# The Scoring Procedure for a Competitive Research Competition Influences the Usefulness of the Results in Real-World Applications

Kristopher M. Thornburg and Geb W. Thomas, PhD

*Abstract*— **In the 2006 RoboCup Virtual Rescue competition, teams from different research labs developed methods for controlling teams of mobile robots in a simulated urban search and rescue scenario. This paper reviews the strategies and scores from the top six competitors. The scoring procedure used in this inaugural competition rewards participants for the number of victims found, the amount of area explored in the environment, the quality of the maps created by the robot teams and penalties participants for colliding with a victim or relying on human operators. The analysis of the strategies and scores suggests that the scoring procedure may lead teams to adopt strategies that are not consistent with the needs of a real search and rescue scenario. Individual robot contributions to the system were reviewed to account for the costs associated with adding a robot to the environment, indicating that value added per robot is an important measure that is overlooked. The analysis of the impact of human operator penalties on scoring revealed an overemphasis on fully autonomous robotic systems. The analysis also revealed substantial performance variation, depending on the behavior that was being rewarded, which may indicate a lack of focus for evaluative performance measures of robotic urban search and rescue systems. The competition has the potential to provide influential research in this area if a proper scoring procedure that reflects actual research needs is implemented. In order to ensure that research gains made as a result of the competition process are useful to the application community, it is essential that the rules be tuned to the application needs. It is likely that, as competitions and games are becoming a growing part of the research community, this sensitivity is managed along with the other political, social and interactive demands involved in setting rules for research competitions.**

## I. Introduction

USING robots for urban search and rescue activities first occurred in 2001 in response to the World Trade Center disaster in New York City [1], [2]. Since that time, there has been much interest in using robots as part of urban search and rescue teams, though much of the research has resulted in anecdotal observations [1]-[3]. There is a genuine need for quantitative and repeatable research in this area. RoboCup, in an effort to encourage research, innovation, and advancement in urban search and rescue, recently introduced a new competition focused on developing control mechanisms for robots in a virtual setting [4]. By introducing a simulation competition, the costly and time-consuming mechanical aspects of the robot are eliminated, allowing the competition to focus on robotic control. The simulation also allows for repeatable trials, quantitative data collection, and cross-institutional cooperation that did not previously exist. The competition requires an open source policy for competition algorithms, further supporting the advancement of development in robotic urban search and rescue.

This effort to use a research competition to stimulate interest and progress in a research area is a growing trend. RoboCup, in which different teams compete to develop teams of robots to compete in different leagues, is an outgrowth of a desire to rapidly advance and share results within the autonomous robotics community [4]. Recently DARPA has adopted the same approach with the DARPA Grand Challenge to encourage researchers to build autonomous off-road navigation robots for a large cash prize. NASA has also experimented with offering prizes for various competitions to encourage students and researchers to focus on problems relevant to NASA's needs. These competitions are a useful and fun way to advance various research goals, but their introduction is rather recent and it is unclear how quickly and effectively the competitions will achieve the research goals of the people behind their introduction.

This paper examines the details of one competition, the scoring procedure, and the implications of that scoring procedure on the strategies and consequent technologies that might be developed for that competition.

### A. The Urban Search and Rescue Simulator (USARSim)

The simulated disaster environments and robots are powered by the high fidelity Unreal Tournament 2004 game engine (Epic Games, Inc., Raleigh, N.C., USA), interfaced by an open source software package called USARSim. USARSim, originally developed at the University of Pittsburgh and now supported by NIST, allows researchers to develop realistic models of robots and control them within the Unreal Tournament 2004 architecture. Unreal Tournament contains its own environment editor that allows researchers to develop their own environments to exacting specifications.

### B. The Virtual Rescue Competition

The 2006 competition relied on two simulated environments relevant to a real urban search and rescue situation. The first environment, a damaged, multilevel office building, contained rubble, uneven surfaces, and flames. The second environment, a rubble-filled city street, featured

Manuscript received March 16, 2007.

