Evaluation of RoboCup Maps

Benjamin Balaguer, Stefano Carpin
School of Engineering
5200 North Lake Road
University of California, Merced, USA
+1(209)228-4152
{bbalaguer,scarpin}@ucmerced.edu

Stephen Balakirsky
NIST
100 Bureau Drive
Gaithersburg, MD, USA
+1(301) 975-4791
stephen@nist.gov

Arnoud Visser
Universiteit van Amsterdam
Science Park 107
1098 XG Amsterdam, NL
+31 (20) 525-7532
A.Visser@uva.nl

ABSTRACT

This paper describes the steps taken to create the score criteria aimed at measuring the quality of maps produced by teams participating in the RoboCup Rescue Virtual Robots competition. Since metrics have already been developed by a few research groups, we start by highlighting the most popular solutions to this problem, emphasizing their strengths and weaknesses. Having put the difficulty of creating map benchmarks into perspective, we present our map benchmark suite, appropriate for Urban Search and Rescue missions, along with examples taken from former competitions.

Categories and Subject Descriptors

F.2.3 [**Theory of computation**]: Analysis of Algorithms and Problem Features—*Tradeoffs among complexity measures*

General Terms

Performance, Measurement

Keywords

Map evaluation, RoboCup, Performance Metrics

1. INTRODUCTION

One of the liveliest competitions in RoboCup is the Virtual Robot Rescue league, where participants are called upon to deploy teams of robots capable of locating victims and hazards in highly unstructured areas. Differently from other RoboCup competitions, the Virtual Robot Rescue league asks robot teams to map completely unknown environments, with little or no apriori information. The theme behind the league is Urban Search and Rescue (USAR), where robots have to not only work cooperatively as unified teams but also have to consider humans, whether they be victims or first responders. As such, the maps that are generated by the robots need to incorporate useful information that first responders can exploit, adding a new dimension to robot mapping. Indeed, robots now have to generate multiple maps, some of which are for their own needs (e.g. navigation) while others are explicitly for first-responders (e.g. victim locations with safest paths to reach them).

The Virtual Robot Rescue League uses USARSim [5] to simulate disaster scenarios. The simulation is extremely realistic thanks to a dynamic community of users and developers who strive to validate each robot, sensor, or other physical properties. This community-based involvement translates into a remarkably accurate simulation capable of modeling multifaceted disaster environments ranging from traffic accidents to earthquakes and explosions, each possibly exploiting the effects of smoke, fires, debris, water, to name a few. In addition to the near-zero participation cost and the ability to create realistic city-sized disasters, the simulation offers ground truth data that would otherwise be awfully difficult to gather. The large amount of robotic platforms and sensors that USARSim proposes results in a challenging situation for map scoring. Indeed, each team solves the mapping problem differently using a diverse set of robots and sensor configuration, resulting in a massive mismatch between maps, from scaling to rotational differences. This observation put us in the unique position of having to come up with a map benchmark robust enough to take into account all of these differences along with the opportunity of having a tremendous amount of data to test our solution to this difficult problem.

Evidently, and despite the fact that it is still frequently employed, a qualitative approach is fundamentally insufficient for a competition where results have to be both repeatable and reliable. Not wanting to develop a map benchmark from scratch, and optimistically hoping that a solution had already been published, we performed an extensive case study, a subsection of which is shown in Section 2. Realizing that no current solution was robust enough for the problem at hand (i.e. teams would be able to take advantage of the metrics' weaknesses), we developed a mapping benchmark suite comprised of standards and categorized metrics, which are described in Section 3 and 4, respectively. It is worthwhile to note that the standards were so well received that they have subsequently been implemented as part of the Real Robot Rescue League. We close the paper with concluding remarks and possible future work in Section 5. While this paper focuses on mapping, a companion paper highlights the overall RoboCup 2009 competition [2].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 2009 ACM X-XXXXX-XX-X/XX/XX ...$10.00.

