Evaluating Maps Produced by Urban Search and Rescue Robots: Lessons Learned from RoboCup

Benjamin Balaguer, Stephen Balakirsky, Stefano Carpin, and Arnoud Visser*

Abstract

This paper presents the map evaluation methodology developed for the Virtual Robots Rescue competition held as part of the international RoboCup 2008 robotics event. The procedure aims to evaluate the quality of maps produced by multi-robot systems with respect to a number of factors, including usability, exploration, annotation and other aspects relevant to robots and first responders. In addition to the design choices, we illustrate practical examples of maps and scores coming from the latest RoboCup contest, outlining strengths and weaknesses of our modus operandi. We also show how a benchmarking methodology developed for a simulation testbed effortlessly and faithfully transfers to maps built by a real robot. A number of conclusions may be derived from the experience reported in this paper and a thorough discussion is offered.

1 Introduction

RoboCup has demonstrated itself to be an inspiring event capable of accelerating research in a variety of robotic tasks beyond the original robotic soccer scope. It now reaches out to other domains such as service robotics and Urban Search And Rescue (USAR). These additional areas in general, and USAR in particular, call for the deployment of fieldable systems capable of mapping unknown environments. These maps are intended to not only be valuable for robot navigation, but also to provide useful information for first responders trying to reach victims promptly and without exposing themselves to unnecessary risks. RoboCup USAR competitions are divided into two branches. The Rescue Robot League aims to deploy physical robots operating in *arenas*, the purpose of which is to provide repeatable test methods including a variety of mobility and sensing challenges [28]. To date, arenas are, due to logistical difficulties, fairly limited indoor environments extending up to a few hundred square meters. The Rescue Simulation League has the goal of promoting research in cooperative problem-solving by utilizing numerous agents operating in a simulated city-sized scenario. On one hand, the Rescue Robot League evaluates single robotic platforms, or small teams, on their low-level capabilities, like mobility, sensing, mapping, safe navigation and human-robot interfaces. On the other hand, the Rescue Simulation League focuses on high-level tasks, like real-time multi-agent planning, agent heterogeneity, learning and complete autonomy. Due to the evident discrepancy between the research agendas of these two research communities that are pursuing the same final goal (i.e. minimizing casualties in urban disasters), the Virtual Robots Rescue League was introduced to bridge these two groups and promote a fruitful cross fertilization of ideas [7, 15, 18, 39]. The Virtual Robot Rescue League is based on USARSim [16], a high fidelity multi-robot simulator that includes models of numerous robots and sensors, the majority of which are used in the Rescue Robot League (additional applicative scenarios are the ICRA Space Robotics Challenge [12] and the IEEE/NIST Virtual Manufacturing Automation Competition). Relieving participants from the requirement of owning expensive robots, which creates a lower entry barrier, teams participating in the Virtual Robots Rescue competition routinely deploy groups of robots much larger than those in the Rescue Robot League. Moreover, thanks to the effortless possibility of adding virtual sensors or actuators to the simulated robots, various mapping algorithms have been either developed from scratch, reused as is, or enhanced and incorporated into more complex control

*Benjamin Balaguer and Stefano Carpin are with the University of California, Merced, USA. Stephen Balakirsky is with the National Institute of Standards and Technology, USA. Arnoud Visser is with the Universiteit van Amsterdam, the Netherlands.

systems. As a consequence, while organizing and running this competition, we faced the challenge of ranking maps produced by various multi-robot teams using a tremendous variety of heterogeneous approaches. After the first exploratory stage, where maps were essentially inspected visually, the necessity to quickly adopt an accountable and repeatable method to grade maps according to some criteria became apparent. Effectively, visual inspection is too subjective of a method to be applied in a competitive environment where winners have to be selected. Unfortunately, at that time and still nowadays, no solution was available and we therefore developed a methodology from scratch. Interestingly, some of the ideas generated in the Virtual Robots Rescue competition were later extended to the Rescue Robot League; a testament to the value of the high fidelity simulation infrastructure sustaining the simulation competition.

In this paper, we illustrate the principles governing the scoring methods we developed through the years, and we display some of the results collected during the last competition held in 2008. In no way do we claim the adopted methodology is the best possible at the moment. However, the very nature of the competition has put us in the privileged and challenging position of comparing and ranking maps produced by different research groups embracing a variety of heterogeneous software and algorithmic tools. Another peculiar aspect of our endeavor resides in the simulated nature of the task. Thanks to it, we have the possibility of easily accessing ground truth data, a difficult asset to accurately obtain in real world experiments. The necessity to collect and evaluate input from different competitors has also pushed us to embrace an approach valuing open tools and easy to adopt standards, rather than picking an arbitrary one serving our specific research needs. We believe the contest offered us the opportunity to perform an analytic comparison exercise that, to the best of our knowledge, has never been matched in terms of size and variety. The goal of this paper is to detail our lessons learned and outline some interesting research questions for the future.

The paper is organized as follows. In Section 2 we present literature related to the task of analytic map comparison, a topic that deserved little attention to date, but that is gaining momentum. Section 3 presents the method we used in order to rank maps and accounts for the practical choices we embraced. Detailed results coming from the RoboCup 2008 event are offered in Section 4, where maps produced by the various teams are contrasted and evaluated. In Section 5 we compare our metric with some of the methods cited in Section 2, outlining similarities and differences. One of our main goals while developing the Virtual Robots Rescue competition and USARSim has always been to keep the simulation framework coupled with real robotic systems as much as possible. For this reason, in Section 6, we illustrate how the proposed evaluation procedure ranks maps produced by a real robot and contrast the real results with their simulated counterpart. Finally, in Section 7, we outline the lessons learned and identify some open questions worth additional investigation.

2 Related Work

Robot benchmarking and performance evaluation is a recent research thread so the amount of available scholarly work in this area is generally limited. Localization and mapping is no exception to this trend and we describe the few attempts made in establishing both benchmark problems and evaluation metrics.

In a recent paper, the RawSeed project is presented [23]. RawSeed is an initiative aimed at addressing the very problems dealt with in this special issue, namely to establish independent representation and benchmarks for SLAM algorithms. The project is still in its infancy, but the paper outlines some principles that we subscribe to and that we resume in the next section, namely the importance to tie the evaluation method to the robotics task under examination. In [1], Abdallah et al. propose a benchmarking infrastructure for the special case of visual SLAM in outdoor environments. The authors also stress the importance of tailoring the benchmarking process to the specific robotic task considered, realizing the inherent difficulty in comparing, for example, indoor and outdoor visual SLAM algorithms. The paper, however, does not provide an assessment metric. A similar effort targeting vision-based robotic localization is presented in [24], where the authors distinguish between topological and metric localization and mainly rely on omnidirectional cameras. In contrast to the previous contribution, they propose an evaluation formula taking into account the size of the dataset used for localization, the resolution of the sensor and the incurred errors.

Major contributions to the map benchmarking research area, and one of the few efforts to quantitatively establish the effectiveness of mapping algorithms producing occupancy grids, came from Collins et al. [19–21, 35–37]. The authors establish their benchmarking suite based on three individual components: 1) a map cross correlation as proposed in image processing literature [9]; 2) the Map Score measure invented by Martin and Moravec [30], along with a slightly modified version; 3) path generation on the robot-generated map to see if such paths would be valid on the ground truth map. The map cross correlation in the benchmark could alternatively be replaced by the more extensively exploited Pearson's Correlation coefficient [27]. Similarly, the Map Score measure could have equally been substituted by the Overall Error metric [14], where error is measured rather than accuracy. Nevertheless, and while we agree with the usefulness of breaking down a benchmark into multiple categories, we find the image correlation and Map Score procedures questionable. These two components are likely to negatively affect maps with a single misalignment that propagates through the rest of the map. The path generation component is both creative and useful for robot navigation but is not necessary for USAR applications, where first responders might end up using the robot-generated map (as opposed to another robot). Collins et al. exploit their benchmarking suite in a series of practical applications [20, 21, 36, 37], ranging from the assessment of map quality to the comparison of different mapping algorithms and sensor models. The conclusions made in each paper are supplemented by hundreds of map evaluations thanks to the use of simulation supplemented by real experiments, a mindset that we also support and advocate. Moreover, and in accordance to the spirit embraced in the USAR competition, they realize that it is necessary to assess the utility of a map as a tool for robotic navigation, and that this assessment may substantially differ from a purely metric analysis. They, however, limit the discussion to robot navigation that we extend to first responders where, for example, a robot may produce a bent map that may be penalized from a metric point of view, but that may still be highly valuable for a human operator or for navigational purposes. A very similar contribution was presented by Varsadan et al. in [43], where an image similarity metric [11] is used to compare robot-produced maps against their ground truth equivalents. Additionally, the authors show that the image similarity metric provides comparable results to the use of a Least Mean Squared Euclidean distance metric between map and ground truth points. The same conclusions can be drawn from this paper as what is presented, more extensively, by Collins et al. in [19, 35]

In [8], Baltes offers a general discussion about robot benchmarks, questioning the uncritical use of robot competitions in order to assess robot performance and suggesting three numerical metrics for tasks related to motion. Michel et al., however, claim that ”RoboCup is recognized as a reference benchmark in robotics” [31]. We subscribe to Baltes's statement that competitions do not provide the ultimate benchmark for robotic performance. Indeed, the final result of a competition is a ranking that might wrongfully reward stable control systems to the detriment of innovative ones. Michel et al. also propose the creation of a database of source code that could be used to measure the progress of algorithms over time, a practice that we have already implemented since the finalists of our competition are required to publicly share their code. Calisi et al. also focus on the benchmarking of motion tasks by providing a set of metrics, while promoting the use of simulation, competitions, and data repositories [13]. More notably, the authors bring in the notion of time as part of their performance metric. The notion of time, in addition to power consumption, is also proposed by Basilico et al. as an efficiency metric [10]. Additional time-related metrics have been proposed in [2, 3, 41, 42], where the number of sensing operations (i.e. time steps) is used in conjunction with the robot's path information to compare the performances of different exploration strategies. Even though time and power are not factors for our benchmark, because all maps are generated from a twenty-minute run and fifteen minutes of map processing time, we recognize the importance of including time as part of a map benchmark framework. Evidently, comparing a map created in real-time with one that was post-processed for hours will yield drastically different results. More recently, we have looked into the possibility of having shorter runs for robots with heavy sensor loads that would require more power. We are still in the early stages of developing such a metric and do not include it in this paper. Egerton and Callaghan define the Lost Metric [22], i.e. a metric aimed to measure the ability of a robot to re-localize itself in a given map after it lost localization. This metric is formulated for robots using a perceptive inference map, i.e. a model appropriate for biologically inspired robots, and does not seem immediately usable for USAR tasks.

