in terms of subjective qualities such as naturalness or comfort [30]. Without formalised metrics for social behaviour, different algorithms are hard to compare but the implementation of such a metric is borderline impossible due to the inherently subjective nature of the task. An alternative might be not to evaluate if a robot behaves overall socially compliant, but whether it can perform basic subtasks that are required in daily interactions - just like the clearly defined subtasks of PointGoal Navigation and Coverage based Exploration are used to gauge whether a robot can navigate intelligently.

Possible subtasks for Socially aware robot Navigation might be Human Following [119], handover tasks [312], continuous observation or Embodied Question asking with spoken instructions from a human, see Figure 15. The task of human following in particular is a very specific and self-contained subset of the overall topic but it is one of the most useful and comparatively easy to evaluate. Especially in settings where robots are used as transport vehicles to unburden the human user, it might be beneficial to implement a following strategy.

Petrovic et al. [313] present a solution based on stereo vision and Kalman filters that is capable of following the user in indoor and outdoor scenarios. A solution with cameras in addition to laser range sensors was presented by Shimizu et al. [314], who achieve markerless results in indoor settings. In contrast, Kit Tsun et al. [311] use a depth camera, an active marker and proximity sensors as well as simulated environments. Two approaches using reinforcement learning have been proposed by Wang et al. [315] and Bayomi et al. [316], with the former learning the correct social distance when interacting with the human and the latter learning about the human’s goal so the robot can navigate efficiently even when the human makes detours.

A more formal definition and shared understanding of the Human Following task (for example: defined distance to the human, defined input modalities) and corresponding metrics (for example: percentage of steps with the human in “frame” for visual input, percentage of steps at the correct distance, penalty for invading the person’s space) could elevate this task to a shared benchmark.

**Interaction via Language.** Language is already used for general Human Robot Interaction [314], but it could also be used...
to convey instructions or additional information for Socially Aware Robot Navigation. Recent research on Robot Navigation has seen an uptick in language-guided Navigation \cite{319,319}, since it allows for complex semantic task descriptions. The focus of those works are usually on the conversion of the written language task into a concrete goal formulation for the robot and this scope could be easily extended to a version where the human is in the loop. This would also entail strategies for good information transfer (for example: good listening positions, conveying attention). Similarly, Embodied Question Answering \cite{88}, another well-defined subtask of Robot Navigation, could be adapted to have a human pose the question verbally.

**Gesture and Emotion Recognition.** Gesture Recognition is used in the context of HRI \cite{320} but gestures are also a natural way to enrich navigation behaviours \cite{321}, for example the nonverbal communication of “after you” at bottlenecks or “come here” gestures. An adjacent topic is the integration of emotional cues into Human Robot Interaction \cite{322} and potentially robot navigation. Narayanan et al. \cite{323} inferred overall mood from gaits, which were used to adapt the robot’s path and Bera et al. \cite{324} used trajectories and faces, as did Gines et al. \cite{325}. Yelwande et al. \cite{326} compared different path planning strategies that utilise information about the human’s emotional state.

**Datasets.** Socially Aware Robot navigation is a difficult topic to create realistic datasets for, since they have to encompass the rich input of an embodied agent, the variations of human behaviour and the fact that the behaviour of the robot itself will affect the humans around it. In addition, one dataset will not necessarily fit robots that are different in size and shape, since this also influences how humans react, which makes recording reliable training data a difficult task. Simulated datasets on the other hand are widely used, but they will always suffer bias in terms of the representation of the world as well as the human behaviour. Nevertheless, we have seen the first promising datasets that are explicitly created to model human-aware robot navigation such as the dataset by Tolani et al. \cite{327} or the very recent SEAN environment \cite{328}. However both are still very new and not widely used yet. Another recent work by Baghel et al. \cite{329} also proposes the design of simulation frameworks specifically for social navigation, although it focuses more on trajectory simulation than realistic visual input. Similarly, Puig et al. \cite{330} presented a framework for modeling human activities with the aim of simulating daily task and potentially training collaborative behaviour \cite{331}. All of this is indicative of the desire to pursue this area of research, even when bigger and more influential frameworks such as AI Habitat and AI2 Thor do not support human actors yet. However, a recent announcement from the team of the iGibson simulator suggests that Socially aware Navigation will be of interest to them: starting with the iGibson Challenge 2021, the robot can be trained to navigate in simulated crowds\cite{82}.

