The Evolution of Performance Metrics in the RoboCup Rescue Virtual Robot Competition

Stephen Balakirsky & Chris Scrapper
NIST
100 Bureau Drive
Gaithersburg, MD, USA

Stefano Carpin
University of Calif. Merced
5200 North Lake Rd
Merced, CA, USA

***Abstract*—**This paper presents an overview of the 2007 RoboCup Rescue Virtual Robot Competition and the performance metrics that were used to judge the competition. For this competition, great effort was placed in bringing together researchers with diverse interests to competitively participate. The competition arenas and metrics used for scoring were specifically designed to create a "level" playing field for the various research disciplines. The specific metrics, how they evolved from the prior year's competition, and the way in which the competition was run will be discussed in detail. Defects that were noted in the metrics will also be discussed.

***Keywords*:** *robotics, competition, simulation, performance metrics, RoboCup*

I. Introduction

July 2007 saw the second annual running of the RoboCup Rescue Virtual Robot Competition in Atlanta GA. Robocup [1] provides an international forum where researchers meet to compete against each other in robotic competitions ranging from soccer to dance to urban search and rescue (USAR). Underlying these competitions are basic research thrusts focusing on core robotic technologies such as mobility, multi-agent cooperation, and fine motor control.

This year's USAR virtual robot competition consisted of 9 runs over 7 days and took place in complex indoor and outdoor domains. 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. The use of *a priori* data and carefully constructed worlds allowed the researchers to concentrate their efforts in one or more research areas while maintaining competitiveness among groups performing in different research areas. Examples of the indoor and outdoor worlds that were used for the competition are shown in Figure 1 and Figure 2 respectively.

Figure 1: Example of the cubicle area from the indoor environment used in the RoboCup 07 competition.

Figure 2: Example of the bridge accident scene from the outdoor environment used in the RoboCup07 competition.

II. Background

A. USARSim

The current version of Urban Search and Rescue Simulation (USARSim)[3] is based on the UnrealEngine2[1] game engine that was released by Epic Games as part of

[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 National Institute of Standards and Technology, nor does it imply that the software tools identified are necessarily the best available for the purpose.

Unreal Tournament 2004. This engine may be inexpensively obtained by purchasing the Unreal Tournament 2004 game. The USARSim extensions may then be freely downloaded from sourceforge.net/projects/usarsim. The engine handles most of the basic mechanics of simulation and includes modules for handling input, output (3D rendering, 2D drawing, and sound), networking, physics and dynamics. USARSim uses these features to provide controllable camera views and the ability to control multiple robots. In addition to the simulation, a sophisticated graphical development environment and a variety of specialized tools are provided with the purchase of Unreal Tournament.

The USARSim framework builds on this game engine and consists of:

- standards that dictate how agent/game engine interaction is to occur,
- modifications to the game engine that permit this interaction,
- an Application Programmer's Interface (API) that defines how to utilize these modifications to control an embodied agent in the environment ,
- 3-D immersive test environments,
- models of several commercial and laboratory robots and effectors,
- models of commonly used robotic sensors

USARSim does not provide a robot controller. However, several open source controllers may be freely downloaded. These include the community-developed MOAST controller (sourceforge.net/projects/moast), the player middleware (sourceforge.net/projects/playerstage), and any of the winning controllers from previous year's competitions (2006's winning controllers may be found on the Robocup Rescue wiki at: www.robocuprescue.org/wiki/). A description of the winning algorithms may be found in [2].

B. RoboCup Virtual Robot Competition

RoboCup is an annual competition that was held in 2007 in Atlanta, GA. Nearly 300 teams from 33 countries participated. The virtual robot competition (VRC) is part of the RoboCup Rescue Simulation League. The VRC, which this year saw its second annual running, is designed to foster collaboration and competition between research groups conducting research in the diverse areas of human-computer interfaces, map building, the formation of multi-agent communication networks, complex terrain navigation, and victim search and identification strategies. The competition was run over 7 days and consisted of two preliminary pass/fail rounds followed by three main competition rounds, 2 semi-final rounds, and 2 final rounds.

The preliminary rounds of the competition were designed to verify that teams met a minimum set of competencies. Teams needed to control their robots through the use of a provided communications server (a new requirement for this year in order to mimic the non-line-of-sight nature of a real disaster location), generate maps and find victims, and provide the judges with maps and victim locations that were in the proper format. Eight teams from five different countries participated in the preliminary rounds. All of the teams passed and moved onto the actual competition.

