An Experimental Protocol for Benchmarking Robotic Indoor Navigation

Christoph Sprunk[1*], Jörg Röwekämper[1*], Gershon Parent[2*], Luciano Spinello[1], Gian Diego Tipaldi[1], Wolfram Burgard[1], and Mihai Jalobeanu[2]

[1] Department of Computer Science, University of Freiburg, Germany
[2] Microsoft Robotics, Microsoft Corporation, USA

Abstract. Robot navigation is one of the most studied problems in robotics and the key capability for robot autonomy. Navigation techniques have become more and more reliable, but evaluation mainly focused on individual navigation components (i.e., mapping, localization, and planning) using datasets or simulations. The goal of this paper is to define an experimental protocol to evaluate the whole navigation system, deployed in a real environment. To ensure repeatability and reproducibility of experiments, our benchmark protocol provides detailed definitions and controls the environment dynamics. We define standardized environments and introduce the concept of a reference robot to allow comparison between different navigation systems at different experimentation sites. We present applications of our protocol in experiments in two different research groups, showing the usefulness of the benchmark.

Keywords: benchmark, autonomous navigation, indoor robots, dynamic environments

1 Introduction

Robot navigation is a widely studied topic in robotics due to its cornerstone function for robot autonomy. Prior work on benchmarking robot navigation primarily focused on simultaneous localization and mapping (SLAM) techniques, and in particular on assessing the accuracy of the generated maps [4, 20]. These evaluations are useful when the robot task is to compute a precise map, e.g., for architectural or other surveying purposes. However, when the map is built for autonomous navigation, its metric accuracy does not necessarily relate to the performance of the robot. A robot navigating in a real-world environment must be able to localize and reach destinations in environments that are populated with dynamic objects and that are changed with respect to the initial conditions. This includes environments shared with people or environments where objects may be moved around.

In this paper, we formulate an experimental protocol for benchmarking robot navigation. This fills the void of a missing evaluation method for repeatable,

[*] C. Sprunk, J. Röwekämper and G. Parent contributed equally to this work.

reproducible and comparable tests for autonomous indoor navigation consisting of performance metrics, methodology and baseline. We aim at accommodating for hardware differences between comparable solutions and for differences in sensors. In particular, we aim at reproducing identical environments, including environment dynamics between multiple runs at an experimentation site.

This paper represents the first time that navigation is quantified in a fashion similar to other hard sciences where environmental conditions are key for reproducibility and fair comparison. In other computer science disciplines, such as computer vision and machine learning, benchmarks had a large impact to standardize and to uniform evaluation procedures [1, 11]. Differently from these sciences, robot navigation cannot be evaluated only with datasets. The robot is immersed in the environment and interacts with it. For this reason, we provide to the community ways of measuring ground truth and suggest a reference robot.

In our benchmark, we aim to compute statistics about *a simulated year* of continuous robot operation. For this, we provide detailed definitions for the experimental environment and conditions. The experimental setup consists of definitions about the size, the dynamics, the environmental conditions and the overall duration of an experiment. This includes the number and the size of the rooms, the number of people walking in the scene, the kinds and amounts of objects and furniture that are moved and the number of goals for each environment. As reference robot, we selected the widespread commercial platform Pioneer P3-DX. We applied the benchmarking protocol to conduct experiments in two different research groups by using two different kinds of robots, showing the usefulness of the benchmark. The complete benchmark protocol along with detailed instructions and our evaluation software is publicly available at `http://research.microsoft.com/brin/`.

2 Related Work

Benchmarking plays an important role for comparison and evaluation in science. In particular, there are many benchmarking works in several fields related to robotics, including machine learning, computer vision and artificial intelligence. Machine learning is probably the field that received most attention, thanks to the use of very large evaluation datasets for different tasks [1, 2, 16]. Similarly, computer vision has many procedures and benchmarks available [6, 9, 11, 17].

Despite being one of the most studied field in robotics, there is a relatively small amount of literature related to benchmarking robot navigation. This is probably caused by the fact that robot navigation cannot be evaluated on a dataset. The robot navigates in a dynamic environment that is constantly changing. In NaviGates [14], the authors present an early benchmark for robot navigation. Here, they concentrate on robot skills and architecture but they do not take in account how to systematically evaluate the robot performance in a changing environment. Gutmann *et al.* [12] presented a set of extensive experiments evaluating the accuracy and robustness of localization systems using datasets. Calisi *et al.* [5] propose a benchmark framework that concentrates only on the evalua-

tion of vehicle motion algorithms. Borenstein and Feng [3] introduce a method for measuring odometry errors of mobile robots. Specifically, it focuses on quantitative evaluation of systematic and non systematic errors. The work of Nowack *et al.* [15] presents an investigation for an evaluation of two specific robot tasks, namely path planning and obstacle avoidance. In that work, the environment is considered static. Del Pobil *et al.* [8] and Dillmann *et al.* [10] survey efforts in quantification for a set of robot tasks, including robot cleaning, robot rescue and autonomous driving. Another way of evaluating navigation systems is to let them compete in a challenge such as the DARPA urban challenge [7]. However, such challenges typically require to transport all robot systems to one location and their outcome is rather a ranking of systems than an analysis.