Two initiatives not focused on SLAM benchmarking but central to our vision of robot performance metrics are the web projects OpenSLAM [34] and Radish [38]. The former, OpenSLAM, is a website providing ready-to-use implementations of various SLAM algorithms. A comparison among their performances is however not immediate because they rely on different hypotheses and representations for input data and they produce maps in a variety of formats. The latter, Radish, is a repository of data. As such, it hosts some recordings of robotic runs that may be considered shared benchmarks. What is missing, however, is a consistent data format and a taxonomy of the available data sets, in terms of the algorithmic challenges they offer.

3 A Scoring Methodology for Urban Search and Rescue

3.1 Map Format

It is said in [23] that any map assessment method should be intimately tied to the practical task for which the map will eventually be used. We subscribe to this statement and add that the task itself also influences the map format. The goal of every robot team participating in RoboCup USAR competitions is to produce a single map providing information valuable to both robots and first responders. For robots, the most valuable contribution is to provide information enabling navigation in the environment. It should be noticed that the competition does not mandate the use of fully autonomous robot teams, therefore robots may move autonomously or under human control. First responders, on the other hand, use the maps for different needs, namely to quickly gain access to victims while minimizing their exposure to risky areas. Even though humans could certainly move in the environment solely relying on a robot-generated occupancy grid map, it is clear that distinctive landmarks (e.g. the victim is near the red car) or topological information (e.g. there is a victim in the third room down the corridor) will be more useful than the information as to whether a certain part of the environment is not traversable due to a robot mobility limitation. Landmarks that are useful for human navigation include environmental features (e.g. walls, doors, stairs, slopes), topological information, hazardous areas, and victims' locations. The heterogeneity of the desired information calls for a map embedding different layers. As such, and as an example, one layer might represent the occupancy grid map while other layers might display topological information or victim depiction. First responders can then display only the layers in which they are interested. Teams are therefore required to deliver their maps in GeoTIFF format [25] for raster data and MIF format for vector data. GeoTIFF is chosen for three reasons. Firstly, being built atop the TIFF format, it allows the inclusion of multiple layers, thus serving the purpose of conveying different information channels, as formerly specified. Secondly, GeoTIFF has been designed with the very goal of embedding georeferences inside the file itself. An additional benefit stemming from this fact is the possibility of superimposing map layers to ground truth maps. These ground truth maps come from the simulated environment, when using simulation, or can be hand-drawn or available as blue prints, for real experiments. Finally, GeoTIFF is an open standard, and there exists open source packages and libraries to read, write and visualize this file format. As in GeoTIFF, the MIF vector format was chosen for several reasons. Firstly, it is an ASCII format that is easy to read and simple to parse. In addition, as in GeoTIFF, the MIF format embeds the georeference information in the file and there is a large variety of open source tools and libraries that are able to read and write the format.

For the mandatory layer displaying information pertaining to robot navigation, we embrace an occupancy grid representation since it is the representation mostly used for robots exploring terrestrial areas [33]. Given the task at hand, teams participating in the competition are required to categorize grid cells in four different ways. A cell can be: *untraversable*, *unknown*, *cleared* (i.e. traversable and known to be victim free), or *uncleared* (i.e. traversable but not necessarily victim free). Each cell type is assigned a precise color specified by its RGB components, so that no ambiguity arises while generating or inspecting cells. To be precise, and also to appreciate the results presented in the following sections, untraversable areas are black, cleared are green, uncleared are white, and unexplored are blue. Additionally, locations where victims are present are marked as red. This information has to be included in the navigation layer because these cells are obviously non-traversable.

In addition to the occupancy grid color requirements, the map layer has to be georeferenced, a requisite imposed for two main reasons. First, from a practical aspect, georeferenced maps can be overlaid on a georeferenced ground truth map. This superimposition not only facilitates benchmarking, but also allows for the deduction of where potential mapping problems arise (e.g. by observing misalignments) and, as a consequence, deduce weaknesses in the mapping algorithm utilized. Second, from a more theoretical standpoint, robots map spatial areas that have a physical location. As such, maps should not be floating in free space, but should be referenced with respect to a point, whether it be, in the case of simulation, fictitious, or, in the case of real robots, measured, extrapolated from software, or given by a GPS receiver or some other source of information. All things considered, the method chosen for georeferencing maps does not matter as long as it is performed consistently across all maps. Georeferencing then becomes a formidable tool, allowing the juxtaposition of maps of the same environment gathered using different algorithms, displaying maps produced from different parts of a building, or, on a larger scale, viewing maps produced by distinctive robots deployed in different nearby buildings. Evidently, georeferencing is only marginally beneficial when dealing with a single map.

No specifications have been set for other raster layers. The only requirements are the usage of a color schema not interfering with the colors specified for the navigation layer, and that they should be georeferenced as well. This specification, or lack thereof, is intentional. We have found that by having an open policy for additional layers, participants can freely explore creative ways of representing information. In fact, after the competition the most innovative layers are presented to all participants, strongly recommended, and could become mandatory in subsequent competitions.

For vector information, the only required product is a list of points that represent victims that have been located in the environment. Extra points are available for additional information. Many of the teams include a vector representation of the path that their robots traversed, some form of skeleton representation of the world, and grouped and labeled objects.

We conclude this section observing that the chosen occupancy grid representation for the basic map goes hand in hand with the GeoTIFF format, in the sense that the standard specifies how to reference raster images. In particular, the GeoTIFF format enables one to specify the pixel resolution in meters, thus offering a consistent way to overlay and compare maps produced by different mapping systems based on potentially different scales.

3.2 Map Assessment

Before getting into details about map quality assessment, it is worth remembering that mapping is only one of the components contributing to the final numerical score attributed to teams taking part into the competition. To be precise, teams are scored according to the following formula:

$$S = \frac{E \cdot 50 + M \cdot 50 + \sum_{i=1}^{n} V_i \cdot 20}{O^2} \tag{1}$$

where

- E is a number between 0 and 1 accounting for the *cleared* part of the map (i.e. obstacle and victim free).
- M is a number between 0 and 1 accounting for map quality.
- n is the number of discovered victims.
- V_i is a number between 0 and 1 accounting for the information provided about the ith victim.
- O is an integer greater than 0 indicating the number of human operators supervising the robot team.

It is evident that in the context of a competition where judges eventually need to univocally identify the top three performing teams, the combination of performance measures that are non-commensurable, as evidenced by Eq. 1, is inevitable. The relative weight given to the various components is not necessarily the best or the only possibility, and it grew out of feedback collected from the participants after every competition. In general, one could envision the use of multi-criteria evaluation functions, and to identify to Pareto optimal maps. This option, although valuable, has not been pursued because of its difficult application in the context of a competition.

We now discuss in detail the three components relevant to mapping, i.e. E, M and V_i. We note that the O variable was included in the formula in a successful attempt to promote autonomy by penalizing multi-operator systems. For the purpose of this discussion, and since the number of operators is irrelevant to our mapping benchmark, one can think of O as always being set to 1 (i.e. we assume that one operator is needed to set up the robots and start the control system, regardless of whether the robots are fully autonomous, semi-autonomous, or fully tele-operated).
For the E factor, which stands for *exploration*, maps are scored based on the correctly reported cleared surface, measured in square meters. This evaluation can be done automatically by counting the number of green pixels on the map under evaluation. Thanks to the provided georeferencing information, the counted number of pixels immediately translates to an area measure. For sake of completeness, we mention that the team exploring the largest surface gets an E value equal to 1 (and thus gets the full 50 points available for exploration), and lower values linearly decrease to 0. In order to account for possible errors in the surface classification, a 5-point penalty is subtracted for every victim present in an area declared as *cleared*.
The M factor accounts for *map quality*. Keeping in mind that the map should be equally useful for first responders and robots, five criteria that are detailed in Section 4 are considered. Each criterion is considered equally important. *Utility* rewards the presence of information valuable to first responders, such as the location of hazards. *Skeleton quality* aims to reward the ability to reduce a complex map into a set of connected locations. For example, a corridor with many doors on the sides may be represented by a skeleton with a line and various symbols for the doors. In sense, one may say that the skeleton quality aims to reward the ability to extract a topological representation [29] from the occupancy grid. *Metric quality* measures the accuracy of the map when compared with ground truth. *Attribution* rewards the embedding of attributes added to the map, such as the path followed by robots while exploring the environment. Finally, *grouping* rewards the ability of adding an additional layer displaying that four connected segments make up a room, a hallway, a car, an open area, and so on. Elements pertaining to utility, attribution, skeleton quality and grouping are delivered as additional layers for the TIFF file.
For what concerns V_i, various information is considered for scoring purposes. From a mapping point of view the most important aspect is the location of the victim. Since locations are marked as red pixels in georeferenced maps and victims are embedded with unique identifiers, it is elementary to measure the difference between the reported victim location and the true one. The reported victim location is compared with the ground truth victim location and, if the error falls below a certain threshold, points are awarded.