K. M. Thornburg is with the Department of Mechanical and Industrial Engineering, The University of Iowa, Iowa City, IA 52242 USA (phone: 319-384-0526; e-mail: kristopher-thornburg@uiowa.edu).

G. W. Thomas, PhD, is with the Department of Mechanical and Industrial Engineering, The University of Iowa, Iowa City, IA 52242 USA (e-mail: geb-thomas@uiowa.edu).

uneven surfaces, flaming and overturned vehicles, and a park area with trees. Both environments, developed for the competition by NIST, were vast (several thousand square meters) and contained victims dispersed throughout in a random fashion. Each simulated victim was equipped with a radio frequency identification (RFID) tag which transmitted the victim's name and relative location when the robot was within one meter of the victim. Additional information was transmitted if the robot reached a closer distance threshold to the victim. False alarms (detecting a victim that was not present) were also possible. Additional RFID tags were dispersed throughout each environment for judging and scoring purposes.

All competition participants had access to the same robotic platforms (different sized wheeled and tracked robots) and sensors (such as sonars, cameras, and laser range finders). The robotic platforms had maximum payload specifications that required participants to carefully configure the robots with sensors. Sensor feedback was simulated as closely to real sensor feedback as possible, adding to the simulator's fidelity. The number of robots used by any participating team was not limited and communications between robots was not limited by bandwidth or other constraints.

Each trial, or run, was limited to 20 minutes, in which time each robot was to explore and map as much of the area and locate as many victims as possible. Each team started from approximately the same position and explored the same environment during each run. At the end of the 20 minute run, each team was allowed 10 minutes to compile the files necessary for scoring. The files submitted by each team for scoring a run included an image file of the map created by the robot, integrated with other robot maps if multiple robots were used in a run, a list of victims found with locations, and any additional information about the victims collected by the robot, a list of RFID tags and associated locations detected in the environment, and any images of victims recorded by the robot. Additional performance measures recorded automatically by the simulator server included the amount of area explored in square meters and the number of robot collisions with victims. The judges used this information to determine the score for each run. Table 1 describes point values given for particular aspects of each run. Equation (1) shows how the points were combined to form the final score.

## II. Methods

To determine what types of robotic control worked in the urban search and rescue simulation, the raw scores of the top six competitors of RoboCup 2006 Virtual Rescue were collected. To analyze the scores, specific aspects of the scores, or performance measures, were reviewed, aggregated, and compared as well as the overall scores as calculated by the general scoring algorithm (Equation 1). These performance measures include number of victims found, amount of area explored, overall map quality, and penalizing factors. The number of robots used by each team was also considered to

TABLE I
MERIT AND PENALIZING FACTORS WITH ASSOCIATED POINT VALUES

| Merit Factors | Variable | Point value |
|---|---|---|
| Found victim | V | 10 |
| Victim status reported | Vs | 10 |
| Victim bonus (picture, etc.) | Vb | up to 20 |
| Map visual quality | Mv | up to 50 |
| Map metric quality | Mm | 0 to 1 |
| Map total | | Mv*Mm |
| Area explored (environment dependent) | A | up to 50 |
| **Penalizing Factors** | | |
| Number human operators | N | divide total score by $(N+1)^2$ |
| False victim identification | Vf | -5 |
| Victim collision | Vc | -5 |

$$Score = \frac{[(V*10)+(Vs*10)+Vb+(Mv*Mm)+A]-[(Vf*5)+(Vc*5)]}{(N+1)^2} \quad (1)$$