3 Experimental Protocol

In this section, we provide a detailed description of the proposed experimental protocol. Further details beyond the presentation here are available at `http://research.microsoft.com/brin/`. The goal of the protocol is to evaluate and compare the performance of navigation systems (hardware and software) in real environments over long periods of time. In order to allow comparison between different navigation systems at different physical locations, we devise means for normalizing the performance across environments and platforms and take measures towards standardization and repeatability of evaluations.

First, we define a standard environment composed of four *areas*. Second, we define a set of *challenges* that the robot has to face. These challenges include changes in environment appearance, geometrical configuration, and dynamic obstacles. Third, we introduce the concept of a *reference robot* and a *reference navigation system* that will be identical across evaluation sites. Expressing the performance of the tested system relative to this reference system ensures comparability of results across robots and evaluation sites. Finally, we employ a vision-based *ground-truth system* to evaluate the navigation performance of both the test and the reference robot.

We propose to simulate an entire year of robot operation, defining 12 *loops*, each corresponding to a virtual month of operation. The experimenter defines 8 *way-points*, two for each area, and creates a route that visits all way-points and always changes areas between way-points. The task of the robot is to travel along this route in each loop, facing a different set of challenges for each loop.

Tab. 1, 2 show an overview of the experimental protocol. The rows indicate the challenges, while the columns indicate their category, frequency, and configuration/location with respect to each of the twelve loops. In the remainder of this section, we will explain each element of our protocol in more detail.

3.1 Areas

We devised a standardized test environment consisting of four distinct areas: *atrium*, *lounge*, *office* and *hallway*. These areas are shown in the leftmost column

of Tab. 1, 2, grouping the challenges. The environment should contain at least one doorway and at least two different surfaces (e.g., carpet, tile, wood, cement). Ideally, the environment should not be a dedicated testing facility but rather a real building. Where possible, the test areas should be equipped with artificial lighting and with blinds or drapes to modify the environmental illumination.

The atrium is supposed to be a predominately open space with 90 percent or more of its surface area clear of furniture with a recommended size of above 15 m×15 m. The lounge is a social seating/dining area with an intended size of at least 12 m×12 m. The office is densely occupied by desks, office chairs and shelves and has a recommended minimum size of 10 m×10 m. The hallway has an intended length of at least 15 m and should have a low number of geometric and visual features. The above dimensions are recommendations, the experimenter is encouraged to respect the relative size of the areas in case of space limitations. Figures 2 and 4 show the real environments used in our experiments.

3.2 Challenges

We define a set of common environment dynamics, called challenges, to standardize the comparison with the reference robot and with tests conducted in different environments. Each challenge is listed as a numbered row in Tab. 1, 2. The challenges are representative of events and dynamics that are highly likely to occur at least once over a year-long deployment of a robot in a typical indoor environment. They are divided into three main categories that are shown in Tab. 1, 2 next to the challenge description:

Appearance (A): This category comprises visual appearance changes in the environment such as changing art work, whiteboard contents and lighting conditions. The challenges in this category are meant to test and assess the robustness of vision-based approaches.

Geometry (G): Challenges of this category include movable objects like doors, boxes, chairs, and ladders. These challenges simulate the natural variation of object configurations in environments and the different states of articulated objects such as doors. They test the robustness of navigation systems against geometry changes with respect to the setup and mapping phase. In addition to vision sensors, challenges in this category also affect proximity sensors.

Moving Obstacles (O): This category includes dynamic objects such as moving people, people transporting objects or gathering in groups, potentially (completely) blocking the path of the robot for an extended period of time. These challenges test the capabilities of a navigation system to deal with replanning while moving and to negotiate stalling situations.

All dynamic and moving elements have a designated frequency of occurrence and a designated location. The frequency can be hourly (H), daily (D), monthly (M) or yearly (Y) and is shown in the column next to the challenge category. The designated location/configuration of a challenge is shown in the respective column for each loop of the benchmark. If the navigation system of the robot does not rely on visual appearance (e.g., laser-based) one can skip the environment variations in the protocol that only affect visual appearance (category A).

3.3 Benchmark test grid

To ensure that the robot faces its challenges and the environment variations in a standardized and reproducible fashion, we devise a benchmark test grid that regulates the experimental evaluation. While the robot is traveling along its designated route, the environment is constantly modified according to the test grid shown in Tab. 1, 2. The test grid contains instructions that describe the challenges the robot has to face. For each challenge, the table lists the specific configuration for each of the 12 benchmark loops.

The experimenter has to devise positions for the way-points 1–8. Then, the experimenter defines the order in which the robot has to visit the way-points, taking care to avoid traveling between two way-points in the same area. One complete visit of all way-points counts as one loop, or a benchmarking month, for the evaluation. With the knowledge of the robot's default path the experimenter is then able to provide meaningful positions for the generic configurations of challenges like "Two People Blocking Path (no room to avoid)" (line 14), or "Person in Path" (line 8). It also falls into the responsibility of the experimenter to concretely define configurations for the qualitative settings of the environment dynamics, e.g., a configuration change from "Neat" to "Messy" in an experiment script, see also Sec. 4. Additionally, the experimenter records the lengths of the default path segments of a loop for the evaluation.