4 Examples from Competitions

The metric described in Section 3.2 is currently in use for judging the annual RoboCup Rescue Virtual Competition event. This section and the next one present results from the 2008 competition that took place in Suzhou China in July of 2008. The maps shown in Fig. 2, 3, 4 show examples of the maps that were reviewed during the event. These maps were generated while operating in the environment depicted in Fig. 1 which shows a scene from the simulated world that matches the generated maps. In Fig. 2, 3, 4, the competitors' maps have been overlaid on top of ground truth. The fact that maps were georeferenced made this task simple to perform. Judges were then able to easily see small errors in the structure of the autonomously generated map. For example, in the second map in Fig. 2 the horizontal corridor in the top of the image exhibits drift with respect to the ground truth. In addition to the raster layer that depicts the regions that have been cleared and explored, vector information on such items as the robots' path, a potential path for a first responder to take to reach a discovered victim, and skeleton information is provided (see Fig. 3 and 4). Scores for the maps were not performed on an individual basis. Instead, maps were compared against each other to determine the final scores.

Fig. 2 shows the comparison of a map from Team A to a map from Team C. The first step in the map comparison is to automatically determine the area of the regions of the map that had been explored and cleared. In this case, Team A cleared 211 m^2 and explored 373 m^2 while Team C cleared 257 m^2 and explored 351 m^2. It should be noted that if a team reports an area as explored that could not have been seen by the robot, this area would be removed before the automatic computation was performed.

The next phase of scoring involves determining the mapping quality. Recall that the map quality is made up of criteria including metric quality, skeleton quality, attribution, and utility.

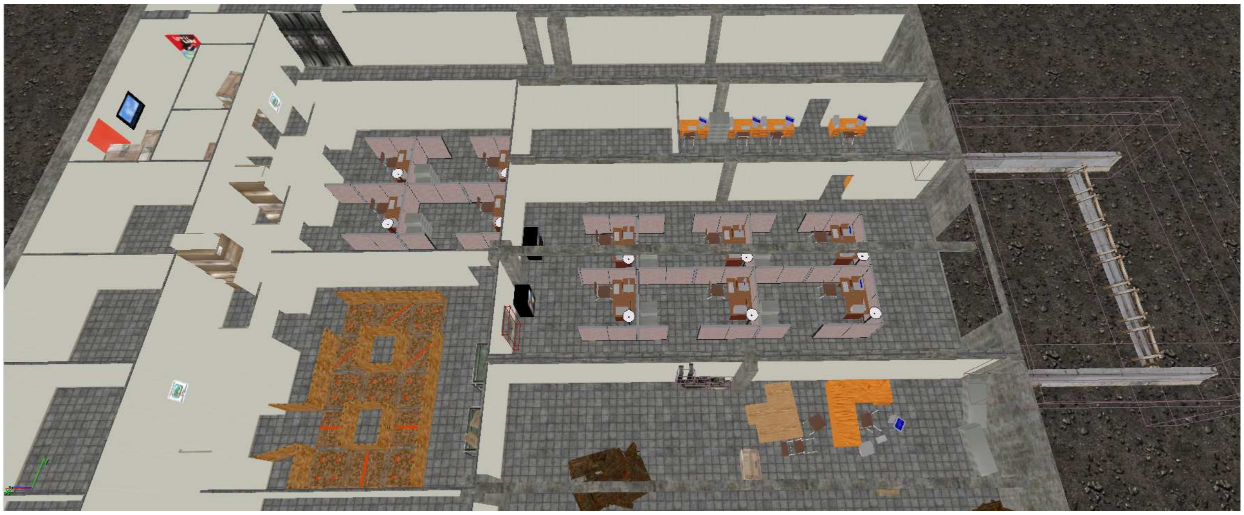

Figure 1: A view of the world that the robots were exploring while creating their maps. The results in Fig. 2, 3, 4 and 5 are all based on this environment.

4.1 Metric Quality

The metric quality measure is designed to reward teams that are able to accurately determine the location of walls, obstacles, and free space in the environment. The determination of a score for this metric is made difficult by the fact that a mistake in robot localization often causes a large error in the metric accuracy of the map. If a localization error occurs early in the run, it will affect the entire map; if it occurs late in the run, it will affect a much smaller percentage of the overall map. However, it is the opinion of the competition's technical committee that both maps may be deserving of the same metric quality score since they both experienced one severe localization error. This notion eliminates simple scoring techniques such as correlation with ground truth (see also the discussion in the following section).

In order to fairly judge the metric quality, it was decided that both local quality and global quality should be assessed. Local map quality is the quality of the map between localization errors. Features such as wall consistency (i.e. the presence of holes or gaps in the walls), double representation of objects, false positives, and false negatives are taken into account. The global assessment examines the number and severity of the localization errors. Several of these classes of errors may be seen in Fig. 2. Team C's map shows a significant global error that occurred in the exploration of the upper hallway. There are also several local errors including missing walls and obstacles in the cubical area of the map (see Fig. 1 for a visual representation of the world that was explored). Points were awarded for metric quality by assigning the best map 100% of the available points and decreasing the points awarded by map rank. In this case, Team A received a 100% score while Team B received 75% of the available points.

4.2 Skeleton Quality

Teams were asked to provide a vector layer that contains a map skeleton for judging. Fig. 4 shows Team A's skeleton map overlaid on top of ground truth with the addition of a suggested path to the victim. The idea behind a skeletal map is that it provides all of the information that would be necessary for a human to navigate in the space. For example, from examining the map in Fig. 4, one could give instructions to reach the victim as follows: Enter the large door to the west, followed by a left turn and go straight until you reach the wall. Make a right turn and enter through the second door on your right to find the victim.

Unfortunately, there is a wall missing in this map and the responder should have entered the third doorway to locate the victim. Scoring for the vector map was also performed by ranking the maps from best to worst and assigning credit proportional to the rank. In the judges' opinions, no team was fully successful in delivering a skeleton layer.

4.3 Attribution and Grouping

One of the reasons to generate a map is to convey information to first responders. This information is often represented as attributes on the map. Teams were required to denote areas explored (gray color on map examples), areas cleared of victims (green color on map examples), and victim locations. The competition's definition of "cleared" means that no undetected victims exist in that area. Therefore, teams received penalties for any victims that were located in cleared areas and that were not reported. Other than for these restrictions, teams were free to include any additional map attributes that they found useful and, overall, teams were very successful in providing a layer that contained the map attributes and groups. Grouping is a higher order mapping task used to recognize that discrete elements of a map constitute larger features. For example the fact that a set of walls makes up a room, or a particular set of obstacles is really a car. Fig. 3 shows Team A's and C's fully detailed map. This map layer was generated by the human operator while performing the mission. Most teams provided a tool that allowed the operator to sketch bounding areas on the map and label them as the robots proceeded. Team A added the additional innovation of being able to include georeferenced snapshots of interesting regions that were captured by the robot. A large number of teams produced equivalent feature classes for the attribution, such as cubicle, hallway, step field and open space. Team A was awarded additional points for being the only team with the innovation of including snapshots. It is anticipated that most teams will incorporate this feature in subsequent competitions.

4.4 Utility

The idea behind judging the utility of the maps was to tie all of the other metrics together and imagine that a first responder was using the map to find an object of interest in the world. The basic question posed was "how useful will the map be in helping to accomplish the task?". Fig. 3 shows the combined map constructed from the various data layers of Team A and Team C. Team A performed an excellent job of delivering a map that could be used by a first responder to reach the detected victim. It provides the precise path that should be taken, and shows landmarks along the way (e.g. the cubicle) that could be used to track the responders position. While Team C also provides numerous landmarks that would be useful for navigation, it is not clear how a responder would get to the detected victim (the red 'X' on the left of the figure). Would the better path be to take the top hallway or go around the cubicles in the center of the map?

5 Comparison with Previous Metrics

In this section, we quantitatively compare our benchmark results against a variety of metrics, all of which have been discussed in Section 2. Each metric evaluates occupancy grid maps, which were created by post-processing the submitted basic maps into binary images where each cell is either occupied (i.e. a value of 1) or free (i.e. a value of 0). Similarly, the ground truth maps were easily extracted from the simulation environment, with each cell also having one of two values: 1 for occupied space and 0 for free space. It is worthwhile to note that all comments made in this section regarding the metrics are based on our map representation and that different observation could be made if cells indicated probabilities (i.e. a value between 0 and 1) rather than certainties (i.e. either 0 or 1). We name our results *Map Set 1* and *Map Set 2* and will use these names throughout the rest of this section. The occupancy grid maps are illustrated in Fig. 5 and Fig. 6 with the metrics comparison being displayed in Table 1 and Table 2 for *Map Set 1* and *Map Set 2*, respectively.

Metric	**Best**	**Team A**	**Team B**	**Team C**	**Team D**
Area Explored (m^2)	1000	373	286	351	**679**
Map Score [30]	539524	489363	491321	**496818**	496758
Overall Error [14]	0	50161	48203	**42706**	42766
Normalized Map Score [35]	0	32513	**26148**	37384	30971
Occupied Map Score [35]	0	17648	22055	**5322**	11795
Baron's Correlation [35]	1	0.2499	**0.361**	0.2243	0.3337
Pearson's Correlation [27]	1	0.7247	**0.8155**	0.0039	0.6422
Picture-Distance-Function [11]	0	22.8908	**13.0171**	25.7268	19.6739
Skeleton Quality	12	12	12	12	12
Metric Quality	12	**12**	10	9	9

Table 1: Metrics comparison for *Map Set 1*. The first row indicates area explored, calculated by our benchmark. The next seven rows are the different metrics used, taken from different publications. The last two rows display our metric scores for the Skeleton and Metric Qualities. Bold numbers represent the best result, as per the metric used. The first column gives the best score possible. The actual maps for this data set can be seen in Fig. 5.