In summary, we have covered motivations for socially-aware robot navigation, ranging from industrial contexts, such as assembly and packing, to healthcare and social companionship. A brief introduction of the Active Vision paradigm was given, since it is the basis for most of the highlighted research and for Robot Navigation in general. We also looked into the current and previous trends for Robot Navigation that are most relevant for socially-aware applications. In addition, the topics Human Behaviour Modeling and Analysis and Human Robot Interaction were briefly introduced, again with a focus on works that are relevant for our topic. The last section on Human-aware Motion Planning in particular has highlighted how different areas of research can come together naturally to model behaviour that is suitable for a human environment.

We have found a variety of promising approaches and datasets that aim to solve the complex problem of socially-aware Robot Navigation, but we have also observed that at the current point in time, there are few works that focus specifically on this challenging topic. Especially with the current popularity of simulation-based learning, a variety of datasets and frameworks exist that could in theory be augmented to include humans. Instead, most approaches are confined to only Robot Navigation without humans, Human Robot Interaction without robot movement etc.

We hope that the overview given in this survey will serve as inspiration to other researchers for more interdisciplinary research and we look forward to future work in this topic.

7. Acknowledgements

This research has been supported by Marie Skłodowska-Curie Innovative Training Networks - European Industrial Doctorates - PhilHumans Project (Contract no. 812882, Web: \url{http://www.philhumans.eu}, MIUR AIM - Attrazione e Mobilita Internazionale Linea 1 - AIM1893589 -

\footnote{https://sean.interactive-machines.com/setup.html} \footnote{http://svl.stanford.edu/igibson/challenge.html}
References

[1] M. Ford, The Rise of The Robots, Oneworld Publications, UK, 2015.

[2] M. L. Knapp, J. A. Hall, T. G. Horgan, Non-Verbal Communication in Human Interaction Cengage Learning, 2013. URL https://books.google.de/books?id=gTh23RlMwCoC

[3] M. Mori, K. MacDorman, N. Kageki, The Uncanny Valley [From the Field], IEEE Robotics & Automation Magazine 19 (2012) 98–100. doi:10.1109/MRA.2012.2192811

[4] E. Paccheriotti, H. Christensen, P. Jensfält, Evaluation of Passing Distance for Social Robots, in: The 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN’06), 2006, pp. 315–320. doi:10.1109/RO-MAN.2006.314430

[5] Z. Henkel, C. Bethel, R. Murphy, V. Srinivasan, Evaluation of Proxemic Scaling Functions for Social Robotics, Human-Machine Systems, IEEE Transactions on 44 (2014) 374–385. doi:10.1109/TMHMS.2014.2304075

[6] Y. Kim, B. Mulh, How social distance shapes human–robot interaction, International Journal of Human-Computer Studies 72 (2014) 783–795. doi:10.1016/j.ijhcs.2014.05.005

[7] P. Anderson, A. X. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva, A. R. Zamir, On Evaluation of Embodied Navigation Agents, ArXiv abs/1807.06757 (2018).

[8] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, Y. Zhang, Matterport3D: Learning from RGB-D Data in Indoor Environments, International Conference on 3D Vision (3DV) (2017).

[9] P. Ammirato, P. Poiron, E. Park, J. Kosecka, A. C. Berg, A Dataset for Developing and Benchmarking Active Vision, 2017. doi:10.1109/IJCAI.2017.7985816

[10] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, E. Topol, Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again, 1st Edition, Basic Books, Inc., USA, 2019.

[11] M. Mori, K. MacDorman, N. Kageki, The Uncanny Valley [From the Field], IEEE Robotics & Automation Magazine 19 (2012) 98–100. doi:10.1109/MRA.2012.2192811

[12] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, E. Topol, Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again, 1st Edition, Basic Books, Inc., USA, 2019.