2. CASE STUDY

Map benchmarking is a relatively novel effort, and so is robot benchmarking in general. Therefore, the amount of formerly published scholar work is rather limited (the reader is referred to a forthcoming special issue of the Autonomous Robots journal on *Characterizing mobile robot localization and mapping*). In this section, we quantitatively compare some of the most popular benchmark metrics that have been previously published. We run the metrics with two binary occupancy grid maps, one generated by a robot and the other being ground truth. Each grid cell can only have a value of 1 for occupied space or a value of 0 for free space. Please note that the discussion in this section is entirely based on our binary map representation and that results might be different with a probabilistic occupancy grid map. Even though we have performed a full case study on different environments, we only present a representative example in Fig. 1 and Table 1 due to space constraints.

The first set of four metrics, namely the Map Score [8], Overall Error [4], Normalized Map Score [10], and Occupied Map Score [10], represents an approach requiring pixel-to-pixel comparisons between the ground truth and robot-generated maps. The Map Score metric counts the number of ground truth and robot map pixels that are the same. As such, the total number of pixels in the maps would be a perfect score. The Overall Error metric counts the number of ground truth and robot map pixels that are different, where a perfect score would be zero. It is worthwhile to mention that the Map Score metric measures accuracy whereas the Overall Error metric measures error and that adding both metrics together will equal the total number of pixels. The two aforementioned metrics are utilized over all the pixels, regardless of what they represent (i.e. occupied or free space). Consequently, the two metrics are biased towards maps with large regions of correct free space, as shown in Fig. 1 and Table 1. From the table, Team A and Team E have the best scores and, looking at the figure, the bias is clear: the two maps with the smallest amount of discovered walls receive a higher score. Research groups have attempted to remove this bias by introducing the Normalized Map Score and Occupied Map Score metrics. They work the same way as the Overall Error metric (i.e. looking for pixel mismatches) but are only run on the occupied space of the maps. The Normalized Map Score runs on the occupied space of the ground truth map whereas the Occupied Map Score runs on the occupied space of the robot-generated map. Unfortunately, these metrics only move the bias, which is now dependent on the occupied space. Using the Normalized Score metric, the robot maps that have thick walls do better, as shown by Team C and Team E, since they do a better job in replicating the wall thickness of the ground truth map. In contrast, Team A and Team B do better with the Occupied Map Score metric, thanks to their thin walls that allow for a greater margin of error when compared to the thicker ground truth walls.

Another interesting pixel-to-pixel approach is presented through the Picture-Distance function [3]. In this metric, the score represents the Manhattan-distance between an occupied pixel in the ground truth map and the closest occupied pixel in the robot-generated map. The process is repeated over all the occupied pixels of 1) the ground truth map and 2) the robot-generated maps. Finally, the result is normalized by dividing it by the total number of pixels considered. The Picture-Distance function is a measure of map error and, as such, the best possible score is zero. A look at Fig. 1 and Table 1 quickly shows that the two teams who have explored the most, Team C and Team D, do better with this metric. From both the method used and the experiment performed, it is clear that the method is also biased, towards exploration (i.e. wall discovery).

Moving away from the bias of pixel-to-pixel comparisons brings us to correlation coefficients, a comparison measures valued between -1 and 1, with -1, 0, and 1 representing perfect inverse correlation, no correlation, and perfect correlation, respectively. The Baron's Cross Correlation coefficient [10] attempts to correlate two images by using the ground truth and robot-generated pixels' mean and standard deviation. Since averages are used, and the pixel's values can only be 0 or 1, the Baron's coefficient rewards robot maps that have a similar number of occupied and free pixels to the ground truth. Consequently, the coefficient is influenced both by wall thickness and exploration, as can be seen in Fig. 1 and Table 1 where Team C and Team E have the highest scores. The Pearson's Correlation coefficient [7] evaluates the occupied space of the map as a spatial function, trying to linearly describe one map from the other. The Pearson's coefficient requires an approximately similar point distribution between the two map. This drawback is evidenced by the results for Team A and Team E, where, even though both maps are very similar they have extremely different Pearson's coefficients. It is worthwhile to note that both correlation coefficients can be unpredictable, shown by the scores of Team A and Team B.