The Map Score [30], Overall Error [14], Normalized Map Score [35], and Occupied Map Score [35] metrics perform pixel-to-pixel comparisons between the ground truth and robot-generated maps. More specifically, the Map Score metric starts with a score of 0 and it increases by one for every pixel that is the same in the ground truth map and the robot-generated map. Consequently, the best possible score for this metric is the total number of pixels in the maps. The Overall Error metric is the opposite of the Map Score metric since it increases by one for each pixel that does not match within the two maps. As such, the best Overall Error is zero, and adding the Map Score and Overall Error metrics together will yield the total number of pixels in the maps. One of the problems with the two aforementioned metrics stems from the fact that they are utilized over all the pixels, regardless of whether or not they represent occupied or free space. Since maps have a majority of free space, these two metrics tend to be biased towards correct free space rather than correct occupied space, a fact that is exemplified by *Map Set 1*. The table shows that Team C gets the best scores for these two metrics and, looking at the figure, we can clearly understand why. Team C produces the map with the most unoccupied cells and the thinnest walls. This problem is seen again in *Map Set 2*, where the second-best map as per the metrics, produced by Team E, is almost unusable. The high score for that map is entirely attributed to the tremendous amount of free space that it encompasses. Alternatively, Team F produces the best map for this metric, a fair result even though only a small portion of the ground truth map has been explored. In an attempt to correct the unoccupied space problem observed in the previous two metrics, the Normalized Map Score and Occupied Map Score metrics yield the same results as the Overall Error metric (i.e. the result increases for each pixel that does not match between the two maps) but are only run on the occupied space of the maps. Two metrics are required since the metric is performed once using the occupied space of the ground truth map and a second time using the occupied space of the robot-generated map. The results between the two are usually quite different since ground truth maps tend to incorporate more occupied space (e.g. larger maps, thicker walls) than their robot-generated counterparts. For *Map Set 1*, we can clearly see that the Normalized Map Score metric promotes thick walls and the amount of occupied space discovered as evidenced by the top two teams, as per this metric, Team B and Team D, respectively. For *Map Set 2*, the same behavior can be observed with the top three teams, Team B, Team A, and Team D, respectively, having the thickest walls and the largest number of walls uncovered. Conversely, the Occupied Map Score metric is the exact opposite measure, by providing better scores for maps that have less occupied space and smaller walls. This fact is evidenced by a complete reversal of the best maps: Team C and Team A for *Map Set 1* and Team E, Team F, and Team C for *Map Set 2*. All things considered, these four metrics are very similar. In fact, the addition of the Normalized Map Score and Occupied Map Score metrics will yield the Overall Error metric.

The next two metrics that we evaluate are correlation coefficients. The first, Baron's Cross Correlation coefficient [35], comes from the template matching research community and evaluates all the pixels in the map. A nice feature of this metric is the fact that it is normalized between 0 and 1, where 1 means perfect correlation. Since Baron's Correlation heavily relies on averaging, it recompenses equal, or close, number of occupied and free cells between the two maps, to the detriment of accuracy between the two maps. This drawback is clearly seen in both Map Sets. For *Map Set 1*, the best Baron's Correlation is achieved by Team B thanks to its thick walls that nicely mimic the ground truth map. *Map Set 2* shows this drawback in even greater detail, where Team A and Team B are the top two teams using this metric. As can clearly be seen from Fig. 6, the map generated by Team B is distorted, but gets the second-best score for its large number of occupied pixels. The second correlation coefficient, Pearson's Correlation coefficient [27], comes from statistics and only evaluates the occupied space in the maps. The Pearson's Correlation coefficient gives a measure, between 0 and 1, of how likely it is possible to infer a map from another, strictly using linear equations. This correlation coefficient comes with two important drawbacks since it requires a similar number of occupied pixels between each map and is easily perturbed by outliers. The striking number for this correlation coefficient in *Map Set 1* comes from Team C, with an extremely low Pearson's coefficient, due to the difference in the number of occupied pixels between ground truth and the map. In *Map Set 2*, another extremely low coefficient is observed for Team D, due to the large amount of noise and outliers in the map. Evidently, neither map is as bad as the Pearson's correlation make it seem. The correlation coefficients provide new metrics that do not follow the overused pixel-to-pixel comparisons, along with normalized values that are easy to understand, but they can easily give unpredictable results.

Finally, we explore the Picture-Distance-Function [11], as a different approach to what has been presented so far. For each occupied pixel in the ground truth map, the closest Manhattan-distance to an occupied pixel in the robot-generated map is calculated. The process is repeated for each occupied pixel in the robot-generated map, using the closest Manhattan-distance to an occupied pixel in the ground truth map. The metric is then the sum of all the Manhattan-distances for each pixel, using both the occupied space of the ground truth and robot-generated maps, divided by the total number of pixels used. Evidently, this metric is not normalized and a perfect map would receive a Picture-Distance-Function score of zero. Even though this metric is a very good attempt at removing the inherent problems of pixel-to-pixel comparisons, it is biased towards exploration, an observation that is substantiated by both Map Sets. Indeed, in *Map Set 1*, the two best teams as per this metric, Team B and Team D, respectively, have discovered the greatest amount of walls through exploration. Similarly, for *Map Set 2*, the two worst teams, Team E and Team F, respectively, have explored the less area or discovered less occupied space.

6 Closing the Loop Between Simulation and Real Robot Systems

A common criticism raised towards robotic simulators is that results obtained in simulation are hard to generalize because proofs of concepts are deemed to succeed in the simplified simulated scenario. Indeed we agree that results obtained in simulation should carefully be considered and we do not view the simulator as a tool to prove concepts, but rather as a development and debugging framework to safely ease the evolution of software that will eventually run on real robots. For this reason, and since the very inception of the USARSim project, special care has been devoted to create experiments aiming at assessing the accuracy of results obtained in simulation [5, 17, 40].

In order to verify that the presented map evaluation schema extends to real systems, we contrast two maps as shown in Fig. 7 and Fig. 8 with their respective assessments performed according to the presented methodology. The first map (left hand side of Fig. 7 and Fig. 8) has been created with a robotic system, while the second one (right hand side of Fig. 7 and Fig. 8) is obtained with the USARSim simulator operating in a corresponding environment. We stress that the two experimental scenarios are *aligned.* With this expression we indicate that we use exactly the same control software, the same robot, and we take

Metric	**Best**	**Team A**	**Team B**	**Team C**	**Team D**	**Team E**	**Team F**
Area Explored (m^2)	1000	352	**720**	488	467	365	488
Map Score [30]	510952	470275	469373	471312	470619	471864	**472174**
Overall Error [14]	0	40677	41579	39640	40333	39088	**38778**
Normalized Map Score [35]	0	34279	**33969**	36542	35204	38074	36570
Occupied Map Score [35]	0	6398	7610	3098	5129	**1014**	2208
Baron's Correlation [35]	1	**0.1861**	0.1839	0.1305	0.1644	0.0684	0.1468
Pearson's Correlation [27]	1	0.3247	0.3992	**0.4363**	0.0593	0.2517	0.3764
Picture-Distance-Function [11]	0	48.9046	44.3993	38.4483	**36.043**	61.9765	58.2449
Skeleton Quality	**12**	12	12	12	12	8	10
Metric Quality	**12**	11	9	11	10	**12**	9

Table 2: Metrics comparison for *Map Set 2*. The first row indicates area explored, calculated by our benchmark. The next seven rows are the different metrics used, taken from different publications. The last two rows display our metric scores for the Skeleton and Metric Qualities. Bold numbers represent the best result, as per the metric used. The first column gives the best score possible. The actual maps for this data set can be seen in Fig. 6.

care that the simulated environment faithfully reproduces the real one. Even though only one example is presented for the sake of space and clarity, many experiments have been run under similar conditions and have shown to be consistent. Fig. 9 presents pictures contrasting the real and simulated robots in their respective environments.

6.1 Robot

The robot platform is a P3AT [1] equipped with a SICK LMS200 proximity range finder and a webcam. The control software [4] is a two-layer application where the user interface is implemented with Microsoft Robotics Studio and the lower layer is implemented using the Mobility Open Architecture Simulation and Tools (MOAST) developed at NIST [32]. For mapping purposes the control application exploits the GMapping algorithm [26], an open source package that produces a probabilistic occupancy grid map using range finder and odometry data. A human supervisor detects and localizes victims based on the webcam feedback, and places them manually on the map provided by GMapping. Finally, the human supervisor manually produces an additional layer containing grouping information, either during or at the end of the run.

In the testbed presented in this section two experiments are performed in sequence, one with the real robot

[1] Certain commercial software and tools are identified in this paper in order to explain our research. Such identification does not imply recommendation or endorsement by the authors, nor does it imply that the software tools identified are necessarily the best available for the purpose.

and one in simulation. Even though it would be possible to drive the two robots in parallel with the same joystick[2], as was done in [5, 6], we do not follow this approach because the single operator would then be subject to a too demanding scenario (i.e. managing two Graphical User Interfaces at the same time). Instead, care has been taken by the operator to ensure the two robots follow the same topological path. This choice is evidenced by the paths displayed in Fig. 7.

6.2 Map Evaluation

If judged in the RoboCup Rescue Virtual Robot Competition, the two maps displayed in Fig. 7 would receive remarkably similar scores. The real robot explored and cleared an area of 185 m^2 while the simulated robot explored and cleared an area of 195 m^2. In terms of metric quality, Fig. 8 show a comparison of the two maps. This figure illustrates the difficulty in judging metric quality. Both maps appear to have similar local accuracy as evident from the bottom corridor where the robots began their missions. However, the simulated robot appears to have had a localization error before reaching the second room on the right (thus affecting global accuracy), while the real robot did not experience a localization error until the third room on the right. In addition, both robots appear to have experienced a consistent drift in their solution, and have a similar error where non-existent portions of the corridor near the top of the map are marked as cleared. Due to the similar nature of the mapping errors, the two maps would have received identical metric scores.