[13] P. Ammirato, P. Poiron, E. Park, J. Kosecka, A. C. Berg, A Dataset for Developing and Benchmarking Active Vision, 2017. doi:10.1109/IJCAI.2017.7985816

[14] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, E. Topol, Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again, 1st Edition, Basic Books, Inc., USA, 2019.

[15] M. Mori, K. MacDorman, N. Kageki, The Uncanny Valley [From the Field], IEEE Robotics & Automation Magazine 19 (2012) 98–100. doi:10.1109/MRA.2012.2192811

[16] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, E. Topol, Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again, 1st Edition, Basic Books, Inc., USA, 2019.

[17] M. Mori, K. MacDorman, N. Kageki, The Uncanny Valley [From the Field], IEEE Robotics & Automation Magazine 19 (2012) 98–100. doi:10.1109/MRA.2012.2192811

[18] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, E. Topol, Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again, 1st Edition, Basic Books, Inc., USA, 2019.

[19] M. Mori, K. MacDorman, N. Kageki, The Uncanny Valley [From the Field], IEEE Robotics & Automation Magazine 19 (2012) 98–100. doi:10.1109/MRA.2012.2192811

[20] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, E. Topol, Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again, 1st Edition, Basic Books, Inc., USA, 2019.

[21] M. Mori, K. MacDorman, N. Kageki, The Uncanny Valley [From the Field], IEEE Robotics & Automation Magazine 19 (2012) 98–100. doi:10.1109/MRA.2012.2192811

[22] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, E. Topol, Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again, 1st Edition, Basic Books, Inc., USA, 2019.

[23] M. Mori, K. MacDorman, N. Kageki, The Uncanny Valley [From the Field], IEEE Robotics & Automation Magazine 19 (2012) 98–100. doi:10.1109/MRA.2012.2192811

[24] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, E. Topol, Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again, 1st Edition, Basic Books, Inc., USA, 2019.

[25] M. Mori, K. MacDorman, N. Kageki, The Uncanny Valley [From the Field], IEEE Robotics & Automation Magazine 19 (2012) 98–100. doi:10.1109/MRA.2012.2192811

[26] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, E. Topol, Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again, 1st Edition, Basic Books, Inc., USA, 2019.

[27] M. Mori, K. MacDorman, N. Kageki, The Uncanny Valley [From the Field], IEEE Robotics & Automation Magazine 19 (2012) 98–100. doi:10.1109/MRA.2012.2192811

[28] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, E. Topol, Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again, 1st Edition, Basic Books, Inc., USA, 2019.
Y. F. Chen, M. Everett, M. Liu, J. How, Socially aware motion planning.

N. Akalin, A. Loutfi, Reinforcement Learning Approaches in Social Robotics (2020).

J. Lukasiewicz, On three-valued logic, Ruch Filozoficzny 5 (170-171) (1920).

D. Gordon, A. Kembhavi, M. Rastegari, J. Redmon, D. Fox, A. Farhadi, A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, D. Batra, Embodied Question Answering in Photorealistic Environments With Point Cloud Perception, 2019, pp. 6652–6661.

M. Montemerlo, S. Thrun, D. Koller, B. Wegbreit, FastSLAM 2.0: An Improved Particle Filtering Algorithm for Simultaneous Localization and Mapping that Provably Converges, Proceedings of IJCAI International Joint Conference on Artificial Intelligence (2003).

R. Isermann, On fuzzy logic applications for automatic control, supervision, and fault diagnosis, IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 28 (2) (1998) 221–235.

Y. Y. Lin, M. Martinez, Fuzzy tools in recommender systems: A survey, International Journal of Fuzzy Systems Applications 4 (3) (2017) 257–276.

J. Xiao, K. A. Ehinger, A. Oliva, A. Torralba, Recognizing scene viewpoints from viewpoint-specific training sets, Computer Vision and Pattern Recognition (2012) 2695–2702.