The competition rounds consisted of extensive indoor or outdoor terrain. The goal of the competition was to find as many victims while clearing as much area as possible before the batteries of the robot expired. Robots were given a battery life of approximately 20 minutes. In order to support a wide variety of research interests and lower the competition entry barriers by assuring that teams did not need to be experts in all fields, *a priori* data was provided on the difficulty of terrain traversal, the difficulty of communicating with the base station, and the difficulty of finding victims. Each of these areas had three levels of difficulty as defined as:

Mobility:
- Easy – Flat floors
- Moderate – Sloped floors, rolling areas, narrow passageways, small steps
- Difficult – Stairs, rough terrain, drops and holes that can damage the robots

Communications:
- Easy – Use of communications server required
- Moderate – No direct communication between robot and base station
- Difficult – No direct communication to base station and robot is prevented from reentering moderate or easy communication area

Victim Finding:
- Easy – Static, exposed victims with minimum false alarms
- Moderate – Dynamic (moving limbs), partially occluded, many false alarms
- Difficult – Dynamic, significant victim occlusion (entombed or hidden), many false alarms

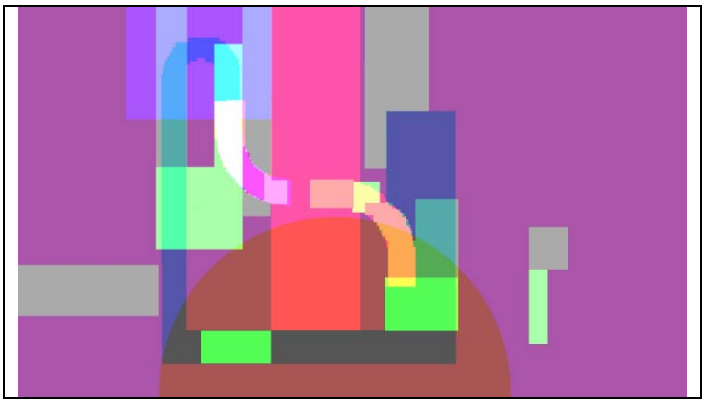

Figure 3: Example of a priori data from competition.

To balance the possible points that were awarded, an approximately even number of points were available in each area. For example, if we form the tuple (mobility difficulty, communications difficulty, victim difficulty), then the area covered by (Moderate, Easy, Easy) would have the same

points available as the area (Easy, Moderate, Easy), and (Difficult, Easy, Easy). This allowed for teams with higher levels of competency or multiple competencies to score more points.

An example of the composite *a priori* data is shown in Figure 3 where mobility information is encoded in red, victim information is encoded green, and communication information is encoded in blue. The color ranges were from 0-255 with 85 being easy, 170 being moderate and 255 being difficult. In the figure, larger values for colors appear more saturated.

III. Performance Metric

The primary goals of the competition are to report the location of victims in the environment and to form accurate, attributed maps of the explored area. These two distinct areas have separate techniques that are used for judging competency, and the performance metrics utilized have evolved.

A. Victim finding

Since one of the primary goals of the competition is to locate victims (and in 2006 to determine the victim's health status), a technique for determining a team's competency needed to be developed. However, what does it mean to "locate" a victim? How does one autonomously obtain health status? Several possible interpretations exist ranging from simply requiring a robot to be in proximity of a victim (e.g. drive by the victim) to requiring the robot to employ sensor processing to recognize that a victim is located nearby (e.g. recognize a human form in a camera image) and then examine that victim for visually apparent injuries. While recognizing a human from a camera image is the solution most readily portable to a real hardware, it places an undue burden on both the competitors and the evaluation team. For the competitors, a robust image processing system would need to be developed that could recognize occluded human forms. No matter how exceptional the mapping and exploration features of a team were, failing to produce the image processing module would result in a losing effort. In addition, the evaluation team would need to develop an entire family of simulated human forms so that teams could not "cheat" by simply template matching on a small non-diverse set of victims.

It was decided that robots should be required to be "aware" of the presence of a victim, but that requiring every team to have expertise in image processing was against the philosophy of lowering entry barriers. Therefore, a new type of sensor: a victim sensor, was introduced. To allow for the metrics to be portable to real hardware, this new sensor would need to be based on existing technology.