The remainder of the map score is based on skeleton quality (no skeleton was provided), attribution, and utility. In this case the attribution of both maps was identical. Thus the maps would have received identical scores. In addition, the utility of both maps for reaching the located victims is excellent in both cases. This experiment has shown that this mapping algorithm produces almost identical results in both the USARSim simulator and on a real robotic platform.

7 Open Problems, Future Work and Conclusions

A few conclusions can be drawn from the presented results, and we can also identify a set of limitations that may serve as stimulus for future research work.

Throughout our experience, we committed to the use of planar occupancy grid maps in order to model spatial information acquired by exploring robots. This choice proves to be appropriate for the scenario at hand (i.e. USAR applications) but starts to show its limitations due to the increasing popularity of heterogeneous teams of robots capable of overcoming three-dimensional obstacles, or even flying. In contrast to the two-dimensional case, for which occupancy grid maps are well accepted by the research community, notably less agreement exists for an appropriate representation of three-dimensional scenarios, also because the natural extension of grids to three dimensions carries a certain space complexity overhead. In order to promote research towards this very promising area, convergence towards an agreed representation can only be beneficial to compare results and promote reuse of third party code. We conclude this discussion about representation issues by mentioning that, for certain classes of robots, an occupancy grid may not be appropriate at all. For example feature-based maps embedding sparsely detected features may be more useful for aquatic robots.

We believe that one success story of our experience is the commitment to open standards, like GeoTIFF, and we are pleased to see that having enforced participating teams to release their code under open source terms is leading to a rewarding exchange of algorithms and software components among participants. We are convinced advocates that in order to converge to community-accepted benchmarking tests for mapping, but also for other robotic tasks, it is necessary that researchers make their algorithms and data publicly available to the research community for, if nothing else, evaluation purposes. While this requirement may face some resistance in the beginning, this effort will yield big dividends in the long run. Repeatability of experiments is one of the cornerstones of scientific investigation and the research robotics community now has its chance to embrace this much belated change.

[2] Experiments like these have been successfully performed in the past. The reader is referred to the videos available on https://robotics.ucmerced.edu/Robotics/multimedia for some examples.

The quest for a mapping benchmark is far from being close to an end, but increased awareness of this problem's importance leads to some optimism. A set of benchmarking algorithms with no or few parameters to set is still missing. While running the RoboCup competition, where teams eventually are rewarded with prizes for their efforts, we saw that the evaluation of the E and V_i components of Eq. 1 are well received by participants because no subjective human judgment is involved. On the contrary, and understandably, the M value is still evaluated according to some subjective deliberation, and does not lead to the repeatability we strive for. The definition of a community-accepted formula for contrasting the metric quality of two occupancy grid maps, or of a map against ground truth is in our opinion the most challenging issue.

As a first attempt of objective evaluation, in 2009 the occupancy grid maps will be automatically judged on usefulness. This approach is inspired by [21]. The usefulness of the map for robot navigation will be tested by planning a number of paths to predefined reachable locations. The fraction of correct paths as a function of the possible path will be used as measure. Path planning can fail due to the following map characteristics:

- no occupancy information is available for the target point (i.e. map is too small).

- every possible path is blocked by occupied space, due to observed obstacles that are actually not present in the world.

- a path is found to the target point, but the path is dangerously long due to obstacles observed or gateways that are missed (a detour). This classification will be based on a relative threshold (e.g. 20% longer than the optimal path).

- a path is found to the target point, but the path is dangerously short due to missed obstacles or observed gateways which are not present in the world (a shortcut). Executing this path would crash a robot. This classification will be based on a relative threshold (e.g. 5% shorter than the optimal path).

Even with this objective criterion, the problem is not solved. In this approach, maps are treated as images and analyzed with machinery coming from computer vision. As evidenced in Section 5, treating occupancy grid maps as pictures has a number of drawbacks. Yet, the navigational skills of the robot team are still an important aspect since they determine the size and accuracy of the map. The robot team has full control over its input data, engaged in active perception of its surroundings. There is no way to directly assess the real power of a mapping algorithm that, in real-time, adjusts to the availability, accuracy and power of its sensing devices. A comparison with algorithm complexity may be helpful to get our point. One of the reasons for the success of the famous big-O notation is its relationship to a well accepted computational, i.e. the Von Neumann model. We believe a similar approach is needed in order to establish accepted benchmarks, i.e. it is necessary to ground these tests on sound abstractions of the robot components sustaining the algorithms under evaluation.

Finally, we would like to stress that carefully engineered simulation environments can be instrumental in perfecting evaluation methodologies and to perform preliminary trials aimed to select the most appropriate mapping solution for the task at hand.

References

[1] S. M. Abdallah, D.C. Smar, and J.S. Zelek. A benchmark for outdoor vision SLAM systems. *Journal of Field Robotics*, 24(1/2):145–165, 2007.

[2] F. Amigoni. Experimental evaluation of some exploration strategies for mobile robots. In *IEEE International Conference on Robotics and Automation*, pages 2818–2823, 2008.

[3] F. Amigoni and A. Gallo. A multi-objective exploration strategy for mobile robots. In *IEEE International Conference on Robotics and Automation*, pages 3850–3855, 2005.

[4] B. Balaguer and S. Carpin. Uc mercenary team description paper: Robocup 2008 virtual robot rescue simulation league. In *Team Description Paper for RoboCup 2008*, 2008.

[5] B. Balaguer and S. Carpin. Where Am I? A Simulated GPS Sensor for Outdoor Robotic Applications. In S. Carpin et al., editors, *Proceedings of the First International Conference on Simulation, Modeling and Programming for Autonomous Robots*, pages 222–233. Springer, 2008.

[6] B. Balaguer, S. Carpin, and S. Balakirsky. Towards quantitative comparisons of robot algorithms: Experiences with slam in simulation and real world systems. In *Workshop on "Performance Evaluation and Benchmarking for Intelligent Robots and Systems" at IEEE/RSJ IROS*, 2007.

[7] S. Balakirsky, S. Carpin, A. Kleiner, M. Lewis, A. Visser, J. Wang, and V.A. Ziparo. Towards heterogeneous robot teams for disaster mitigation: Results and performance metrics from robocup rescue. *Journal of Field Robotics*, 24(11-12):943–967, 2007.

[8] J. Baltes. A benchmark suite for mobile robots. In *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems*, pages 1101–1106, 2000.

[9] R.J. Baron. Mechanisms of human facial recognition. *International Journal of man machine studies*, pages 137–178, 1981.

[10] N. Basilico and F. Amigoni. On evaluating performance of exploration strategies for autonomous mobile robots. In *Proceedings of the Performance Evaluation and Benchmarking for Intelligent Robots and Systems Workshop at IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2008.

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

[12] A. Birk, J. Poppinga, T. Stoyanov, and Y. Nevatia. Planetary Exploration in USARsim: A Case Study including Real World Data from Mars. In L. Iocchi, H. Matsubara, A. Weitzenfeld, and C. Zhou, editors, *RoboCup 2008: Robot WorldCup XII*, Lecture Notes in Artificial Intelligence (LNAI). Springer, 2008. in press.

[13] D. Calisi, L. Iocchi, and D. Nardi. A unified benchmark framework for autonomous mobile robots and vehicles motion algorithms (movema benchmarks). In *Workshop on Experimental Methodology and Benchmarking in Robotics Research (RSS 2008)*, 2008.

[14] 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.

[15] S. Carpin, M. Lewis, J. Wang, S. Balakirsky, and C. Scrapper. Bridging the gap between simulation and reality in urban search and rescue. In *Robocup 2006: Robot Soccer World Cup X*, number 4434 in LNCS, pages 1–12. Springer, 2007.

[16] 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.

[17] S. Carpin, T. Stoyanov, Y. Nevatia, M. Lewis, and J. Wang. Quantitative assessments of usarsim accuracy. In *Proceedings of the Performance Metrics for Intelligent Systems Workshop*, 2006.

[18] S. Carpin, J. Wang, M. Lewis, A. Birk, and A. Jacoff. High fidelity tools for rescue robotics: results and perspectives. In *Robocup 2005: Robot Soccer World Cup IX*, LNCS, pages 301–311, 2006.

[19] J.J. Collins and S. O'Sullivan. Developing a benchmarking framework for map building paradigms. In *Ninth International Symposium on Artificial Life and Robots*, 2004.

[20] T. Collins, J.J. Collins, M. Mansfield, and S. O'Sullivan. Evaluating techniques for resolving redundant information and specularity in occupancy grids. In S. Zhang and R. Jarvis (Eds.), editors, *Proceedings of AI 2006: Advances in Artificial Intelligence*, pages 235–244. Springer, 2005.

[21] T. Collins, J.J. Collins, and C. Ryan. Occupancy grid mapping: and empirical evaluation. In *Proceedings of the Mediterranean Conference on Control and Automation*, pages 1–6, 2007.

[22] S.J. Egerton and V. Callaghan. A benchmark for measuring mobile robot environment modelling performacne. In *Proceedings of the IEEE Conference on Robotics, Automation and Mechatronics*, pages 407–412, 2004.

[23] G. Fontana, M. Matteucci, and D.G. Sorrenti. The RAWSEED proposal for representation-independent benchmarking of SLAM. In *Workshop on Experimental Methodology and Benchmarking in Robotics Research (RSS 2008)*, 2008.

[24] E. Frontoni, A. Ascani, A. Mancini, and P. Zingaretti. Performance metric for vision based robot localization. In *Workshop on Experimental Methodology and Benchmarking in Robotics Research (RSS 2008)*, 2008.