3.4 Reference robot and navigation system

For the baseline, we deploy the Pioneer P3-DX as *reference robot* in the same environment, running a *reference navigation software*. The software builds on the ARNL navigation stack shipped with the Pioneer, and is available at `http://research.microsoft.com/brin/`. We use ARNL 1.7.5.1 and BaseARNL 1.7.5.2 and change from the default values only the parameters *SecsToFail* to 90, *GoalOccupiedFailDistance* to 500 and *UseSonar* to "false".

The reference robot will visit the same way-points in the same order as the robot under evaluation. Thanks to the test grid introduced in the previous section it will also face the same challenges and configuration changes in a comparable manner. Fig. 1 (left) shows one of the reference robots used in the experiments.

3.5 Ground-truth evaluation

We developed a cheap and affordable ground-truth system [13] to automatically detect when and if the robot has successful reached a way-point. The system consists of visual markers placed on the ceiling and an upward-pointing camera mounted on the robot. A dedicated software component, independent of the navigation system, is responsible for capturing the images from the camera at the way-points and for determining the positioning accuracy. It is available free of charge at `http://research.microsoft.com/brin/`. The system requires an initial calibration in which the user manually drives to the way-points and registers

		Challenge	Cat.[a]	Freq.[b]	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6
All Areas	1	Artificial Lighting	A	D	Off	On	On	Off	Off	On
	2	Lamps On/Off	A	D	On	On	On	On	On	On
	3	Blinds or Drapes Open/Closed	A	D	All Closed	All Open	50/50	All Closed	All Open	50/50
	4	Wall Art Changes	A	Y	Wall Art 1					
	5	Door Open/Closed	G	H	Constantly	Constantly	Constantly	Constantly	Constantly	Constantly
	6	Wall Color Changes	A	Y	Color 1					
Atrium	7	Large Display Monitors Change Content	A	D	Image 1	Image 2	Image 3	Image 1	Image 2	Image 3
	8	Person in Path (ample room to avoid)	O	D		Position 1	Position2	Position 3	Position 4	
	9	Small Group in Path (ample room to avoid)	O	D		Position 1	Position 2	Position 3		Position 1
	10	Person Pushing Cart (ample room to avoid)	O	D		Position 1	Position 2		Position 1	Position 2
Hallway	11	Shipping Boxes on Floor	G	D		1 Box	2 Boxes	3 Boxes	2 Boxes	1 Box
	12	Cart Moves	G	D	Position 1	Position 2	Position 3	Position 4	Position 1	Position 2
	13	Ladders, Tools, Cables	G	Y			Position 1			
	14	Two People Blocking Path (no room to avoid)	O	D	Position 1				Position 2	
	15	Path Completely Blocked (door) for 1 Minute	O	D		Position 1				Position 2
	16	Path Completely Blocked (people) for 1 Minute	O	D			Position 1			
	17	Person Pushing Cart (no room to avoid)	O	D				Position 1		
Lounge	18	Dining Chairs Shift	G	H	Neat	25% Messy	50% Messy	75% Messy	100% Messy	Neat
	19	Coats/Jackets on Coat Racks	G	D		1/2 Full	Full		1/2 Full	Full
	20	Cart Moves	G	D	Position 1	Position 2	Position 3	Position 4	Position 1	Position 2
	21	Caution Sign (Janitor)	G	D		Position 1	Position 2		Position 2	Position 1
	22	Garbage/Recycling Bags	G	D		Black	White			2 Black
	23	Reconfigure Furniture	G	Y	Configuration 1					
	24	Person Vacuuming or Mopping	O	D		Position 1	Position 2		Position 1	Position 2
	25	Large Work/Social Gathering (20-30 people)	O	M			Position 1	Position 1		
Office	26	Whiteboard Contents Change	A	D	Clean	5%	10%	20%	30%	40%
	27	Desk Chairs Shift (less than 1.5 meters)	G	H	Neat	25% Messy	50% Messy	75% Messy	100% Messy	Neat
	28	Coats/Jackets on Chairs	G	D			5%		10%	2%
	29	Bags on Floor Near Desks	G	D		20%	40%	60%		20%
	30	Loose Paper on Floor	G	D	0 Pieces	5 Pieces	0 Pieces	5 Pieces	0 Pieces	5 Pieces
	31	Shelves Contents Change	G	M	20% Full		40% Full		60% Full	
	32	Shelves Move	G	Y	Position 1					
	33	Small Gathering in Work Area (4-8 people)	O	D		Position 1	Position 2		Position 1	Position 2
	34	Social Gathering (10-15 people)	O	M			Position 1			

[a] Challenge category. A: Appearance, G: Geometry, O: (moving) Obstacle

[b] Frequency of occurence. H: Hourly, D: Daily, M: Monthly, Y: Yearly

Table 1. The benchmark test grid proposed in this work. The table lists the configuration of each challenge for every loop of the benchmark, see Tab. 2 for the second part covering months/loops 7–12.

