Evaluating The RoboCup 2009
Virtual Robot Rescue Competition

Stephen Balakirsky
NIST
100 Bureau Drive
Gaithersburg, MD, USA
+1 (301) 975-4791

stephen@nist.gov

Stefano Carpin
University of California, Merced
5200 N Lake Rd
Merced, CA, USA
+1 (209) 228-4152

scarpin@ucmerced.edu

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

A.Visser@uva.nll

ABSTRACT

The 2009 RoboCup Competitions took place in Graz Austria in July of 2009. The Virtual Robot Rescue Competition included 11 competitors from 10 different countries. The main objective of this competition is to utilize teams of robots to perform an urban search and rescue (USAR) mission over both indoor and outdoor terrains. For the first time, elemental tests were performed in autonomously generated map quality, multi-vehicle tele-operation, and communication's system deployment. In addition, a comprehensive USAR scenario was performed. This year's competition featured new performance metrics and automatic scoring programs. This paper presents an overview of the metrics for the competitions and lessons learned from their application during a high-intensity international competition.

Categories and Subject Descriptors

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

General Terms

Measurement, Performance, Experimentation, Human Factors, Standardization.

Keywords

Robotics, Evaluation, Competition, Simulation, Performance Metrics, RoboCup

1. INTRODUCTION

July 2009 saw the fourth annual running of the RoboCup Rescue Virtual Robot Competition in Graz, Austria. The RoboCup competition [1] provided an international forum where approximately 400 teams, with 2000 participants from 35 countries, came together to compete in the areas of robot soccer, rescue, service robotics, and junior leagues. For the Virtual Robot Competition (VRC), 11 teams from 10 countries (Austria, Brazil, China, England, Germany, Iran, Italy, the Netherlands, Spain and the USA) participated.

Figure 1: Example of the bridge accident scene from the outdoor environment used in the RoboCup07 competition.

In the past, the VRC was run as several rounds of Urban Search and Rescue (USAR) scenarios. Each scenario consisted of teams of robots striving to find as many victims as possible in an indoor or outdoor accident scene such as the one depicted in Figure 1. The scoring performance metrics were specifically designed to award research advances in the general areas of multi-agent cooperation, human-computer interfaces (HCI), and map building. Specific emphasis was placed on the formation of multi-agent communication networks, complex terrain navigation, and victim search and identification strategies. While certain aspects of the scoring were computed automatically, a significant part of the scoring metric was computed by hand by the technical committee of the competition. This scoring procedure was very time consuming and placed a large burden on the committee, thus limiting the number of teams that would be able to participate in the event. More information on the scoring metrics utilized in past competitions may be found in Balakirsky et al. [2].

While these performance metrics proved useful in determining the overall winner of the competition, it was not possible to get deep insight into why a team won. The individual components that constitute a team's capabilities were not evaluated; only the composite results. Since strength in several different areas is required to successfully carry out the mission, it may be stated that the team with the strongest weakest link would win the competition. The goal of this year's event was to change that. We wished to be able to determine which team had the strongest mapping, which had the best communications strategy, which had

the best human-computer interface, AND which had the best overall system. To this end, we implemented the SCORE framework of evaluation [3] and evaluated three elemental tests as well as the overall USAR scenario. In addition, an effort was made to automate as much of the scoring procedure as possible. This allowed our three person technical committee to simultaneously work with four teams (two teams competing in an actual event and two teams setting up to compete in the next event).

The remainder of this paper is organized as follows: Section 2 provides an overview of the elemental tests and the final scenario. Section 3 describes the scoring metrics and the automated scoring tools that were utilized during the competition. Section 4 presents a summary of the results and lessons learned about the scoring metrics and Section 5 presents conclusions and future work.