M. Deitke, W. Han, A. Herrasti, A. Kembhavi, E. Kolve, R. Mottaghi, R. Martín-Martín, S. Savarese, Interactive Gibson: A Benchmark for Interactive Navigation in Cluttered Environments, IEEE Robotics and Autonomous Systems (2021).
in: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI ’15, Association for Computing Machinery, New York, NY, USA, 2015, p. 295–302. doi:10.1145/2696454.2696462

[249] H. Koppula, A. Saxena, Anticipating Human Activities Using Object Affordances for Reactive Robotic Response, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (2015) 1–1. doi:10.1109/TPAMI.2015.2430335

[250] K. Kitani, B. Ziebart, J. Bagnell, M. Hebert, Activity Forecasting, in: Proceedings of the 12th European conference on Computer Vision, VCIP 2012, 2012, pp. 201–214. doi:10.1007/978-3-642-33765-9_15

[251] D.-A. Huang, K. Kitani, Action-Reaction: Forecasting the Dynamics of Human Interaction, in: Computer Vision (ECCV), Springer International Publishing, Cham, 2014, p. 489–504.

[252] J. Gao, Z. Yang, R. Nevatia, RED: Reinforced Encoder-Decoder Networks for Action Anticipation, in: Proceedings of the British Machine Vision Conference (BMVC), BMVA Press, 2017, pp. 92–1–92.12. doi:10.5244/C.31.92

[253] N. Rhinehart, K. Kitani, First-Person Activity Forecasting with Online Inverse Reinforcement Learning, in: IEEE International Conference on Computer Vision (ICCV), 2017. doi:10.1109/ICCV.2017.399

[254] A. Furnari, G. Farinella, What Would You Expect? Anticipating Ego-centric Activities With Rolling-Unrolling LSTMs and Modality Attention, 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019) 6251–6260.

[255] J. Rio-Martínez, A. Spalanzani, C. Laugier, From proxemics theory to socially-aware navigation: a survey, International Journal of Social Robotics 7 (2014) 137–153. doi:10.1007/s12369-014-0251-1

[256] J. Lim, H. Pirsiavash, D. Ramanan, Detecting Activities of Daily Living in First-Person Camera Views, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018) 5431–5438.

[257] J. Song, C. Ponce, B. Selman, A. Saxena, Human Activity Detection from RGBD Images, in: Proceedings of the 16th AAAI Conference on Plan, Activity, and Intent Recognition, AAAI’WS-11-16, AAAI Press, 2011, p. 47–55.

[258] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, Y. Sheikh, Panoptic Studio: A Massively Multiview System for Social Interaction Capture, IEEE Transactions on Pattern Analysis and Machine Intelligence PP (2016).

[259] F. J. Liu, A. Shahroudy, J. Liu, T. Ng, G. Wang, NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding, IEEE transactions on pattern analysis and machine intelligence (2019).

[260] A. Shahroudy, J. Liu, T. Ng, G. Wang, NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1010–1019. doi:10.1109/CVPR.2016.115

[261] M. Ryoo, L. Matthies, First-Person Activity Recognition: What Are They Doing To Me?, in: Proceedings/ CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2012, pp. 263–270. doi:10.1109/TPAMI.2017.282743

[262] P. Beltran, Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) Database, in: Tech. report CMU-RI-TR-08-22, Robotics Institute, Carnegie Mellon University, 2008.

[263] I. Olatunji, Human activity recognition for mobile robots, in: Conference on Intelligent Robots and Systems, IROS, 2014.

[264] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, M. J. Black, Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape for a Single Image, in: European Conference on Computer Vision (ECCV), 2016.

[265] M. Fabbi, F. Lanzi, S. Calderara, A. Palazzi, R. Vezzani, R. Cucchiara, Learning to Detect and Track Visible and Occluded Body Joints in a Virtual World, in: European Conference on Computer Vision, 2018.

[266] Z. Cao, H. Gao, K. Mangalam, Q. Cai, M. Vo, J. Malik, Long-term human motion prediction with scene context, in: European Conference on Computer Vision ECCV, 2020.