		Challenge	Cat.[a]	Freq.[b]	Month 7	Month 8	Month 9	Month 10	Month 11	Month 12
All Areas	1	Artificial Lighting	A	D	On	Off	Off	On	On	Off
	2	Lamps On/Off	A	D	Off	Off	Off	Off	Off	Off
	3	Blinds or Drapes Open/Closed	A	D	All Closed	All Open	50/50	All Closed	All Open	50/50
	4	Wall Art Changes	A	Y	Wall Art 2					
	5	Door Open/Closed	G	H	Constantly	Constantly	Constantly	Constantly	Constantly	Constantly
	6	Wall Color Changes	A	Y	Color 2					
Atrium	7	Large Display Monitors Change Content	A	D	Image 1	Image 2	Image 3	Image 1	Image 2	Image 3
	8	Person in Path (ample room to avoid)	O	D	Position 1	Position 2	Position 3	Position 4		Position 1
	9	Small Group in Path (ample room to avoid)	O	D	Position 2	Position 3		Position 1	Position 2	Position 3
	10	Person Pushing Cart (ample room to avoid)	O	D		Position 1	Position 2		Position 1	Position 2
Hallway	11	Shipping Boxes on Floor	G	D		1 Box	2 Boxes	3 Boxes	2 Boxes	1 Box
	12	Cart Moves	G	D	Position 3	Position 4	Position 1	Position 2	Position 3	Position 4
	13	Ladders, Tools, Cables	G	Y		Position 2				
	14	Two People Blocking Path (no room to avoid)	O	D			Position 3			
	15	Path Completely Blocked (door) for 1 Minute	O	D				Position 3		
	16	Path Completely Blocked (people) for 1 Minute	O	D	Position 2				Position 3	
	17	Person Pushing Cart (no room to avoid)	O	D		Position 1				Position 3
Lounge	18	Dining Chairs Shift	G	H	25% Messy	50% Messy	75% Messy	100% Messy	Neat	25% Messy
	19	Coats/Jackets on Coat Racks	G	D		1/2 Full	Full		1/2 Full	Full
	20	Cart Moves	G	D	Position 3	Position 4	Position 1	Position 2	Position 3	Position 4
	21	Caution Sign (Janitor)	G	D		Position 1	Position 2		Position 2	Position 1
	22	Garbage/Recycling Bags	G	D	2 White			2 Black	2 White	
	23	Reconfigure Furniture	G	Y	Configuration 2					
	24	Person Vacuuming or Mopping	O	D		Position 1	Position 2		Position 1	Position 2
	25	Large Work/Social Gathering (20-30 people)	O	M			Position 2	Position 2		
Office	26	Whiteboard Contents Change	A	D	50%	60%	70%	80%	90%	100%
	27	Desk Chairs Shift (less than 1.5 meters)	G	H	25% Messy	50% Messy	75% Messy	100% Messy	Neat	25% Messy
	28	Coats/Jackets on Chairs	G	D	20%	4%	30%	6%	40%	8%
	29	Bags on Floor Near Desks	G	D	40%	60%		20%	40%	60%
	30	Loose Paper on Floor	G	D	0 Pieces	5 Pieces	0 Pieces	5 Pieces	0 Pieces	5 Pieces
	31	Shelves Contents Change	G	M	20% Full		40% Full		60% Full	
	32	Shelves Move	G	Y	Position 2					
	33	Small Gathering in Work Area (4-8 people)	O	D		Position 1	Position 2		Position 1	Position 2
	34	Social Gathering (10-15 people)	O	M				Position 2		

[a] Challenge category. A: Appearance, G: Geometry, O: (moving) Obstacle

[b] Frequency of occurence. H: Hourly, D: Daily, M: Monthly, Y: Yearly

Table 2. Continuation of Tab. 1, the benchmark test grid proposed in this work.

P3-DX (reference robot)

omniRob

MSR-P1

Fig. 1. Robots used in our experiments. All robots have an up-facing camera mounted for ground truth marker detection. *Left:* The reference robot, a Pioneer P3-DX with a SICK LMS 200 laser scanner. *Middle:* The omniRob used in the environment ALU-FR. *Right:* The Microsoft Robotics Prototype 1 (MSR-P1), used in the environment MS.

their position within the reference software. The visual markers are black-and-white checkerboards printed on foam-boards, and thus cheap and disposable, see Fig. 5. Whenever the robot reports an arrival at a way-point, the ground-truth system determines whether the way-point is reached, the accuracy with respect to the marker and the time elapsed from the last way-point.

We compute the following statistics: total number of failures, time to failure, distance to failure, average speed, accuracy at goal. The total number of failures is the number of segments in which the navigation system has been unable to arrive at a way-point. The time to failure is the operational time between consecutive failures, counted from the last restart to the last successfully visited way-point.