For the 2006 competition, the victim sensor was based on Radio Frequency Identification Tag (RFID) technology. False alarm tags were scattered strategically in the environment, and each victim contained an embedded tag. At long range (10 m), a signal from the tag was readable when the tag was in the field of view (FOV) of the sensor. At closer range (6 m), the sensor would report that a victim or false alarm was present. At even closer range (5 m) the ID of the victim would be reported. Finally, at the closest range (2 m), the status of the victim (e.g. injured, conscious, bleeding, etc.) was available. Points were subtracted for reporting false alarms, and were awarded for various degrees of information collected from the victims. Bonus points were awarded for including an image of the victim with the report. This technique worked well for scoring the 2006 competition. However, several deficiencies were noted with this sensor system:

- The RFID tag was located in the victim's torso and operated on a line-of-sight basis. Therefore, it was impossible to have largely occluded victims.
- The operation of the sensor encouraged teams to drive quickly through the environment and did not require any user input or additional behaviors when a victim was located.
- While the sensor was based on existing technology (RFID tags), no actual victim locating system works in this way.

To rectify these problems, the victim sensor was significantly revamped for the 2007 competition. The new sensor is modeled after template based human form detection. The sensor performs a line-of-sight calculation to the victim and reports which of the 7 body parts identified in Figure 4 are visible. In the right side of the figure, the points represent the possible sensor hit-points. Yellow points are non-victim, and green points represent victim hits. The worlds also contained false alarms that would be consistent with a template matching algorithm.

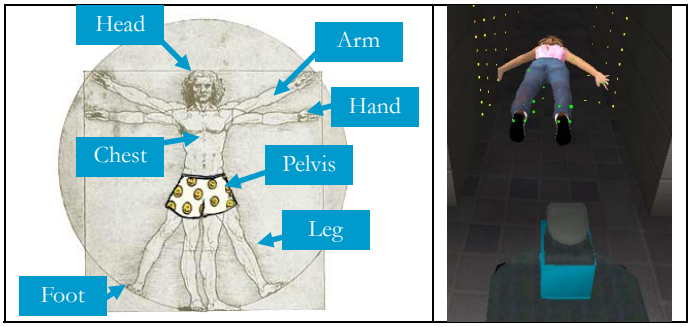

Figure 4: New victim sensor based on template matching of body parts.

The new sensor configuration required teams to attempt to gather multiple body parts from a victim (or have user involvement) in order to make a victim/false alarm determination. This usually required teams to pause upon finding a victim location in order to either alert an operator or to conduct a scan in an effort to find more body parts.

B. Map building

While knowing that a victim is located inside of a structure is useful, having a map of where this victim is located adds even more utility. Therefore, building a map of the environment is a basic requirement of the competition and

performance metrics were developed to evaluate the maps. A major change between the 2006 and 2007 competitions was the requirement that all maps be delivered as geo-registered images with specific color mappings or vector files. This allowed the judges to directly compare competitor's maps to ground truth using geographic information services (GIS) software. The map quality score is based on several components; most of which have evolved from 2006 to 2007.

- Feature quality – In 2006, there was no technique available to overlay team generated maps with ground truth. Therefore, the feature quality of a map was scored automatically by examining the reported locations of "scoring tags". Scoring tags were single shot RFID tags (they could only be read once). A requirement of the competition was for the teams to report the global coordinates of these tags at the conclusion of each run. The automatic scoring program then analyzed the deviation of the perceived locations from the actual locations. The use of these tags had the undesirable result that errors occurring early in a run were penalized more than late errors (the error affected the locations of a greater number of tags). In 2007, feature quality was evaluated subjectively. As shown in Figure 5, geo-registered maps were overlaid on ground truth and were examined for the number of discrete errors. For example, on some maps it was obvious that a single error led to a piece of the map being rotated. False obstacle reports (a single wall being reported in multiple locations) and scaling issues were also noted. The maps were ranked from best to worst and then assigned points based on their ranking.

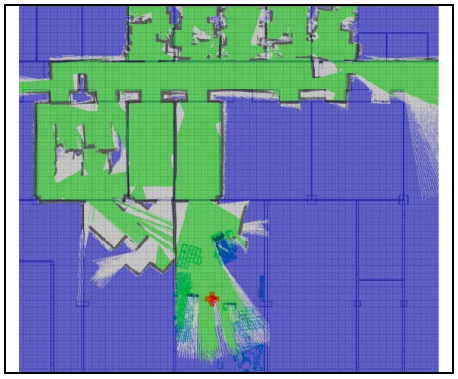

Figure 5: Competitors map overlaid on ground truth from an indoor scenario.