TABLE II
TOTAL SCORE FOR EACH RUN FOR EACH COMPETITOR

| Team | Run 1 Score | Run 2 Score | Run 3 Score | Run 4 Score | Run 5 Score | Total Score | Standard Deviation |
|---|---|---|---|---|---|---|---|
| Yellow | 120.00 | 86.00 | 125.00 | 220.00 | 442.09 | 993.09 | 144.93 |
| Red | 56.50 | 68.25 | 115.36 | 326.83 | 266.07 | 833.01 | 122.46 |
| White | 35.50 | 43.40 | 60.00 | 235.23 | 99.00 | 473.13 | 82.32 |
| Black | 42.88 | 33.63 | 54.43 | 88.58 | 122.34 | 341.85 | 36.65 |
| Blue | 14.81 | 31.49 | 7.05 | 58.66 | 95.49 | 207.49 | 36.08 |
| Green | 10.94 | 10.06 | 30.12 | 41.98 | 46.13 | 139.22 | 16.89 |

analyze performance measures per robot. The final analysis changed the definition of human operator in an effort to reflect reality by charging every team one human operator whether or not the robots operated autonomously. The competitors' names were changed for ease of analysis.

## III. Results

Table 2 shows the total scores received during each run and overall score for each team with the standard deviation of those scores. The table is sorted from highest total score for the five runs to the lowest.

The type of control utilized by each team (autonomous, combination of autonomous behavior and teleoperation, and teleoperation) and the average number of robots used for each run is shown in Table 3. Once again, the table is sorted from highest total score to the lowest. Table 3 also indicates the average score accounted for by each robot. Only one team utilized one fully teleoperated robot, two teams utilized a combination of teleoperation and autonomous activities, and three teams were fully autonomous systems.

TABLE III
TYPE OF CONTROL AND NUMBER OF ROBOTS USED BY EACH TEAM

| Team | Type | # Robots | Total Score | Score / Robots |
|---|---|---|---|---|
| Yellow | Autonomous | 8 | 993.09 | 124.14 |
| Red | Autonomous | 6 | 833.01 | 138.84 |
| White | Autonomous | 6 | 473.13 | 78.85 |
| Black | Teleop & Auto | 6 | 341.85 | 56.98 |
| Blue | Teleop & Auto | 4 | 207.49 | 51.87 |
| Green | Teleoperated | 1 | 139.22 | 139.22 |

The totals from the five runs for the number of victims found, area explored, and mapping score are shown in Figs. 1, 2, and 3 respectively.

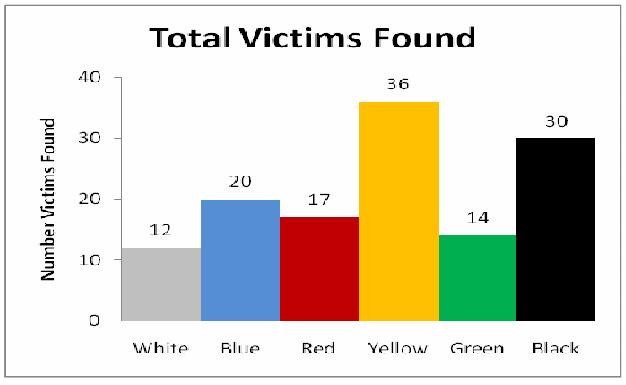


Figure 1. The total number of victims found over five runs by each team.

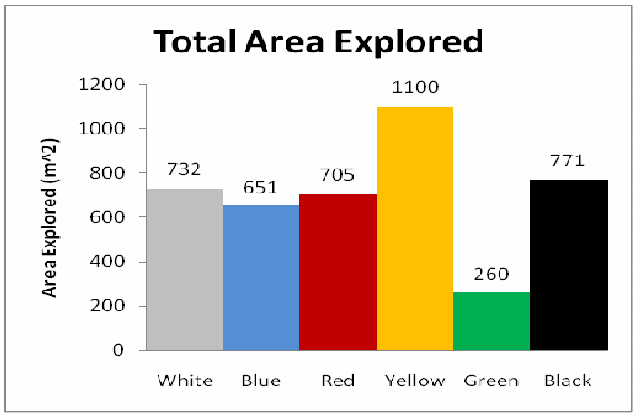


Figure 2. The total area explored over five runs by each team.