2. COMPETITION OVERVIEW

The competition consisted of three preliminary events that were designed to test the individual team's capabilities in elemental skills followed by four comprehensive events that presented an overall search and rescue scenario. The points gained from the preliminary events (0-50 for each elemental test) were totaled to determine who would proceed to the semi-final round. The semi-final round presented two disaster scenarios (one indoor and one outdoor) to the remaining teams. Points were again summed to determine who would proceed to the final round. The final round was run in the same manner as the semi-finals. All of the environments used in the competition have been released to the public via our sourceforge website.[1]

An enhancement in the 2009 worlds included the use of elements that were directly borrowed from the RoboCup Rescue Physical Robot League. This league features real robots competing in physical arenas to provide maps of the environment and locate victims. For this year's competition the physical league's maze area was virtually constructed and replicated in order to fill the upper right room of the mapping challenge world (see Figure 2). Additional elements, such as step fields, appeared in all of the virtual competition worlds.
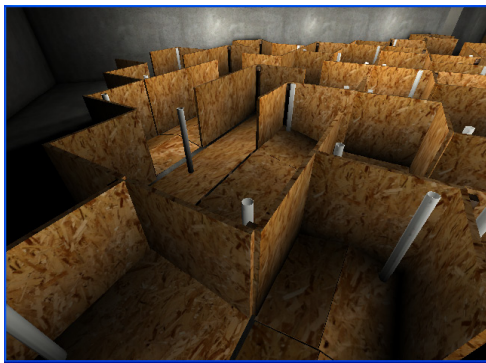

Figure 2: View on the physical league's maze inside the mapping challenge environment.

2.1 Mapping Challenge

The first elemental skill test consisted of a mixed-autonomy (both tele-operation and autonomous operation) mapping challenge. The idea behind this event is that robots are given 20 minutes to map out the environment before the emergency responders enter the building. Once they enter, the emergency responders need to know the best routes to take to various newly discovered points of interest. No *a priori* data was provided to the teams for this event. Teams were allowed to use up to four robots to explore this indoor environment. During this exploration, the robots must communicate with each other through the use of a communications simulator and with the outside world through a communications station. The simulated sensors included a realistic noise model for both external (laser range finder and sonar) and internal (wheel encoders and IMU) sensors. The world was designed to stress the robot's sensors and algorithms by featuring both flat floored and sloped floor mazes, large featureless spaces, and various lighting conditions. The overall size of this world was 45m x 55m. No team was able to explore the entire environment with 4 robots during the 20 minute period.
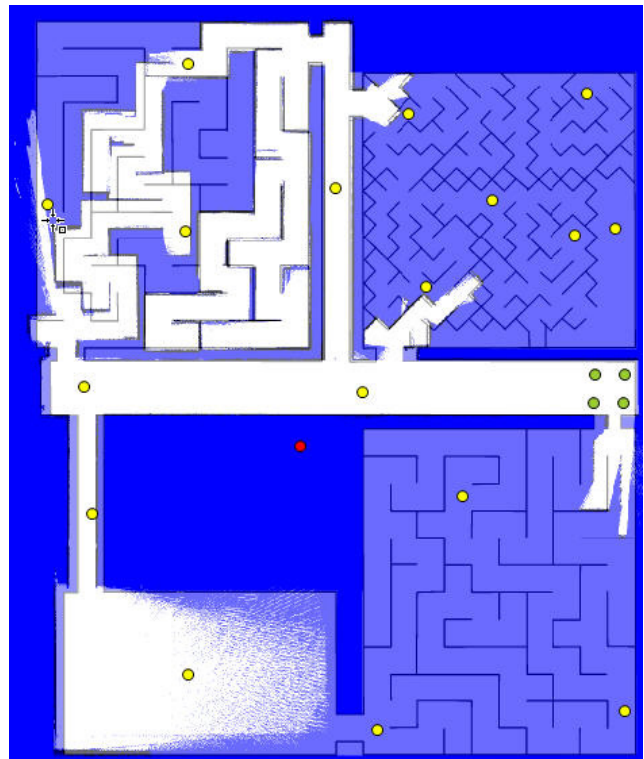

Figure 3: Team 1's mapping challenge map.