[25] The GeoTIFF website. http://trac.osgeo.org/geotiff, 2009.

[26] G. Grisetti, C. Stachniss, and W. Burgard. Improved techniques for grid mapping iwht Rao-Blackwellized particle filters. *IEEE Transactions on Robotics*, 23(1):36–46, 2007.

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

[28] A. Jacoff, E. Messina, B. Weiss, S. Tadokoro, and Y. Nakagawa. Test arenas and performance metrics for urban search and rescue robots. In *International Conference on Intelligent Robots and Systems*, pages 3396 – 3403, 2003.

[29] B. Kuipers. Modeling spatial knowledge. *Cognitive science*, 2:129–153, 1978.

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

[31] O. Michel, Y. Bourquin, and J.C. Baillie. Robotstadium: Online humanoid robot soccer simulation competition. In *Workshop on Experimental Methodology and Benchmarking in Robotics Research (RSS 2008)*, 2008.

[32] MOAST project. http://sourceforge.net/projects/moast, 2009.

[33] H.P. Moravec. Sensor fusion in certainty grids for mobile robots. *AI Magazine*, 9(3):61–74, 1988.

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

[35] 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.

[36] S. O'Sullivan, J.J. Collins, M. Mansfield, D. Haskett, and M. Eaton. Linear feature prediction for confidence estimation of sonar readings in map building. In *Ninth International Symposium on Artificial Life and Robots*, 2004.

[37] S. O'Sullivan, J.J. Collins, M. Mansfield, D. Haskett, and M. Eaton. A quantitative evaluation of sonar models and mathematical update methods for map building with mobile robots. In *Ninth International Symposium on Artificial Life and Robots*, 2004.

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

[39] K. Sato, S. Yotsukura, and T. Takahashi. To a Rescue Simulation Applicable to Anywhere - Team Description Hinomiyagura, July 2008. Proceedings CD.

[40] T. Schmits and A. Visser. An Omnidirectional Camera Simulation for the USARSim World. In *Proceedings of the 12th RoboCup International Symposium*, July 2008. Proceedings CD. To be published in the Lecture Notes on Artificial Intelligence series.

[41] R. Sim, G. Dudek, and N. Roy. Online control policy optimization for minimizing map uncertainty during exploration. In *IEEE International Conference on Robotics and Automation*, 2004.

[42] C. Stachniss and W. Burgard. Exploring unknown environments with mobile robots using coverage maps. In *18th Joint Conference on Artificial Intelligence*, 2003.

[43] I. Varsadan, A. Birk, M. Pingsthorn, S. Schwertfeger, and K. Pathak. The jacobs map analysis toolkit. In *Workshop on Experimental Methodology and Benchmarking in Robotics Research (RSS 2008)*, 2008.

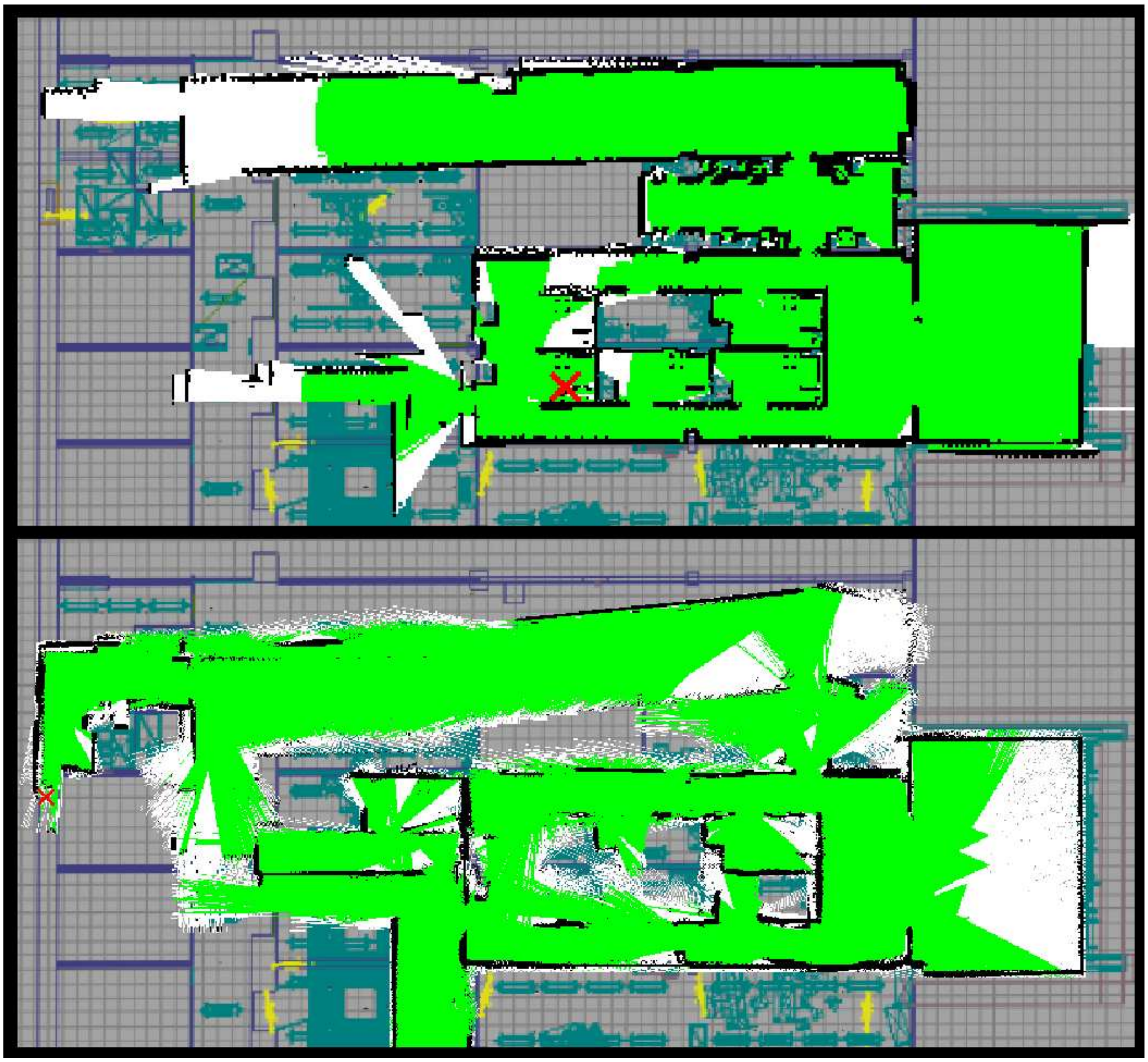

Figure 2: Examples of basis maps generated in the environment shown in Fig. 1. The top image shows the GeoTIFF submitted by Team A while the bottom image shows the GeoTIFF submitted by Team C during the finals of the 2008 RoboCup Virtual Robot Rescue competition. Behind each basis map is a layer with a raster and ground-truth data for this environment.

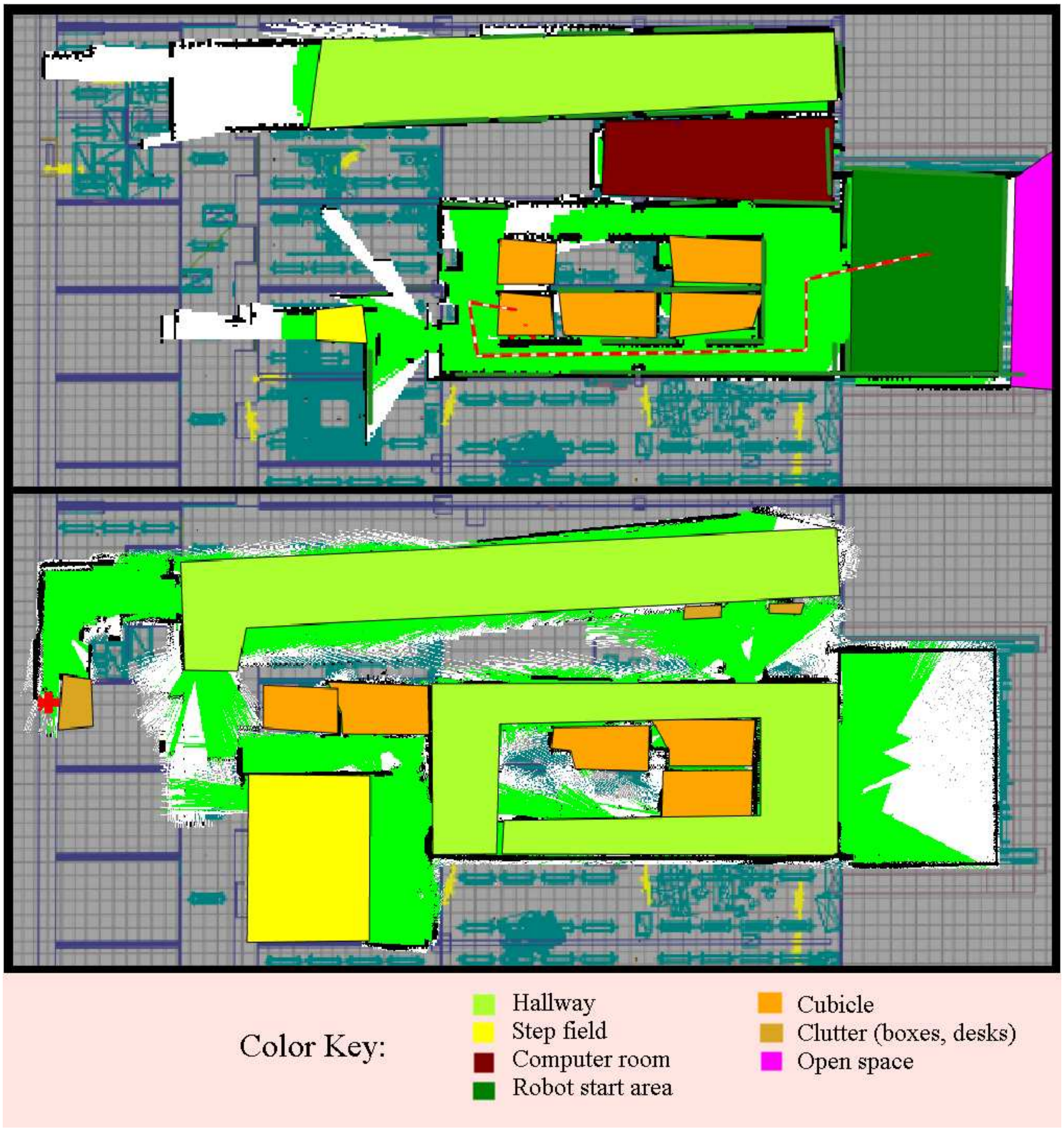

Figure 3: Map with vector overlays and ground truth of the maps submitted by Team A and Team C.