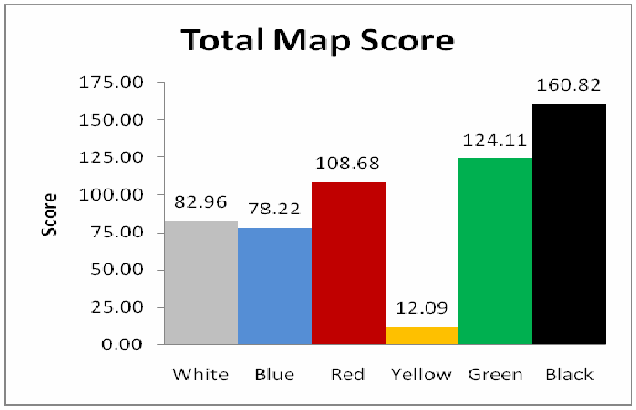


Figure 3. The total mapping score over five runs by each team.

The total average contribution from each robot on each team was calculated. The number of victims found on average by each robot, the average area explored by each robot, and the total average score contribution by each robot are shown in Figs. 4, 5, and 6 respectively.

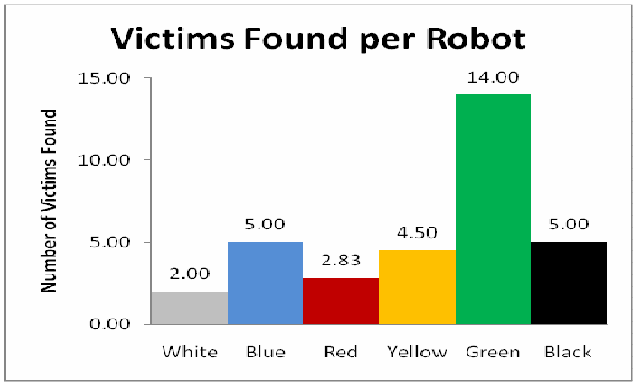


Figure 4. The average number of victims found per robot for each team.

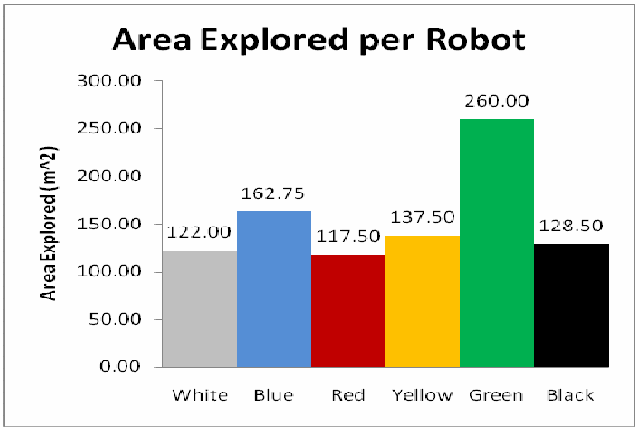


Figure 5. The average area explored per robot for each team.

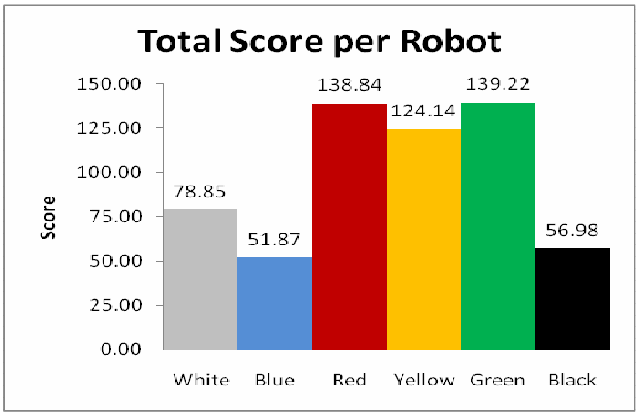


Figure 6. Total average score per robot for each team.