A sample of "Team 1's" map is shown in Figure 3. The green circles in the figure represent the starting locations of the four robots. Yellow circles are utilized for scoring and will be discussed in 3.1. The team's map (white) has been overlaid on the ground truth for the event. The upper-left and lower-right rooms consisted of flat floored mazes. The upper right room was an enlarged model of one of the physical robot's mapping mazes. The room on the lower-left was a large featureless and dark space. The robots were able to maintain a good connection with the

communication's station (shown as the red dot) from anywhere in the world. Teams from this event also participated in the interleague challenge which featured real data collected from the physical leagues maze. For the interleague challenge, the teams ran their identical code from the simulation events on real data.

2.2 Deployment Challenge

The second elemental skill test consisted of a deployment challenge. Teams were allowed to use combinations of up to eight robots in this event. For this event, a radio propagation model was utilized to determine if robots were able to communicate with each other. This model computes radio signal attenuation based on a combination of the distance between the two robots and the number of objects (walls and obstacles) that the signal needs to pass through. A uniform signal loss was applied for each object penetration.

The idea behind this event was that emergency responders needed to enter and work in a building that was too large to have continuous communications coverage without repeaters. Each of the team's robots had a repeater mounted on it, and the teams were required to establish a communications network that covered as much of the building's interior as possible. Scoring for this event was based on the number of square meters of the building that had network connectivity with a communications base station. This measure was automatically generated.

Figure 4: Section of Deployment map from Team 2.

Approaches to accomplish this task ranged from pre-planning the locations for the robots based on an estimate of the communication strength, to driving a robot until it reached the end of its communication range and then extending this range by driving a new robot into the frontier. No operator involvement was allowed during this test, so all robots had to navigate autonomously through the environment.

Figure 4 shows Team 2's map for this challenge. Team 2 pre-computed expected coverage and then autonomously navigated their robots to the computed locations. The blue dots in the figure represent the robot's starting locations and the white dots represent the robot's final locations. The green dot is where the communication station was located. The red areas are obstacles or outside of the building structures, while the green area represents the radio coverage. Teams were provided with *a priori* data for this event, but the data had several errors with some rooms being blocked by collapses. Some teams miscomputed the radio coverage and drove their robots too far thus disconnecting them from the overall network. Team 2 was one of those teams. As may be seen in Figure 4, the two top robots are not connected to the overall network (i.e. there is no path from them to the communication's station) and thus did not generate any points for the team. In addition, not all of the robots were able to reach their desired ending locations. This was due to blockades and navigational challenges present in the terrain, which the robots had to overcome autonomously.

2.3 Tele-Operation Challenge

The final elemental skill test performed was a tele-operation challenge. For this challenge, teams were permitted to use 8 robots in order to reach 8 predetermined goal points. Each goal point was selected such that a particular class of robot was best suited to reach it. For example, there were elevated goal points, as well as goal points in small spaces. The teams did not have *a priori* data on which robot should attempt which goal location.

The idea behind this challenge is that emergency responders have knowledge of interesting locations that must be remotely examined. The teams need to reach these locations and provide feedback. In addition, past experience has shown that few teams were experimenting with novel robots and teaming arrangements. Therefore, another idea behind the tele-operation challenge was to introduce teams to a number of different platforms and to stress their human computer interfaces. Teams could use up to 8 robots, but could only use 2 of each robot class. This rule was put into effect in order to encourage teams to experiment with multiple types of robots and to form heterogeneous teaming arrangements.

An additional challenge for the teams was the lighting condition. All ground robots had to navigate through a small maze before they could reach the target points. In that maze it was quite dark, forcing teams which rely on visual feedback to fall back to the other sensors present on the robots.
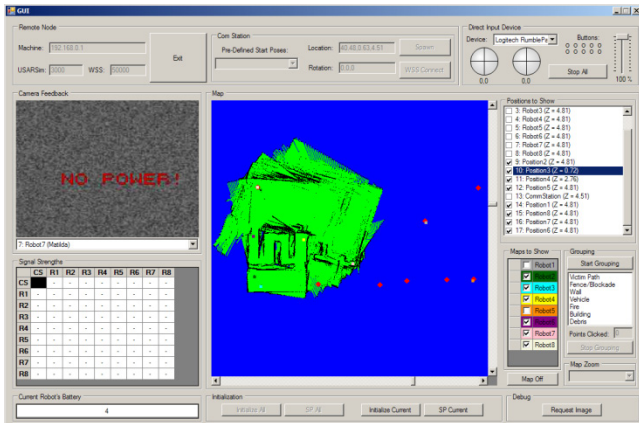
Figure 5: Example of GUI from the Tele-operation challenge.