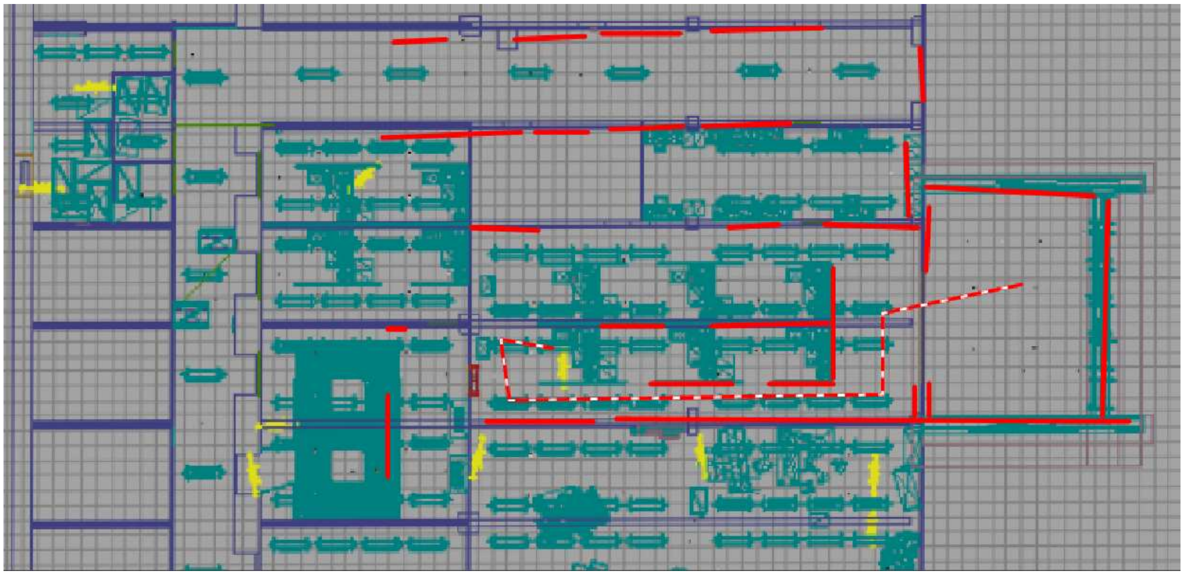

Figure 4: Example of map skeleton and ground truth. The red lines make up the skeleton.

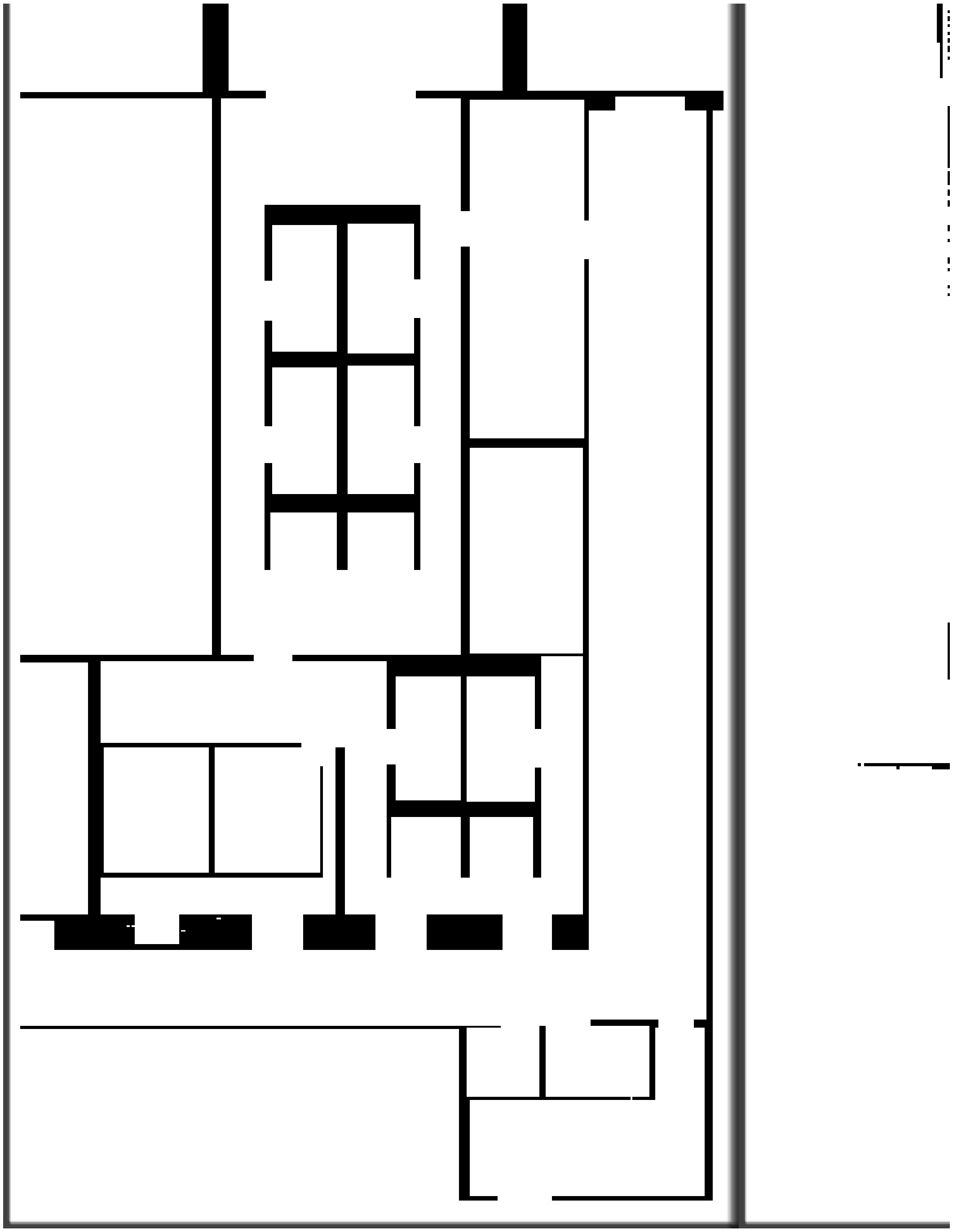

Figure 5: *Map Set 1*, comprised of maps on which the metrics comparison have been made in Table 1. From left to right, the maps represent ground truth, Team A, Team B, Team C, and Team D. Occupied cells are displayed in black and free cells are displayed in white.