The effectiveness of each robot for locating victims can be determined with a new metric by calculating the ratio of victims found to the area explored, revealing a measure of exploration quality. Since each team's robots started from approximately the same position within the same environment on any given run, a general effectiveness comparison can be made. Fig. 7 indicates the ratio of the total number of found victims to the total area explored for each team.

To determine the effect of the scoring penalty of human operators, the total scores were recalculated, charging each team with one operator. The results of this recalculation are shown in Table 4 in which the original score and rank are shown along with the adjusted score and adjusted rank for each team.

## IV. Discussion

Preliminary conclusions based solely on the final scores indicate a preference to autonomous systems (Table 3) as an effective means to searching for and finding victims in an urban search and rescue environment. Table 3 also indicates a preference toward using more robots during a search and

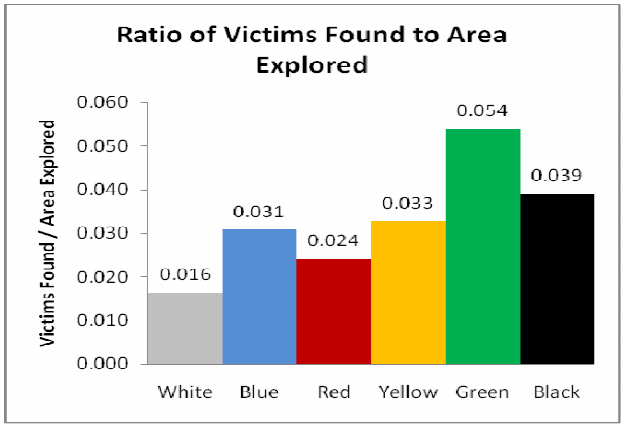


Figure 7. The ratio of victims found to area explored by each team.

TABLE IV
ORIGINAL SCORES AND RANKS WITH ADJUSTED SCORES AND RANKS

| Team | Total Score | Rank | N=1 Operators Score | New Rank |
|---|---|---|---|---|
| Black | 341.85 | 4 | 341.85 | 1 |
| Yellow | 993.09 | 1 | 248.27 | 2 |
| Red | 833.01 | 2 | 208.25 | 3 |
| Blue | 207.49 | 5 | 207.49 | 4 |
| Green | 139.22 | 6 | 139.22 | 5 |
| White | 473.13 | 3 | 118.28 | 6 |

rescue mission, as the higher scoring teams had higher number of robots. However, it seems that, depending on the control algorithms used, increasing the number of robots will not increase the contribution of each robot. Fig. 6 shows the average score accounted for by each robot for each team. The Green team, only using one robot, had essentially the same robot score contribution as the Red team which was using six robots and a higher robot score contribution than the Yellow team which used eight robots. The other three teams seemed to be relatively ineffective in using their robots.

Perhaps the most important aspect of search and rescue is locating victims within the environment. Fig. 1 displays the total number of victims found by each team over the five runs. The Yellow team found the most victims (36), followed by the Black team (30). Interestingly, the Green team, running only one robot, found 14 victims, which is fairly close to the Red team's 17 with six robots and overcomes the White team's six robot count of 12. Additionally, Fig. 4 indicates the average number of victims found by each robot for each team. Since the Green team only ran one robot, that robot gains all the finds (14), while the robots on the other teams found somewhere between two and five victims each. Interestingly, the Blue team and the Black team, both utilizing a combination of teleoperation and automation, found more victims per robot (5) than the fully autonomous teams.

Another important aspect of search and rescue is the exploration of the environment. The maximal amount of area should be explored to ensure all victims have been located. The maximum amount of area explored was 1100 square meters by the Yellow team with 8 fully autonomous robots, as shown in Fig. 2. The maximum area explored per robot was performed by the Green team, as shown in Figure 5, with one teleoperated robot.