This challenge was automatically scored by a metric that evaluated the number of goals reached and the distance from the goal that the robot was able to achieve. A screen shot from a sample GUI is shown in Figure 5. This figure shows the map that the robots are generating as well as unreached goal locations (the red dots).

2.4 Semi-Finals

The semi-finals took the competition back to its roots of performing a tele-assisted multi-robot rescue mission. The top 5 teams ran through both an indoor and outdoor disaster environment. The environments were challenging with large areas of uneven terrain. The semi-finals incorporated a complex scoring metric that included automatic and hand-generated scores. The metric included the amount of area cleared by the robots (guaranteed to be victim free), victim scoring that included the number of victims found and various attributes of the victims (location of injuries, physical description, …), and map quality points.

Figure 6: Example of the outdoor disaster environment.

The original idea behind the semi-finals was to award the 3rd place prize and have only the top two teams precede to the finals. This idea was abandoned, because after the semi-finals there was only a one point difference between 3rd and 4th places Therefore, a decision was made to have a 3rd place runoff the next day.
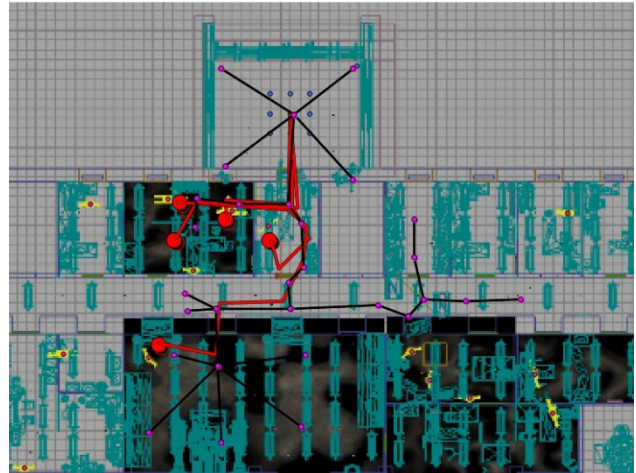

Figure 7: Vector components from final's indoor map.

2.5 Finals and Run-off

Another set of worlds was utilized for the finals and run-offs. These were the most complex worlds that have ever used since the start of the competition in 2006. The top two teams were able to provide both pixel maps (in our standard color scheme) and MapInfo Data Interchange Format (MIF) formatted vector skeletons. One such skeleton (with black arcs and purple nodes) is shown in Figure 7. This was a new feature of this year's competition and shows one way that we are encouraging innovation amongst the teams. Figure 7 also displays other innovative vector information such as responder paths to located victims (red lines) and detailed victim information (each red dot is tagged with information on victim location and vital signs). Raster map components included geo-referenced victim pictures that can be displayed on the map as well as a raster obstacle map.

3. AUTOMATIC SCORING

As previously stated, it was desired to provide automatic scoring programs whenever possible. During this competition, several new scoring techniques were examined with various degrees of success. These techniques are evaluated in the following sections.

3.1 Mapping Challenge

The evaluation of autonomously generated maps is an open question in the current literature. Many approaches tend to treat the map as an image and apply various image processing techniques to the map in order to judge its quality. One such technique is presented by Varsadan et al. in [5], where an image similarity metric is used to compare robot-produced maps against their ground truth equivalents. Past VRC events utilized a combined metric that had such a comparison at its heart. More information on previous year's map evaluation metrics may be

found in [5]. A problem with such a metric is that it is likely to negatively affect maps with a single misalignment that propagates through the rest of the map. Collins et al. [6] augment a purely image based evaluation approach by adding a measure that assess if a path generated on the robot-generated map would be valid on the ground truth map. This approach is useful if another robot will be utilizing the autonomously generated map as ground truth for planning its own routes.