4 Experiments

We prepared two environments for the experiments. The first setup (environment ALU-FR) has been prepared in a large experimental area at the University of Freiburg, Germany. The second (environment MS) is a large real office environment in the Microsoft Research building in Redmond, Washington, USA.

In the environment ALU-FR, we have benchmarked the navigation method proposed in [18, 19] installed on the omnidirectional robot omniRob shown in Fig. 1 (middle). In the environment MS, we evaluated an in-house experimental Microsoft navigation software, on the Microsoft Research Prototype 1 (MSR-P1) shown in Fig. 1 (right). The robot performs both SLAM and navigation by using only the Microsoft Kinect depth stream, gyroscope, and wheel odometry. In both environments, we have run the reference software on the reference platform Pioneer P3-DX, see Fig. 1 (left) and Sec. 3.4.

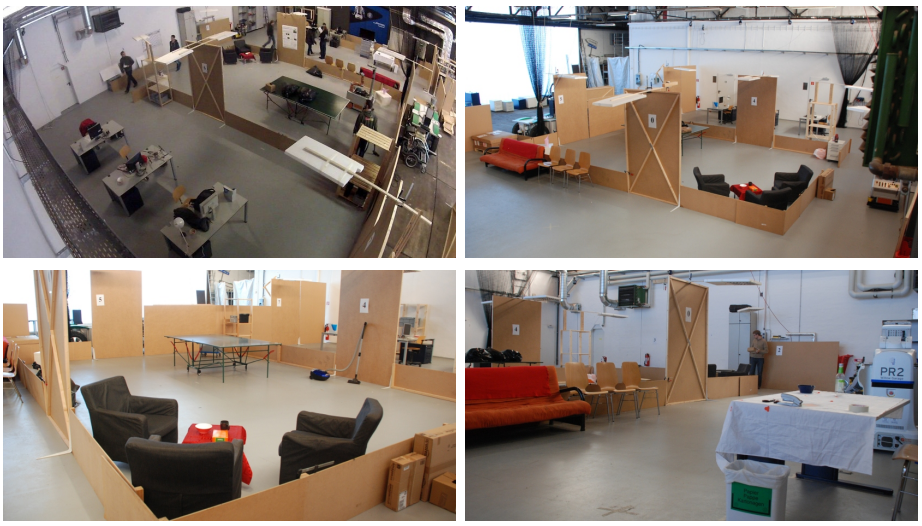

Fig. 2. Overall views of the ALU-FR environment: office (top-left), atrium (top-right) and detail views of the lounge (bottom-left) and the atrium (bottom-right).

4.1 Environment ALU-FR

We furnished the environment to make each dedicated area verisimilar. This includes tables, cupboards, chairs, couches and computers. In particular, we have used wooden panels to subdivide the environment and fixed the fiducial markers at the way-points at a height of approximately 2.45 m. The complete environment measures 19 m×12 m, the atrium 7.5 m×11 m, the lounge 6 m×9 m, the office 5.5 m×12 m, and the hallway is 7 m long, see Fig. 2.

We instantiated the test grid from Tab. 1, 2 into a concrete test script for our experiments. This is important to ensure that the test robot and the reference robot face the same challenges at the same time of each run. The laser-based occupancy grid map used for localization and navigation of omniRob shown in Fig. 3 displays the eight way-points and some of the devised challenge positions. We specified a route by ordering the way-points as follows: 0→2→4→6→3→5→1→7→0. This order has succeeding way-points in different areas and the travel distance between way-points is varying from short to long. We devised positions for people to gather at and move to. Marking these positions on the floor is helpful for the participants during the experiments and to ensure repeatability.

Creating an experiment script from the test grid in Tab. 1, 2 requires particular care on how to design the challenges and which of them can be omitted. The environment and the challenges have to be designed in a way that a path exists for the robot. As the omniRob is larger than the reference robot, we had to increase the size of doors and hallways. The navigation systems of the omniRob and the reference robot are not based on vision sensors but make only use

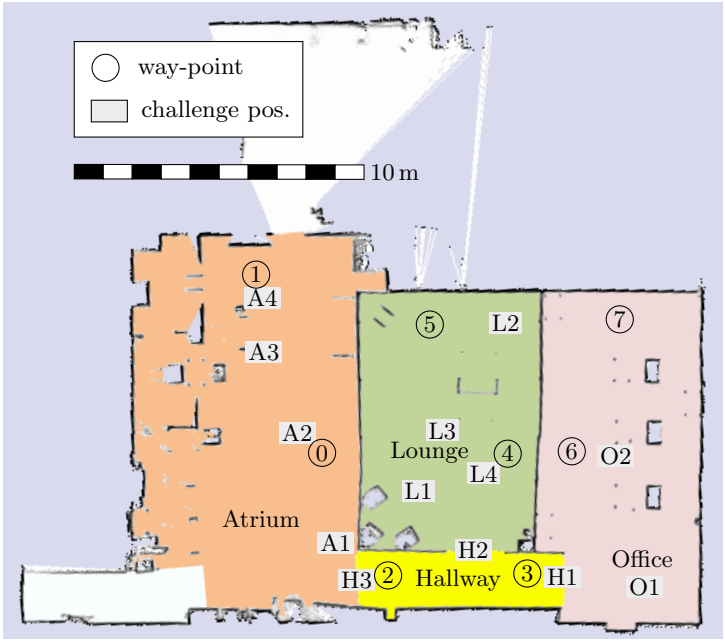

Fig. 3. The occupancy grid map used for the omniRob experiments in the ALU-FR environment. The four areas are marked by color and the map also shows the locations of way-points and some of the test grid challenges.