An important output from robotic search and rescue activities is the production of a map of the environment, indicating the position of all the detected victims. This map is particularly important to search and rescue personnel, who will base their rescue operation plans on the map produced by the robot. The total map scores, as indicated in Fig. 3, show the Black team producing the highest scoring maps, followed by the Green team. It is interesting to note that the Yellow team, although exploring the most area and ultimately pronounced the winner of the competition, produced significantly lower scores for map production than any of the other teams. This team elected not to produce a map in the first four runs.

The ratio of found victims to explored area presents an interesting measure of robot effectiveness to detect victims within the environment. Fig. 7 presents the results of these calculations. The Green team has the highest victims-found-to-area-explored ratio. This is possibly because a human was in control of the robotic system during the entire run, rather than operating from a supervisory position like the operator of the Black team, which had the second highest ratio. With the exception of the Yellow team, all the teams with a human in some sort of control capacity had a higher ratio than fully autonomous teams, indicating that the human had a direct role in locating victims within the environment.

The scoring penalty for human operators introduced in the competition is has an interesting effect on the final scoring of the robotic systems in this competition. As seen in Table IV, if each team is charged with at least one human operator, the rankings of the teams changes considerably. In this case, the Black team wins by a considerable amount. These results warrant further investigation into the scoring algorithm itself to determine whether it reflects reality in the proper weighting of mission attributes (number of victims found and area explored) versus the penalties incurred, such as the number of human operators.

### *A. Consequences of the Scoring Procedure*

Given a good search algorithm, the more robots involved in the search result in more area being explored and more victims being located. This can be seen in Fig. 1 referring to number of victims found and Fig. 2 referring to the total area explored. The Yellow team, which ran 8 robots, located a total of 36 victims and explored a total of 1100 square meters. Interestingly, the Black team, which located the second highest number of victims (30) and explored the second highest amount of area (771 square meters), placed fourth in the final score calculation, indicating a severe overemphasis on autonomous robotic systems. Removing the penalty for a human operator from the scoring procedure would correct this effect.

Different performance measures may indicate different levels of success. If the overall robotic system is considered with the current scoring algorithm, the Yellow team created the best robotic system, as indicated in Table II. If each team is charged an operator, which mirrors reality because operators are required to set up the robots at the disaster site, then the Black team had the best approach to robotic urban search and rescue, as indicated in Table IV. If each individual robot contribution is considered alone, then the Green team had the best approach, as seen in Figs. 4 and 5. The overall score contribution per robot, as shown in Fig. 6, indicates that ineffective robot use should be a penalizing factor to reflect the cost of adding robots to the environment. It is quite obvious from these differing measures that the performance measures for robotic urban search and rescue need to be standardized to focus on the most important part of a search and rescue mission.

Both of these lessons lead into an analysis of the scoring procedure. If the current scoring procedure remains, the systems produced for this competition will begin gravitating toward the control of large teams of autonomous robots. However, it seems likely that, at least for the foreseeable future, USAR will be dominated by the need to use robots to search for victims in cooperation with human teams. The ability of the robot to exhaustively and reliably search a region and to provide excellent documentation about what areas were searched and exactly where victims were found is likely to be the most important criteria for many years.

### *B. Real-World Implications*

As the analysis above demonstrates, the effectiveness of a robotic urban search and rescue system cannot be easily reduced to a single measure. However, the details of the reduction have a tremendous impact on the strategies and directions the research is likely to take.

The scoring strategy for the 2006 RoboCup Virtual Rescue competition induces specific strategies. The first is that, depending upon the interpretation of the scoring algorithm, the best robotic systems have several robots and are fully autonomous, such as the Yellow team, or have a combination of teleoperation and autonomy, such as the Black team. From an individual robot perspective, the best approach is strictly teleoperated, as seen in Figs. 4, 5 and 7 by the Green team.