- Multi-vehicle fusion – Teams were only permitted to turn in a single map file. Those teams that included the output from multiple robots in that single map were awarded bonus points. This metric did not change between 2006 and 2007.
- Attribution – One of the reasons to generate a map is to convey information. This information is often represented as attributes on the map. In 2006, points were awarded for including information on the location, name, and status of victims, the location of obstacles, the paths that the individual robots took, and the location of RFID scoring tags. For 2007, 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 definition of cleared meant 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. Teams were free to include any additional map attributes that they found useful. The best teams had explored space, cleared space, vehicle paths, victim locations, geo-registered victim images, names of grouped areas, confidence in information, and more. An example of an annotated map is shown in Figure 6. Points were once again awarded based on a rank ordering of the maps.

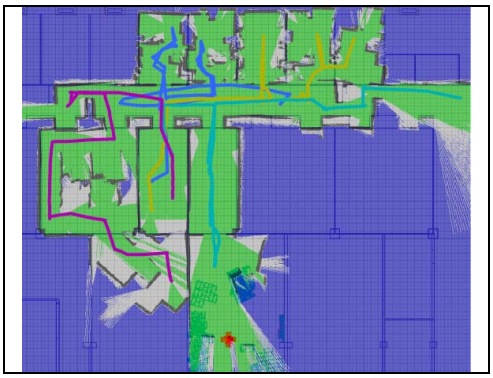

Figure 6: Annotations on map include area explored (gray), area cleared (green), victims located (red cross), and robot paths (multi-colored lines).

- Grouping – A higher order mapping task is 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. Bonus points were awarded for annotating such groups on the map. An example of such groupings is shown in Figure 7. This metric did not change between 2006 and 2007.
- Skeleton quality – A map skeleton reduces a complex map into a set of connected locations. For example, when representing a hallway with numerous doorways, a skeleton may have a line for the hallway and symbols along that line that represent the doors. A map may be inaccurate in terms of metric measurements (a hallway may be shown to be 20 m long instead of 15 m long), but may still present an accurate skeleton (there are three doors before the room with the victim). The category allowed the judges to award points based on how accurately a map skeleton was represented. This metric did not change between 2006 and 2007.

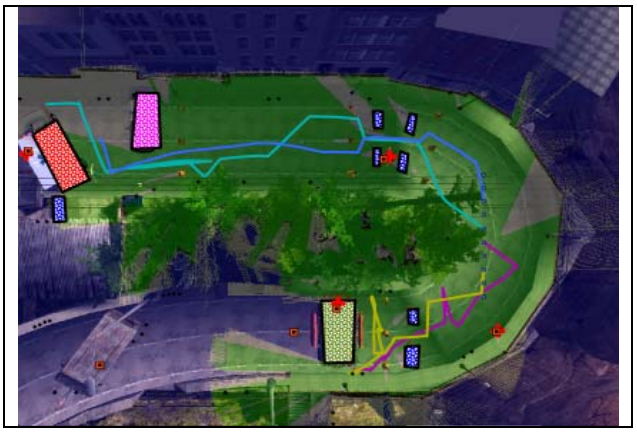

Figure 7: Example of fully annotated and group map. The colored rectangles are keyed to various groups (ambulance, barrier, etc.).

- Utility – One of the main objectives of providing a map was to create the ability for a first responder to utilize the map to determine which areas had been cleared, where hazards may be located, and where victims were trapped. Points were granted by the judges that reflected their feelings on this measure. This metric did not change between 2006 and 2007.

The above mentioned elements were numerically combined according to Equation 1 for 2006 and Equation 2 for 2007.

$$S = \frac{V_{ID} * 10 + V_{ST} * 10 + V_{LO*10} + t * M + E * 50 - C * 5 + B}{(1+N)^2} \quad (1)$$

$$S = \frac{V_{ID} * 5 + V_P * 5 + V_{IP} * 10 + M + E - C * 5 - FA * 5 - V_M * 5}{N^2} \quad (2)$$

The meaning of the variables is discussed below. This equation represents a schema that took into account merit factors that concerned (1) victims discovery, (2) mapping, and (3) exploration. The exact point calculations for each factor are presented below.