It is said in [8] that any map assessment method should be intimately tied to the practical task for which the map will eventually be used. In the case of the VRC, this would be for a emergency responder to utilize the map to find a path to a point of interest. These points would be specified in terms of features (e.g. fourth office on the left) instead of geodetic coordinates. It was hoped that this evaluation technique would reduce the problem of misalignments propagating through the map and distorting the map's score. The actual approach that was implemented was to choose several random destinations from the environment as points of interest (POI). Paths to these POIs would then be computed on the team's maps and evaluated for their topological correctness.

The problem with this approach is that the team's maps are delivered as grid-based images and not as topological structures. Therefore, a technique for extracting the topological information needed to be developed. To solve this task, the POIs were mapped to the team's map and a standard path generating algorithm [9] was run to compute a path solution. The topological properties of this path were then evaluated and used to determine the map's score.

Unfortunately, this automatic procedure was not finished in time for the competition and topological paths were generated by hand on the competitors map. While not the ideal solution of having an automated scoring tool, this procedure was easy and quick to implement and provided valuable insight to the value of this scoring metric.

3.2 Deployment Challenge
The deployment challenge saw the first application of a fully automatic scoring technique applied to the competition. The Java scoring application may be found at sourceforge[2].

The program faithfully replicates the equations used by the Wireless Communication Server (WSS) server in order to determine the signal attenuation between a transmitter and a receiver located at arbitrary positions in the world. The attenuation considers both degradation due to increasing distance, and the presence of obstacles. The overall score is computed as follows. Once the challenge is over, final positions of all robots are retrieved from the log files. These log files are automatically generated by the simulation system and contain full ground truth of the simulation run. The position of the communication base station, instead, is fixed and known to all participants. A *connectivity graph* is then created. The connectivity graph is a graph whose vertices are the robots and the communication base

station. An edge is added between two vertices if the corresponding elements can communicate with each other. In order to determine if two elements (robots or base station) can exchange data, the formerly mentioned equations are used to compute the signal attenuation between the two. If the signal strength is above a given threshold (-93 dB during the competition) an edge is added between the vertices. Once the graph is available, a breadth-first graph search is computed having the base station as the source vertex. All vertices reachable from the source are considered in communication with the base station. The rationale is that they can send information to the base station either directly or indirectly. Once the set of robots connected to the base station is known, the whole environment is sampled on a grid with a given resolution (specified by the user). A point in the grid is considered in communication range if it is connected to one or more robots reachable (directly or indirectly) from the base station. The overall score is the number of sampled points that are connected.

3.3 Tele-operation Challenge
The tele-operation challenge was also scored automatically by a program located at sourceforge[3] using the automatically generated log files. The goal of the challenge was to bring at least one robot to each target location in the allocated time.

A target location was considered reached if at least one robot was within T meters of location where T was a constant determined by the judges and known to the teams before the competition. For the 2009 event, this constant was set to 2 m. 50 points were awarded for each point reached. In addition, the program automatically summed the distance of the target location from the deployment site. This distance was an indication of the difficulty to reach this point. To be able to rank teams which reached exactly the same targets, the distance of the robots to those target locations was subtracted from the score.

The scores of all three challenges were normalized relative to the score of the best team, which gives each challenge an equal weight.

3.4 Semi-finals and Finals
The scoring of the full scenarios involved a combination of automatic scoring and hand scoring. Points for world exploration and victims were computed with automatic programs[4] while map quality assessments were performed by hand. For world exploration, a team's map was first trimmed to remove any out-of-bounds areas or poorly covered areas that were claimed to be explored. An example of a poorly covered area may be seen in the upper right corner of the lower left room of Figure 3. Here a team presents stripes of explored area mixed with stripes of unknown areas. An automatic program was run on the resulting image that computed the area of map that was explored (seen by the robot) and the area that was cleared (guaranteed to be victim free). Exploration points were normalized to a maximum of 50 points.