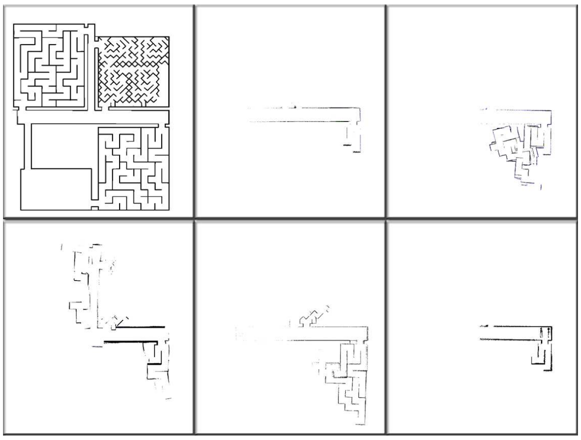


Figure 1: Example set of maps used for the Case Study, the results of which are in Table 1. The first image is the ground truth with the remaining images being, from left to right and up to down, Team A, Team B, Team C, Team D, and Team E, respectively.

3. MAP REPRESENTATION STANDARDS

One of the principal obstacles impeding the development of a consistent map benchmark comes from the lack of standards between the incredible amount of mapping algorithms that have been developed, through the years, by various research groups. Indeed, each algorithm works differently, from the way they represent maps (e.g. occupancy grids, topological, feature-based, etc...) to the different scales and

Metric	Team A	Team B	Team C	Team D	Team E
Map Score [8]	**586779**	586192	585049	585297	**586815**
Overall Error [4]	**56577**	57164	58307	58059	**56541**
Normalized Map Score [10]	56065	55785	**55227**	55363	**54367**
Occupied Map Score [10]	**512**	**1379**	3080	2696	2174
Baron's Correlation [10]	-0.005	0.017	**0.036**	0.032	**0.098**
Pearson's Correlation [7]	0.298	-0.060	**0.479**	0.295	**0.591**
Picture-Distance [3]	210.09	254.37	**129.89**	**189.44**	221.61

Table 1: Metrics comparison for the maps shown in Fig. 1. The seven rows represent each metric taken from different publications. The bold font shows the two best results for a specific metric.

rotations that they may encompass. Having to rank maps generated by many different robotics groups and, as a consequence, facing the same map representation problems, we have imposed two mapping standards on participants, the GeoTIFF image format and the MIF vector format. We have found, over the years, that participants embrace them, primarily for their ease-of-use, while giving the administrators powerful tools to generate a fair mapping benchmark.

3.1 GeoTIFF Image Format

The GeoTIFF image format embeds geographical information as an integral part of the map. The power of GeoTIFF lies in its ease of use, open standard, and layer friendliness. Indeed, it is very simple to geo-reference any map, by providing an additional file comprised of six parameters, namely the X and Y positions of the upper-left pixel, the scale of a pixel in the X and Y directions, the rotation, and the skew. These six parameters take into account any potential differences in scale, translation, and rotation between maps. GeoTIFF is an open standard, a fact that translates into a plethora of open tools that work across different platforms and programming languages. Last but not least, it is very easy to embed multiple layers on top of the original map, a powerful way to display varied information on the maps. Evidently, from a map benchmark standpoint, the GeoTIFF image format allows every map, including ground truth, to be overlaid on top of each other, as shown in Fig. 2; making it straightforward to evaluate the maps either quantitatively or qualitatively.

3.2 MIF Vector Format

The MIF vector format is similar to the GeoTIFF format in that it possesses the same qualities of allowing georeferencing, remaining easy to use, being an open standard, and working well with layers. The difference between the two, however, lies in what can be represented. Whereas GeoTIFF represents images, MIF works with geometric primitives (e.g. points, lines, polygons) that can have an arbitrary number of attributes. The MIF vector format can be best exploited to display topological or feature-based maps, where labeled nodes or features can give high-level information or particular landmarks of interest to first responders. Fig. 3 shows some examples of what can be achieved with a MIF vector file.