of laser range finders. Therefore, we omitted challenges which have no or only minor effects on laser range finders such as changing artificial lighting, opening/closing blinds, wall art changes, wall color changes, and whiteboard content changes, i.e., lines 1–4,6,7,26 from Tab. 1, 2.

Furthermore, we did not put ladders, tools, cables and the cart in the hallway because of the omniRob footprint and the particular manufacturing of its wheels (lines 12,13). Due to the omniwheels of the omniRob, we skipped also the loose paper challenge (line 30). Moreover, we skipped the constant opening and closing of doors (line 5), the lounge coat racks (line 19), the janitor sign (line 21), modified the garbage bags to only be black (line 22) and limited the size of the biggest social gathering to 8 people (lines 25,34).

The test grid only defines the challenges per loop but not at what time in the loop they occur. It is up to the experimenter to define when the robot faces the challenges in each loop. An excerpt of our experiment script is shown in Tab. 3. It shows all the travel segments for month/loop 3 of our test script that we derived from Tab. 1 and specifies which challenge configurations are applied for each loop segment. It is a detailed instruction procedure for the experimenter on how to modify the environment during the evaluation to ensure repeatability and reproducibility of the experiments: For example, while the robot travels between 2→4, it encounters two parcel boxes in the hallway and two people block the

Loop Segment	Area	Challenge/Configuration	Row in Tab. 1, 2
0→2	Atrium	person with cart at A2	10
2→4	Hallway	2 boxes on the floor	11
	Hallway	2 people block at H2 for 1 min	16
	Lounge	move chairs by 0.2 m	18
	Lounge	cart at L3	20
	Lounge	1 garbage bag on the floor	22
4→6	Office	move chairs by 0.2 m	27
	Office	1 jacket on chair	28
	Office	2 bags next to desks	29
	Office	group of 4 people at O2	33
	Office	shelves 40% filled	31
6→3			
3→5	Lounge	group of 8 people at L1	25
	Lounge	person vacuuming at L2	24
5→1	Atrium	group of 4 people at A2	9
1→7	Atrium	person at A2	8
	Office	group of 8 people at O1	34
7→0			

Table 3. Excerpt of the instantiation of the test grid (see Tab. 1 and 2) to an evaluation script for loop 3 of environment ALU-FR. The specific challenges and their locations are shown for each segment of the loop, see also Fig. 3 for challenge locations.

door H2 for 1 min. All chairs in the lounge are moved by 0.2 m with respect to their position while mapping the environment. The cart of the lounge is placed at L3 and one garbage bag was placed on the ground, see also Fig. 3.

4.2 Environment MS

The second environment consists of several areas of the Microsoft Research building 99 in Redmond, Washington, see Fig. 4. The atrium measures 25 m×20 m, the lounge 20 m×12 m, the office 10.5 m×7.8 m and the hallway 17 m×1.75 m. This environment includes an open floor plan in the atrium and lounge areas. It has substantial daylight coming in through the glass ceilings and the entrance. The lounge area includes a coffee shop, with multiple round tables and chairs, as well as tall rectangular tables with high chairs, couches and armchairs. The areas have carpet, linoleum, rough tile and hardwood as floor surface. Where practical, we chose the landmark locations close to interesting or meaningful locations when creating the test script for this environment, such as adjacent to the coffee stand, in front of the elevators and near the receptionist desk. The environment included a doorway between the hallway and the office as well as one additional doorway into an unmapped adjacent space that was alternately opened or closed for each loop. As we used a Microsoft Kinect depth sensor for mapping and navigation, we omitted the challenges involving lighting or appearance changes from the script, including lines 1–4,6,7, and 26 from Tab. 1 and 2.