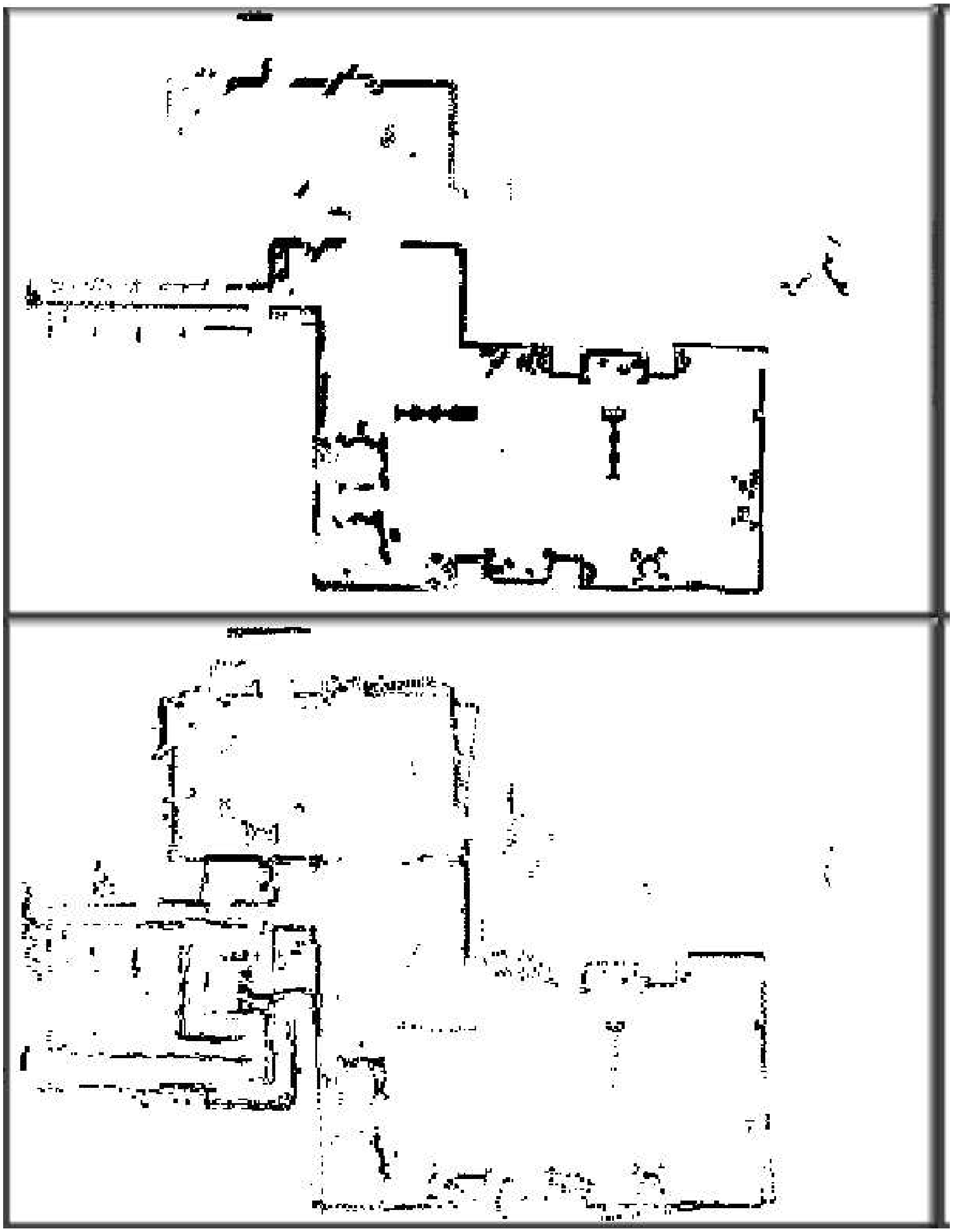

Figure 6: *Map Set 2*, comprised of maps on which the metrics comparison have been made in Table 2. The image in the first row is the ground truth. For the second row, from left to right, the maps represent Team A, Team B, and Team C. For the last row, from left to right, the maps represent Team D, Team E, and Team F. Occupied cells are displayed in black and free cells are displayed in white.

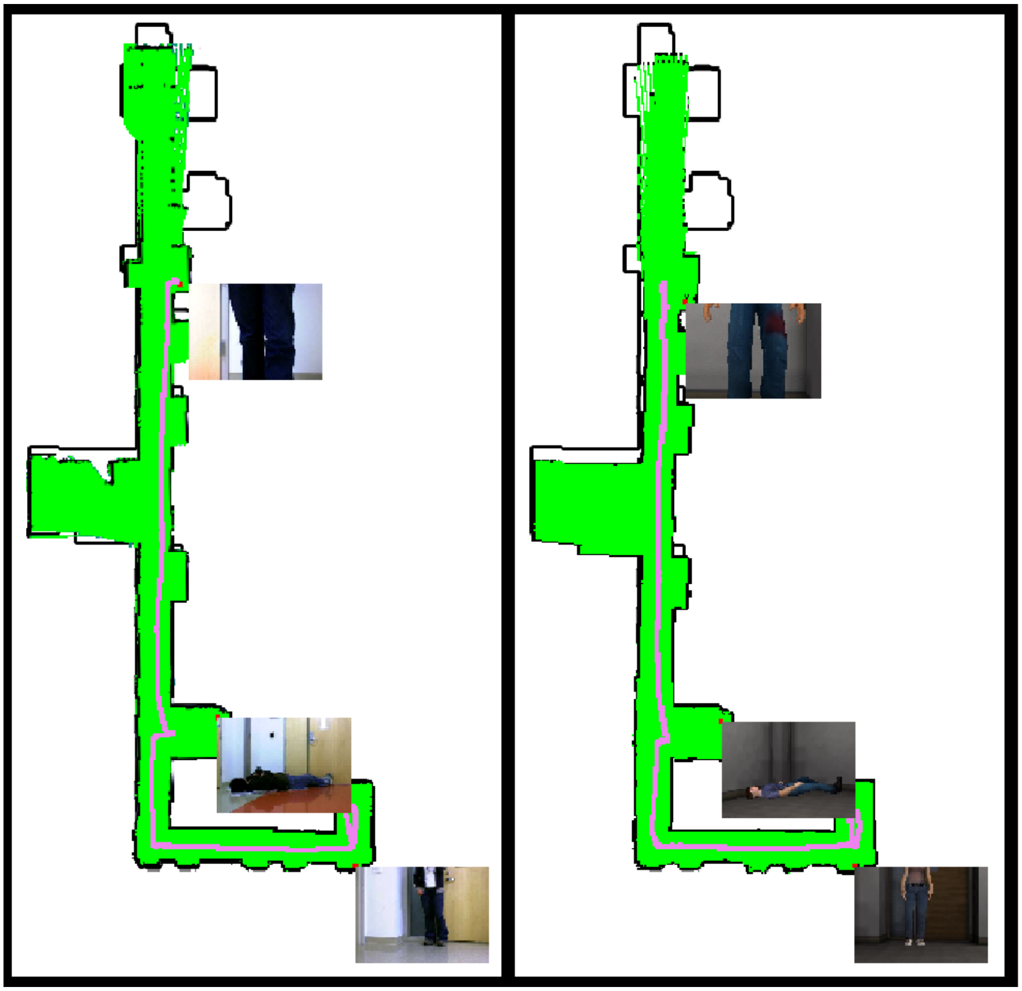

Figure 7: This figure shows the composite map generated by the real platform (left) and the simulated platform (right). The attribute overlay is not displayed.

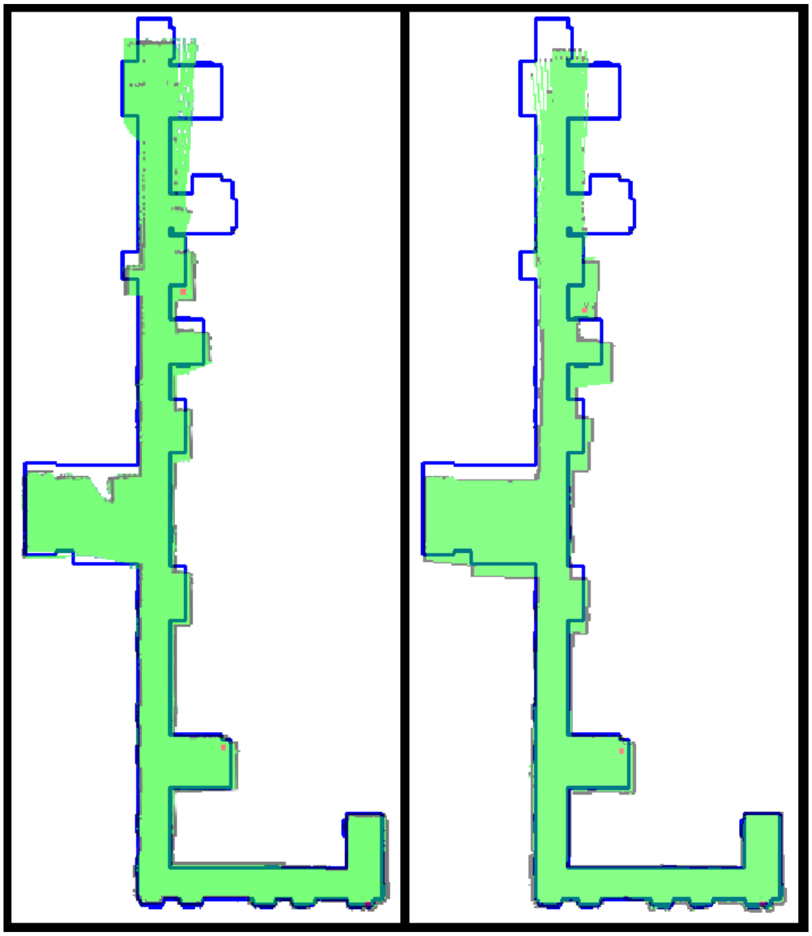

Figure 8: This figure shows the raster map generated by the real platform (left) and the simulated platform (right), each of which is superimposed by the ground truth map (blue). These layers are used to compare the metric quality of the solutions.

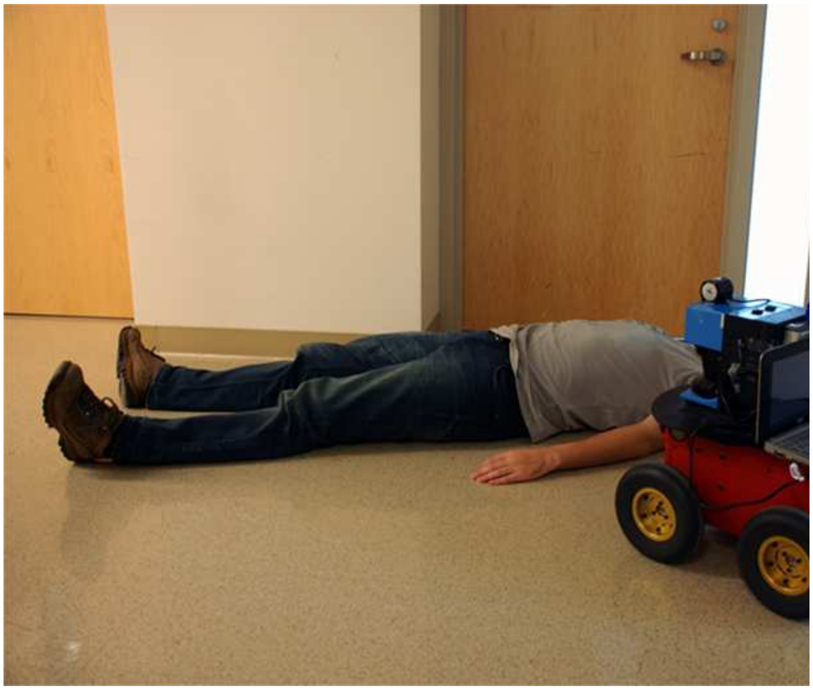

Figure 9: The real robot operating in the environment (left), along with its simulated counterpart (right). The figure shows the fidelity of the robot, environment and victim models inside the simulation.