Victim points included points for correct victim localization and attribution, and subtractions for incorrect localizations and victims

[2] http://usarsim.cvs.sourceforge.net/viewvc/usarsim/usarsim/Tools/ScoreRadio/

[3] http://usarsim.cvs.sourceforge.net/viewvc/usarsim/usarsim/Tools/ScoreTeleOp/

[4] http://usarsim.cvs.sourceforge.net/viewvc/usarsim/usarsim/Tools/ScoreVictims/

that resided in "cleared" areas and were not detected. Victim localization was computed automatically while attribution needed to be hand computed. Victim points were again normalized to a maximum of 50 points.

The final area of scoring was in the computation of skeleton quality, metric quality, and attribution of team provided maps. These scores were computed entirely by hand following the procedure outlined in [2].
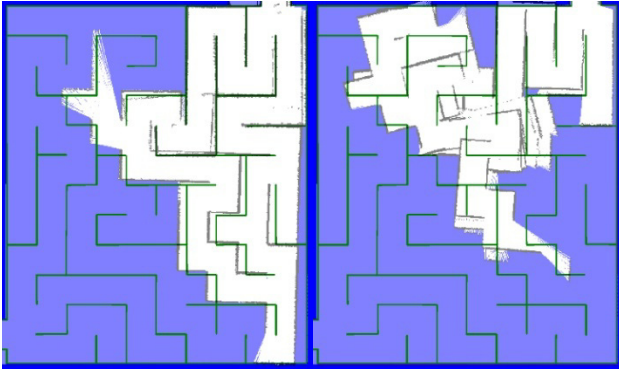

Figure 8: Image of two different team's maps for the same area of the mapping challenge world.

4. SUMMARY OF RESULTS

Overall, we were very happy with the results of the automatic scoring. However, issues did arise with the proposed automatic scoring of the mapping challenge. The intent of the scoring metric was to select several pseudorandom points in the environment and to then compute routes to these points. However, mapping errors made is difficult to place these points on some of the competitor's maps (even by hand). For example, Figure 8 shows the maps from two of the teams. While the map on the left shows slight misalignments, the map on the right presents several rotational errors and scan mis-matches that have caused extra walls to be added.

The problem becomes one of determining where to place our pseudorandom points in the right hand map. This determination must be made before any topological map calculations may be made. One possible solution for this problem is to allow the teams to know the locations of the points before the run. The teams will then need to mark the point locations in their maps and routes will then be generated from the starting location to their marked points. The topological properties of these routes may then be judged against routes created on the ground truth map.

5. FUTURE WORK

While the current automation performed well, there are still several areas that need automating. Techniques need to be developed (or the metrics modified) that will allow for the automatic generation of scores for the semi-final and final rounds. In addition, the mapping challenge scoring program needs to be created and validated.

6. REFERENCE LIST

[1] H. Kitano, RoboCup-97: Robot Soccer World Cup I, Berlin: Springer-Verlag, 1998.

[2] S. Balakirsky, C. Scrapper, and S. Carpin, "The Evolution of Performance Metrics in the RoboCup Rescue Virtual Robot Competition." Proceedings of the Performance Metrics for Intelligent Systems (PerMIS) Workshop, . Aug. 2007.

[3] C. Schlenoff, M. Steves, B. Weiss et al., "Applying SCORE to Field-based Performance Evaluations of Soldier Worn Sensor Technologies," Journal of Field Robotics, vol. 24, no. 8-9. pp.671-698, 2006.

[4] I. Varsadan, A. Birk, and M. Pfingsthorn, "Determining Map Quality Through An Image Similarity metric." In *RoboCup 2008: Robot Soccer World Cup XII*. Springer Berlin/Heidelberg. Lecture Notes in Artificial Intelligence, volume 5339, pp. 355-365. 2009.

[5] B. Balaguer, S. Balakirsky, S. Carpin et al., "Evaluating maps Produced by Urban Search and Rescue Robots: Lesssons Learned from RoboCup," *Autonomous Robots 27*, 2009.

[6] T. Collins, J. Collins, and C. Ryan, "Occupancy Grid Mapping and Empirical Evaluation." , Proceedings of the Mediterranean Conference on Control and Automation, pp. 1-6, 2007.

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

[8] J.C. Latombe, "Robot Motion Planning", Kluwer Academic Publishers, Boston, MA, 1991.