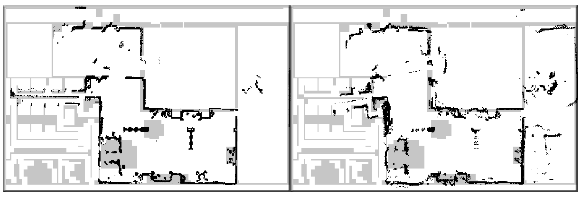
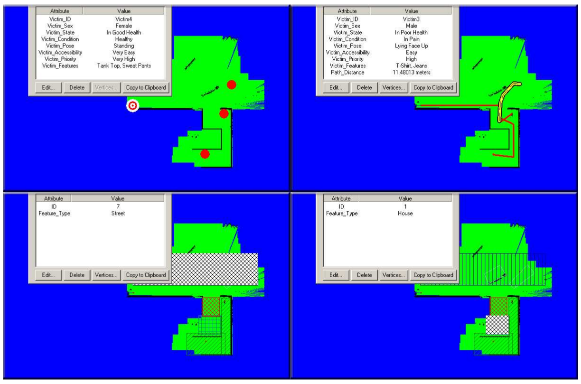


Figure 2: Examples of two robot-generated maps (black) overlaid on top of the ground truth map (gray) for an indoor environment.



Attribute	Value
Victim_ID	Victim4
Victim_Sex	Female
Victim_State	In Good Health
Victim_Condition	Healthy
Victim_Pose	Standing
Victim_Accessibility	Very Easy
Victim_Priority	Very High
Victim_Features	Tank Top, Sweat Pants

Attribute	Value
Victim_ID	Victim3
Victim_Sex	Male
Victim_State	In Poor Health
Victim_Condition	In Pain
Victim_Pose	Lying Face Up
Victim_Accessibility	Easy
Victim_Priority	High
Victim_Features	T-Shirt, Jeans
Path_Distance	11.48013 meters

Attribute	Value
ID	7
Feature_Type	Street

Attribute	Value
ID	1
Feature_Type	House

Figure 3: Four examples of MIF vector files, overlaid on top of the robot-generated map. The upper-left picture shows points representing victims' location labeled with various information about each victim. The upper-right picture shows line segments highlighting the best path to reach each victim, labeled with the victim's information and path's length. The lower pictures display regions of interests, including a street (left) and a house (right).

4. MAP BENCHMARK

It is clear from the Case Study that no published algorithm is adequate on its own or as part of a map benchmarking suite. They each have some sort of bias and cannot solve the problem of error propagation, the toughest challenge when evaluating maps, where similar mapping errors can affect maps differently depending on when the error occurred. For example, an orientation error at the beginning of a mission will result in a map that is wrong through the rest of the mission, whereas the same orientation error at the end of the mission will affect a much smaller portion of the map. It is our belief that the maps should be equally deserving, provided that everything else is equal. Additionally, a map is application-specific and, in our case, USAR and first responders have to be in-the-loop. As such, we devised a categorized benchmark comprised of Metric Quality, Skeleton Quality, Attribution, Grouping, Utility, and Creativity. Each category possesses a weight, the combination of which can be used to steer the competition towards one or more research agendas.

4.1 Metric Quality

The Metric Quality tries to solve the same problem that was studied in the Case Study: the comparison of the robot-generated occupancy grid map to ground truth, from an accuracy standpoint. In order to bypass the aforementioned problem of error propagation, we further divide the Metric Quality into Global and Local Quality. The Global Quality is a measure of the number and severity of mapping errors whereas the Local Quality is a measure of accuracy between these mapping errors. Using Fig. 4 as an example, one can see that both robot-generated maps are similar in terms of Global Quality, each having a small error with the lower hallway. The right map, however, is worst in terms of Local Quality, since it is missing some walls in the center of the map.

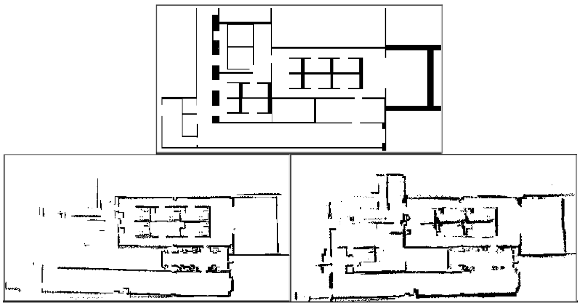


Figure 4: Example for the Metric Quality evaluation, where the upper map is ground truth and the lower-left and lower-right maps are different robot-generated maps.