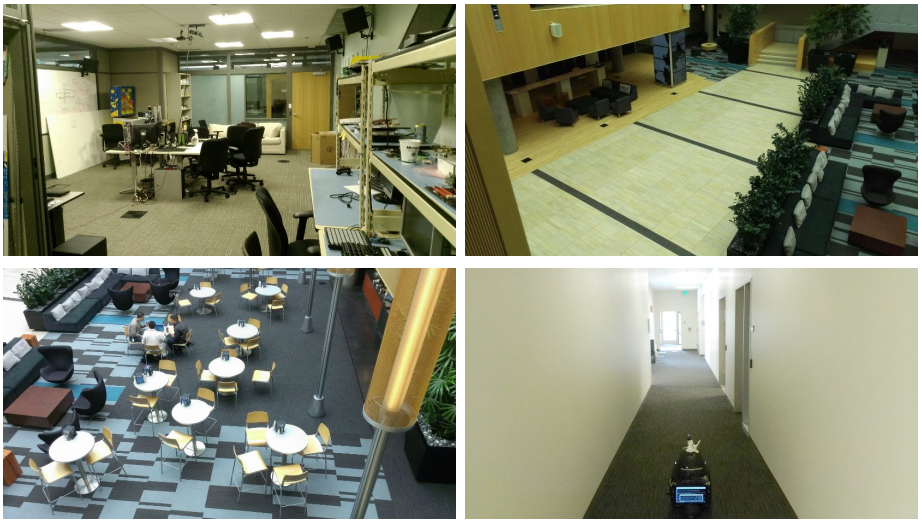

Fig. 4. The four areas of the environment MS: office (top-left), atrium (top-right), lounge (bottom-left) and hallway (bottom-right).

No shelf was available for the office, so we omitted challenges 31 and 32. Challenges 25 and 34 were omitted due to a lack of the required number of people. To avoid disturbances by direct sunlight or non-scripted interactions with people, we started the experiments in the evening.

4.3 Results

The performance of the different systems in the two environments is listed in Tab. 4 and Tab. 5. The last column of each table shows the *relative* performance of a navigation system with respect to the reference one. Thanks to the benchmark protocol, it is now possible to say how accurate is a system with respect to a standardized baseline and environmental conditions. In environment ALU-FR, neither omniRob nor the reference system failed during the ∼1.5 km navigation length in circa 70 min. In environment ALU-FR the robots can always observe sufficient structure to properly localize.

In environment MS, the MSR-P1 and the reference system both encountered failures. The failures for the MSR-P1/reference robot were software problems (1/1), localization inaccuracies (3/1) and divergence (1/1), faulty obstacle perception (0/3), path oscillation for more than 5 minutes (0/1), not finding a path around a new obstacle (0/1) and not detecting a low obstacle (0/1). The benchmark revealed defects in several key areas of navigation including planning, localization, static and dynamic obstacle avoidance, reactive re-planning, remapping, and endurance, consistent with the limitations of each software. The experiments covered ∼2.1 km and took 6 hours to conduct for each robot.

Performance	Freiburg	Reference	Ratio
Number of failures	0	0	-
Mean time to failure	-	-	-
Maximum time to failure	4343 s	5125 s	0.85
Mean distance to failure	-	-	-
Maximum distance to failure	1423 m	1349 m	1.05
Average speed	0.33 m/s	0.26 m/s	1.27
Positioning error	0.005 m ± 0.007 m	0.05 m ± 0.04 m	0.10

Table 4. Benchmark results in the environment ALU-FR.

Performance	Microsoft	Reference	Ratio
Number of failures	5	9	0.56
Mean time to failure	2265 s	726 s	3.12
Maximum time to failure	5023 s	1971 s	2.55
Mean distance to failure	367 m	183 m	2.01
Maximum distance to failure	860 m	472 m	1.82
Average speed	0.16 m/s	0.25 m/s	0.64
Positioning error	0.23 m ± 0.2 m	0.22 m ± 0.1 m	1.05

Table 5. Benchmark results in the environment MS.

Three months prior to the experiments in environment MS, we conducted a stripped down version of the benchmark with older MSR-P1 software. We found that the MSR-P1 showed dramatic improvements (5 failures vs. 12) with respect to the pre-test, consistent with the improvements in navigation and mapping software done in the meantime. We also found that the reference system performed worse in the full benchmark (9 failures vs. 5). This before and after experiment confirms the benchmark's ability to expose the effects of both software and environmental changes.

We believe the results accurately reflect the capabilities and performance of all tested systems. In our observation this is primarily due to the wide coverage of possible failure modes. Moreover, the amount of challenges in our protocol seemed appropriate. The relatively small cumulative runtime seems sufficient to capture a good performance representation. However, as navigation systems get better, the total runtime might need to be increased.

5 Lessons Learned

Comparing autonomous navigation solutions according to their performance in real environments is an arduous task. During the process of setting up and performing the evaluation, we came across two aspects to be considered.

A first aspect is related to the comparison of different systems at different locations. The reference robot is instrumental in providing a sense of the complexity of each environment. However, one must consider that the shape and the size of the robot has a certain degree of influence on the results. The chosen benchmark targets navigation in office environments, thus slightly favoring