The scoring procedure does not weigh the location of victims and the production of an area map in the way search and rescue personnel would [5]. The Yellow team elected to produce no area map for four runs and a low quality map for the fifth run but still placed higher than the Black team which produced the top score in mapping. The ratio of found victims to area explored, which places a value of quality on the victim location effectiveness of each robotic system, indicates that perhaps a number of victims were missed in the runs. Unfortunately, the number of victims missed by teams was not recorded. The number of missed victims, which is an

important aspect from a real-world standpoint, would have provided yet another metric of the usefulness of a particular system in this situation which could be combined with the location effectiveness measure to create a quality of search metric.

Because it is likely that any USAR system that is actually used in the field will involve a human operator, it seems helpful to focus on systems that will effectively use that human operator. The ranking of the competition is does not reflect current urban search and rescue priorities or activities. Human in-the-loop systems in this competition provide just as good, if not better, performance than fully autonomous systems, based on real-world needs.

Overall, the scoring procedure used in this competition must be modified in the following ways, based on the data acquired from the first competition. First, the penalty for human operators must be significantly modified or completely removed. Second, a cost should be associated with utilizing multiple robots to reflect the investment and increased risk associated with using large robot teams. If multiple robots are used, multiple humans are required to pack in the robots to the starting location, thereby increasing the cost of human involvement. Third, appropriate weight must be applied to the creation of a useful map for rescue workers, based on actual urban search and rescue procedures. Fourth, appropriate weight must be applied to locating victims and localizing them on a map, as well as appropriately weighted penalties for missing victims in the environment. And fifth, appropriate evaluative measures must be introduced in order to examine the effectiveness of the robots and the value added to the mission by including more. An example measure is the quality of search metric, which is the ratio of victims found to area explored.

## V. Conclusion

The scoring system for the Virtual Rescue competition clearly favors large teams of autonomous mobile robots that search quickly and find obvious victims. However, the immediate needs of the USAR community are more focused on careful search and victim localization. Assuming that the teams that choose to compete select strategies that will be most favorable for winning the tournament, it is likely that they will favor the strategies that win the tournament rather than those that may be helpful in an applied setting. Over time, this may restrict the cross-fertilization between the two communities and ultimately limit the success of using the competition as a medium to stir interest and motivation towards the practical problems identified for the USAR community.

Although this study is focused on one particularly competition, it is likely that other competitions also face similar challenges, as they seek to make the "game" playable and interesting for a large potential audience. At the same time, the organizers must meet the potentially conflicting needs of the sponsors and players who each have their favored technologies and wish to design the competition to emphasize their own particular interests. Consequently, the organizers must either choose a task with a simple and objective that is effectively beyond debate, or else choose the rules for scoring with great care so that the emergent strategies adopted by the players will match the research needs in the community that is promoting the competition.

## References


[1] J. Casper and R. Murphy, "Human-Robot Interactions During the Robot-Assisted Urban Search and Rescue Response at the World Trade Center," *IEEE Trans. on Systems, Man, and Cybernetics Part B: Cybernetics*, 33 (3), p. 367-385, 2003.

[2] R. Murphy, "Trial by Fire: Activities of the Rescue Robots at the World Trade Center from 11-21 September 2001," *IEEE Robotics and Automation Magazine*, 2004.

[3] J. Burke, R. Murphy, M. Coovert, and D. Riddle, "Moonlight in Miami: A Field Study of Human-Robot Interaction in the Context of an Urban Search and Rescue Disaster Response Training Exercise," *Human-Computer Interaction*, 19, p. 85-116, 2004.

[4] M. Lewis, S. Hughes, J. Wang, M. Koes, and S. Carpin, "Validating USARsim for use in HRI research," *Human Factors and Ergonomics 49th Annual Meeting*, 2005.

[5] S. Dolan, Training Officer for the Iowa City Fire Department, Personal Interview, October 10, 2006.