4.2 Skeleton Quality

The Skeleton Quality evaluates a topological map rather than an occupancy grid map, which can be more useful to first responders. A first responder should be able to follow a skeleton map to reach a chosen point. In this case, the quality is determined from the number of false positives and false negatives. A false positive occurs when a node cannot be accessed whereas a false negative takes place when a clear topological location is available but has not been included in the skeleton map. Fig. 5 shows examples of skeleton maps with similar qualities. The first map has a lot of false positives in the lower and right sections of the map, where topological locations have been identified in unexplored space. The second map contains both false positives, where a topological node is inside a wall, and false negatives, along the left side of the hallway.

4.3 Attribution

The Attribution section of our mapping benchmark aims to reward teams that can successfully deliver a feature-based map with valuable information for first responders. The type of information that can be embedded into the map is fairly open, even though most teams deliver feature-based maps indicating victim locations and information, best paths to reach victims, robot paths, and important landmarks. The Attribution is scored based on the amount and accuracy of the data. As an example, Fig. 6 shows two maps, each providing victim locations and best paths to reach them. Both maps provide accurate victim locations but the left one offers a lot more information about the victim, ranging from the sex, the condition, the priority given to get rescued, the ease of accessibility, etc... Similarly, both maps provide paths to reach the victims but the paths of the left map are inaccurate, going through a section of unexplored space. Based on this example, the right map would get a better score.

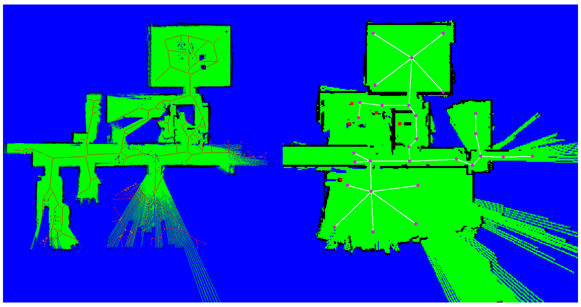
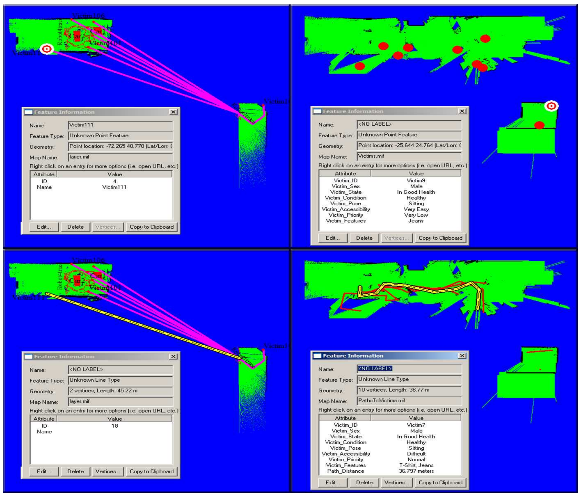


Figure 5: Example for the Skeleton Quality evaluation, with two different robot-generated maps.



Feature Information

Name: Victim111
Feature Type: Unknown Point Feature
Geometry: Point location: -72.265 40.770 (Lat/Lon: (
Map Name: layer.mif
Right click on an entry for more options (i.e. open URL, etc.)

Attribute	Value
ID	4
Name	Victim111

Feature Information

Name: <NO LABEL>
Feature Type: Unknown Point Feature
Geometry: Point location: -25.644 24.764 (Lat/Lon: (
Map Name: Victims.mif
Right click on an entry for more options (i.e. open URL, etc.)

Attribute	Value
Victim_ID	Victim9
Victim_Sex	Male
Victim_State	In Good Health
Victim_Condition	Healthy
Victim_Pose	Sitting
Victim_Accessibility	Very Easy
Victim_Priority	Very Low
Victim_Features	Jeans

Feature Information

Name: <NO LABEL>
Feature Type: Unknown Line Type
Geometry: 2 vertices, Length: 45.22 m
Map Name: layer.mif
Right click on an entry for more options (i.e. open URL, etc.)