1. For victims in 2006, 10 points were awarded for each reported victim ID (V_{ID}). An additional 10 points were granted if the victim's status (V_{ST}) was also provided, and properly localizing the victim in the map was rewarded with an additional 10 points (V_{LO}). Also, at the referee's discretion, up to 20 bonus points were granted for additional information produced (B). For example, some teams managed to not only identify victims, but to also provide pictures taken with the robot's cameras. For this additional information teams were awarded with 15 bonus points. Taking a picture of a victim seemed like a really useful item. Therefore, in 2007, this became a part of the scoring metric (V_P) that was worth 5 points per victim. Correctly geo-referenced victims were worth 5 points if found using the victim sensor (V_{ID}), and 10 points if found using image processing (V_{IP}).
2. Maps were awarded up to 50 points based on their quality (M), as previously described. For the 2006 competition, the obtained score was then scaled by a factor ranging between 0 and 1 (t) that measured the map's feature accuracy. This accuracy was determined through the use of the RFID scoring tags.
3. Up to 50 points were available to reward exploration efforts (E). During the 2006 competition, as the robots were exploring the environment, their poses (on 1 s intervals) were logged. Using the logged position of every robot, the total amount of explored square meters (m^2) was determined and related to the desired amount of explored area. This desired amount was determined by the referees and was based on the competition environment. For example, in a run where 100 m^2 were required to be explored, a team exploring 50 m^2 would receive 25 points, while a team exploring 250 m^2 would receive 50 points, i.e. performances above the required value were leveled off. While this metric was easy to automatically compute, it seemed to reward teams for passing through a location as opposed to actually performing any behaviors while in the location. Therefore a major change was instituted for the 2007 competition.

   For 2007, teams needed to declare where they had explored and where they had cleared. Any victims that existed in a cleared area and were that were not reported by the teams were assessed penalties. The idea being that a map of the environment is useful to responders (therefore award points), and knowing where they do not have to look for victims is even more useful (so award more points). Points were awarded based on a linear scale ranging from 0 – 35 for area cleared and 0 – 15 for area explored. The amount of area that received a top score was the average of the top performing two teams. Exploration above this cutoff was not awarded with additional points. The amount of area explored and cleared by each team was automatically computed based on their maps.

On the penalization side, 5 points were deducted for each collision between a robot and a victim (C). The number of collisions was automatically determined. For 2007, false alarms reported as victims (FA) and victims missed (V_M) in the cleared areas also caused point deductions.

Another parameter that was used to determine the overall score was the number of human operators that were needed to control the robots. The idea was borrowed from the Rescue Robot league with the intent of promoting the deployment of fully autonomous robot teams, or the development of sophisticated human-robot interfaces that allow a single operator to control many agents. In 2006, the overall score was divided by $(1 + N)^2$, where N was the number of operators involved. So, completely autonomous teams, i.e. N=0, incurred no scaling, while teams with a single operator had their score divided by 4. No team used more than one operator. However, for 2007 it was decided that there is no

such thing as a truly operator-less team. At a minimum, an operator must be available to deploy the robots and provide routine maintenance. Therefore, each team was allowed a single operator without a scaling factor.

B. After Action Evaluation

In addition to the scores that teams received during the competition, a large volume of real-time data was logged for post analysis. This information included the actual pose of every robot on a 1 s interval, and a recording of all of the runs. The hope is that teams will be able to combine this information with the environment's ground truth in order to learn from the competition experience.

IV. Future Work

The RoboCup rescue virtual robot competition community remains very active and plans are already underway for the 2008 competition which will take place in Suzhou China. While further evolution of the metrics is inevitable, the main thrust for this year is the automation of the scoring process. Currently, robot-victim bumps are automatically computed as well as the area explored and the area cleared. However, judging the map quality is a manual process. A process that compares competitor generated maps to ground truth and scores map accuracy and utility is an active area of research.

References

1. Asada, M. and Kitano, H., *RoboCup-98: Robot Soccer World Cup II*, Springer-Verlag, Berlin, 1999.

2. Balakirsky, S., Carpin, S., Kleiner, A., Lewis, M., Visser, A., Wang.J., and Ziparo, V., "Towards Heterogeneous Robot Teams for Disaster Mitigation: Results and Performance Metrics from RoboCup Rescue," *Journal of Field Robotics*, Vol. SUBMITTED, 2007.

3. Balakirsky, S., Scrapper, C., Carpin, S., and Lewis, M., "USARSim: Providing a Framework for Multi-robot Performance Evaluation," *Proceedings of the Performance Metrics for Intelligent Systems (PerMIS) Workshop*, 2006.