[267] F. Torre, J. Hodgins, A. W. Bargteil, X. Martin, J. Macey, A. Collado, P. Beltran, Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) Database, in: Tech. report CMU-RI-TR-08-22, Robotics Institute, Carnegie Mellon University, 2008.

[268] A. Shahroudy, J. Liu, T. Ng, G. Wang, NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding, IEEE transactions on pattern analysis and machine intelligence (2019).

[269] S. Savarese, Social LSTM: Human Trajectory Prediction in Crowded Space, in: Conference on Computer Vision (ECCV), 2018.

[270] Y. Li, M. Liu, J. Rehg, In the Eye of Beholder: Joint Learning of Gaze and Actions in First Person Video: 15th European Conference on Computer Vision, Munich, Germany, September 8–14, 2018, Proceedings, Part II, 2018, pp. 639–655. doi:10.1007/978-3-030-01228-1_38

[271] H. Pirsiavash, D. Ramanan, Detecting Activities of Daily Living in First-Person Camera Views, in: Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2012, pp. 263–270. doi:10.1109/TPAMI.2017.282743

[272] A. Shahroudy, J. Liu, T. Ng, G. Wang, NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1010–1019. doi:10.1109/CVPR.2016.115

[273] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, S. Savarese, Social LSTM: Human Trajectory Prediction in Crowded Spaces, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 961–971. doi:10.1109/CVPR.2016.110

[274] S. Kim, S. Guy, W. Liu, R. Lau, M. Lin, D. Manocha, Predicting Pedestrian Trajectories Using Velocity-Space Reasoning, The International Journal of Robotics Research 34 (2014). doi:10.1177/027836491555543

[275] S. Xiao, Z. Wang, J. Folkesson, Unsupervised robot learning to predict person motion, Proceedings - IEEE International Conference on Robotics and Automation 2015 (2015) 691–696. doi:10.1109/ICRA.
[324] A. Bera, T. Randhavane, R. Prinja, K. Kapsaskis, A. Wang, K. Gray, D. Manocha, How are you feeling? multimodal emotion learning for socially-assistive robot navigation, in: IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), 2020, pp. 644–651. doi:10.1109/FG47880.2020.00141

[325] J. Ginès, F. Martín, D. Vargas, F. J. Rodríguez, V. Matellán, Social navigation in a cognitive architecture using dynamic proxemic zones, Sensors 19 (23) (2019) 5189.

[326] S. S. Yelwande, R. M. Dandavate, Study of emotion recognition models for socially aware robots and subsequent path mapping, in: 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1230–1233. doi:10.1109/ICECA49313.2020.9297630

[327] V. Tolani, S. Bansal, A. Faust, C. Tomlin, Visual Navigation Among Humans with Optimal Control as a Supervisor, 2020, eprint: 2003.09354.

[328] N. Tsoi, M. Hussein, J. Espinosa, X. Ruiz, M. Vázquez, SEAN: Social Environment for Autonomous Navigation, in: Proceedings of the 8th International Conference on Human-Agent Interaction, HAI ’20, Association for Computing Machinery, New York, NY, USA, 2020, p. 281–283. doi:10.1145/3406499.3418760 URL https://doi.org/10.1145/3406499.3418760

[329] R. Baghel, A. Kapoor, P. Rachiller, R. R. Jorvekar, D. Rodriguez-Criado, L. J. Manso, A Toolkit to Generate Social Navigation Datasets, Springer, 2020, pp. 80–193.

[330] X. Puig, K. Ra, M. Boben, J. Li, T. Wang, S. Fidler, A. Torralba, Virtual Home: Simulating household activities via programs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8494–8502.

[331] X. Puig, T. Shu, S. Li, Z. Wang, J. B. Tenenbaum, S. Fidler, A. Torralba, Watch-And-Help: A Challenge for Social Perception and Human-AI Collaboration in: International Conference on Learning Representations, 2021. URL https://openreview.net/forum?id=w_7JMpGZRh0