Attribute	Value
ID	18
Name	

Feature Information

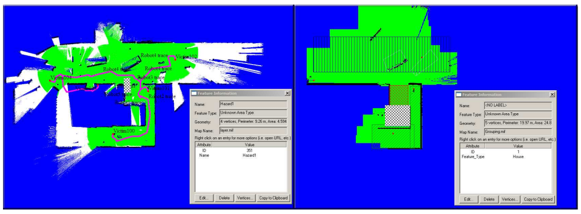
Name: <NO LABEL>
Feature Type: Unknown Line Type
Geometry: 10 vertices, Length: 36.77 m
Map Name: PathsToVictims.mif
Right click on an entry for more options (i.e. open URL, etc.)

Attribute	Value
Victim_ID	Victim7
Victim_Sex	Male
Victim_State	In Good Health
Victim_Condition	Healthy
Victim_Pose	Sitting
Victim_Accessibility	Difficult
Victim_Priority	Normal
Victim_Features	T-Shirt, Jeans
Path_Distance	36.797 meters

Figure 6: Example for the Attribution metric for two different-robot generated maps. The left and right columns each represent a different robot-generated map. The first row shows the victims' attribution while the second row shows the victim paths' attribution.

4.4 Grouping

The Grouping metric is very similar to the Attribution in that it is, essentially, a feature-based map aimed at helping first-responders better navigate the environment. It differs in that instead of being point-based, it groups and labels regions of space. Grouping stems from the fact that a section of occupied pixels represents particular landmarks that can be labeled. Fig. 7 offers a contrasting example, where the left map is comprised of a single group labeled ”Hazard” and the right map contains many different groups labeled as ”House”, ”Street”, ”Vehicle”, among others. Once again, the metric is scored based on the amount and accuracy of the information provided and, in this example, the right map would receive a better score than the first one.



Name: Hazard1
Feature Type: Unknown Area Type
Geometry: 4 vertices, Perimeter: 9.26 m, Area: 4.594
Map Name: layer.mif
Right click on an entry for more options (i.e. open URL, etc.)

Attribute	Value
ID	351
Name	Hazard1

Name: <NO LABEL>
Feature Type: Unknown Area Type
Geometry: 5 vertices, Perimeter: 19.97 m, Area: 24.8
Map Name: Grouping.mif
Right click on an entry for more options (i.e. open URL, etc.)

Attribute	Value
ID	1
Feature_Type	House

Figure 7: Grouping example with two different robot-generated maps.

4.5 Utility

The map Utility takes a look at the overall information provided by the teams. In other words, the map Utility aims at answering the question of how useful are all the layers to a first responder. This metric regroups the other metrics together but looks at a larger scope, where teams have to balance the amount of information they provide with the way it would look on the screen. As more and more information is given, it is harder to display it neatly while still making it easy to understand. The clever use of layers greatly affects the utility of a given map.

4.6 Creativity

For the purpose of the competition, we have added an unorthodox metric that rewards teams for creative new ways of representing valuable information to first responders. Teams are given bonus points for innovative map layers that could help first-responders better do their jobs. In the past, a team came up with the geo-referencing of victims’ pictures, a layer that was quickly adopted by the rest of teams in later competitions. More recently, a team showed the best communication coverage attained while navigating the environment so that first-responders could replicate it should they need to establish a communication network. An example is shown in Fig. 8.

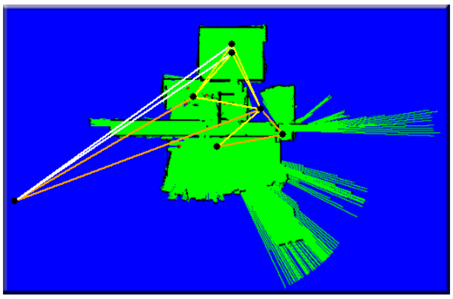


Figure 8: Example of a successful Creativity metric, displaying a communication network. Each transmitter is shown as a point with the lines showing the connections between each link. The point in the left represents the base station.