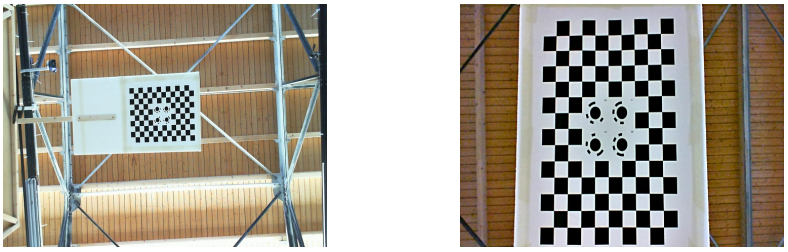

Fig. 5. Influence of camera mounting on marker detection tolerance. The pictures show the marker as seen from the camera for way-point 1 in environment ALU-FR. *Left:* Reference robot P3-DX, the camera is mounted at a height of 0.45 m, see also Fig. 1 (left). *Right:* Freiburg's omniRob, the camera is mounted 90 degrees rotated with respect to the camera of the reference robot and at a height of 1.7 m, see also Fig. 1 (middle).

small and circular robots. When the system under test differs from the reference robot in size, shape or even locomotion principles, the environment and the protocol should be slightly adapted to allow a fair comparison. This happened, for instance, when we evaluated the omniRob system, as described in Sec. 4.1.

A second aspect lies in the fiducial system. The location of the camera on the robot is very important as the relative distance between the markers and the camera defines the *success* range for the failure detection system. A longer relative distance between them allows the marker to be detected from further away, see Fig. 5.

6 Conclusion

With this paper, for the first time, we have presented an experimental protocol to evaluate a robotic indoor navigation system as a whole. Differently from other scientific disciplines, robot navigation cannot be evaluated only with datasets. To ensure repeatability and reproducibility of experiments, our benchmark protocol provides detailed definitions for the environment dynamics. Additionally, we proposed the concept of a reference robot to allow comparison between different navigation systems at different experimentation sites. We applied our protocol and conducted experiments with different robots in two different research groups, showing the validity of the benchmark.

7 Acknowledgment

This work has partly been supported by the EC under FP7-260026-TAPAS, FP7-610917-STAMINA, and FP7-267686-LIFENAV. The authors thank all members of the AIS Lab, the Microsoft Robotics Team, Studio99 and the Building 99 Hardware Lab for their patient help with the experiments.

References

1. K. Bache and M. Lichman. UCI machine learning repository. University of California, Irvine. `http://archive.ics.uci.edu/ml`, 2013.
2. J. Bennett and S. Lanning. The netflix prize. In *KDD cup and workshop at the 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Minig*, 2007.
3. J. Borenstein and L. Feng. Umbmark: A benchmark test for measuring odometry errors in mobile robots. *Proc. SPIE*, 2591:113–124, 1995.
4. W. Burgard, C. Stachniss, G. Grisetti, B. Steder, R. Kümmerle, C. Dornhege, M. Ruhnke, A. Kleiner, and J. D. Tardós. A comparison of SLAM algorithms based on a graph of relations. In *Int. Conf. on Intelligent Robots and Systems*, 2009.
5. D. Calisi, L. Iocchi, and D. Nardi. A unified benchmark framework for autonomous mobile robots and vehicles motion algorithms (MoVeMA benchmarks). In *RSS-Wksp. on experimental methodology and benchmarking in robotics research*, 2008.
6. CAVIAR data sets. `http://homepages.inf.ed.ac.uk/rbf/CAVIAR`.
7. DARPA urban challenge rules. `http://archive.darpa.mil/grandchallenge/docs/Urban_Challenge_Rules_102707.pdf`, 2007.
8. A. P. Del Pobil, R. Madhavan, and E. Messina. Benchmarks in robotics research. In *IROS Workshop on Benchmarks in Robotics Research*, 2007.
9. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2009.
10. R. Dillmann. Ka 1.10 benchmarks for robotics research. `http://www.cas.kth.se/euron/euron-deliverables/ka1-10-benchmarking.pdf`, 2004.
11. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *International journal of computer vision*, 88(2), 2010.
12. J.-S. Gutmann, W. Burgard, D. Fox, and K. Konolige. An experimental comparison of localization methods. In *Int. Conf. on Robotics & Automation*, 1998.
13. H. Kikkeri, G. Parent, M. Jalobeanu, and S. Birchfield. An inexpensive methodology for evaluating the performance of a mobile robot navigation system. In *Int. Conf. on Robotics & Automation*, 2014.
14. R. Knotts, I. Nourbakhsh, and R. Morris. Navigates: A benchmark for indoor navigation. In *Int. Conf. and Exp. on Robotics for Challenging Environments.*, 1998.
15. W. Nowak, A. Zakharov, S. Blumenthal, and E. Prassler. Benchmarks for mobile manipulation and robust obstacle avoidance and navigation. *BRICS Deliverable D3.1*, 2010.
16. B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.
17. PETS 2009 data set. `http://pets2009.net/`.
18. J. Röwekämper, C. Sprunk, G. Tipaldi, C. Stachniss, P. Pfaff, and W. Burgard. On the position accuracy of mobile robot localization based on particle filters combined with scan matching. In *Int. Conf. on Intelligent Robots and Systems*, 2012.
19. C. Sprunk, B. Lau, P. Pfaff, and W. Burgard. Online generation of kinodynamic trajectories for non-circular omnidirectional robots. In *Int. Conf. on Robotics & Automation*, 2011.
20. J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *Int. Conf. on Intelligent Robots and Systems*, 2012.