5. CONCLUSIONS

We have presented the necessary steps to come up with a fair map benchmarking suite capable of scoring maps produced by USAR robots working in close cooperation with first responders. We strongly believe in committing to easily-adoptable, yet powerful, open standards such as GeoTIFF that take little additional work from programmers while providing great benefits. Similarly, we value open-source development by requiring teams to provide public access to their software and encouraging participants to share and reuse code and ideas. In that sense, the competition can be viewed more as an open workshop where teams are equally looking to learn as they are to win. From a benchmarking standpoint, the open-source phenomenon brings an interesting component, where algorithmic progress can easily be measured from year to year thanks to the fact that the software is both available and archived. We hope that the community would follow in our footsteps and make algorithms and data sets public, so that benchmarks can be accepted and evaluated, by an entire community rather than a relatively small research group. Two projects going in that direction are OpenSLAM [9] and Radish [11]. OpenSLAM provides open-source SLAM algorithms and Radish offers data sets. While we praise both initiatives, they are not as extensively used as they should and are missing benchmark tools that would be used to evaluate the quality of the SLAM algorithms (from a localization or a mapping standpoint) for specific applications.

Throughout the years, we have devoted our map benchmarking endeavors to planar occupancy grid maps comprised of certainty values (i.e. either 0 for free space or 1 for occupied space). While this restriction has been reasonable over the last few years, mainly due to the popularity of occupancy grid maps, a surge of newly fashionable robotic platforms ranging from underwater robots to unmanned air vehicles coupled with highly three-dimensional terrain is slowly making two-dimensional occupancy grid maps inadequate. Indeed, teams have already started to explore three-dimensional mapping algorithms [6]. Evidently, the switch from two to three dimensional mapping is not straightforward in terms of map benchmarking and offers an interesting research avenue for future work. Furthermore, it is important to note that three-dimensional mapping does not have a map representation that is well recognized throughout the robotics community and that occupancy grids do not offer an easy transfer from two to three dimensions due to the increase of space and time complexities. We contend that more work needs to be achieved to come up with a community-accepted standard representation for three-dimensional maps.

All things considered, a general ”all-purpose” mapping benchmark is still far from being developed due to the aforementioned problems of map representation, algorithmic differences, lack of open-source data or algorithms, and application dependability. We are convinced that mapping benchmarks need to be tied to the application at hand and, as such, do not see a generalized map benchmark in the near-future. It is rewarding to see, however, that there is an awareness increase as to the importance of the problem and hope that this paper will help steer map benchmarking towards the right direction.

Acknowledgments

An extended version of this paper appeared in [1].

6. REFERENCES

[1] B. Balaguer, S. Balakirsky, S. Carpin, and A. Visser. Evaluating maps produced by urban search and rescue robots: Lessons learned from robocup. *Autonomous Robots*, 2009.

[2] S. Balakirsky, S. Carpin, and A. Visser. Evaluating the robocup 2009 virtual robot rescue competition. In *Proceedings of PerMIS*, 2009.

[3] A. Birk. Learning geometric concepts with an evolutionary algorithm. In *The Fifth Annual Conference on Evolutionary Programming*, 1996.

[4] J. Carlson, R. Murphy, S. Christopher, and J. Casper. Conflict metric as a measure of sensing quality. In *IEEE International Conference on Robotics and Automation*, 2005.

[5] S. Carpin, M. Lewis, J. Wang, S. Balakirsky, and C. Scrapper. USARSim: a robot simulator for research and education. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 1400–1405, 2007.

[6] P. de la Puente, A. Valero, and D. Rodriguez-Losada. 3D mapping: testing algorithms and discovering new ideas with USARSim. In *Proceedings of the IROS workshop Robots, Games, and Research: Success stories in USARSim*, 2009.

[7] I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh. *Feature Extraction, Foundations and Applications.* Series Studies in Fuzziness and Soft Computing. Springer, 2006.

[8] M. C. Martin and H. P. Moravec. Robot Evidence Grids. Technical Report CMU-RI-TR-96-06, Robotics Institute - Carnegie Mellon University, March 1996.

[9] OpenSlam. http://www.openslam.org, 2009.

[10] S. O'Sullivan. An empirical evaluation of map building methodologies in mobile robotics using the feature prediction sonar noise filter and metric grip map benchmarking suite. Master's thesis, University of Limerick, 2003.

[11] Radish – the robotics data set repository. http://radish.sourceforge.net, 2009.