Assessment of the NIST Standard Test Bed for Urban Search and Rescue

Robin Murphy, Jenn Casper, Mark Micire, Jeff Hyams
Computer Science and Engineering
University of South Florida
Tampa, FL 33620
{murphy, jcasper, mmicire, hyams}@csee.usf.edu

Abstract

The USF team in the 2000 AAAI Mobile Robot Competition had the most extensive experience with the NIST Standard Test Bed for USAR. Based on those experiences, the team reports on the utility of the test bed, and makes over 20 specific recommendations on both scoring competitions and on future improvements to the test bed.

Introduction

A team of three operators and two robots from the University of South Florida (USF) tested the NIST standard test bed for urban search and rescue (USAR) as part of the 2000 AAAI Mobile Robot Competition USAR event. The test bed consisted of three sections, each providing a different level of difficulty in order to accommodate most competitors (see Fig. 1). The easiest section, Yellow, contained mainly hallways, blinds, and openings to search through. The course could be traversed by a Nomad type robot. The intermediate Orange Section provided more challenge with the addition of a second level that was reachable by stairs or ramp. Other challenges included those found in the yellow as well as some added doors. The Red Section was intended to be the most difficult. It contained piles of rubble and dropped floorboards that represented a pancake-like structure. The Orange and Red sections clearly required hardware that was capable of traveling such spaces.

In addition to USF, three other teams entered the AAAI competition's USAR event: Kansas State, Swarthmore College, and University of Arkansas. The Kansas State team dropped out due to hardware failures on site. The Swarthmore and Arkansas teams fielded Nomad scout types of robots that operated only in the Yellow Section. The performance of each team is unclear as the judges did not record how many victims were found and how many victims were missed. At the time of publication of this paper, the awards for the event were under protest. Swarthmore had a single robot which attempted to enter a room, perform a panoramic visual scan for possible victims, mark the location on a map, and then enter another room and so on. At the conclusion of their allotted time, the robot was retrieved and the contents

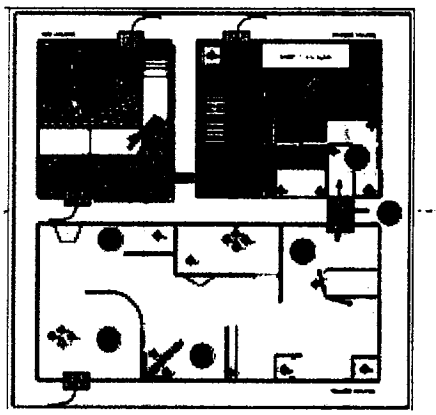

Figure 1: Overview of the NIST USAR arena.

of the map was made available to the judges. They entered one room successfully and it is believed they identified up to two surface victims. The Arkansas team used two Nomad scout type robots; however, each robot was physically placed in a room, and the team was allowed to repeatedly move and reset the robots as needed. The Arkansas team found at least one victim, and communicated this by repeatedly ramming the mannequin.

The USF team used two outdoor robots: 1) a RWI ATRV with sonar, video, and a miniature uncooled FLIR and 2) a customized RWI Urban with a black and white camera, color camera, and sonars. This was intended to be a marsupial pair, but the transport mechanism for the team was still under construction at the time of the competition. The USF team used a *mixed-initiative* or *adjustable autonomy* approach: each platform was teleoperated for purposes of navigation but ran a series of software vision agents for autonomous victim detection: motion, skin color, distinctive color, and thermal region. The user interface displayed the extraction results from each applicable agent and highlighted in color whenever the agent found a candidate. A fifth software agent ran on the ATRV which fused the output of the four agents, compensating for the physical separation between the video and FLIR cameras. It beeped the operator when it had sufficient confidence in the presence of a victim, but the beeping had to be turned off due to a high number of false positives

generated by the audience. The ATRV found an average of 3.75 victims per each of the four runs recorded, while the Urban found an average of 4.67 victims. A fifth run was not recorded and no data is available.

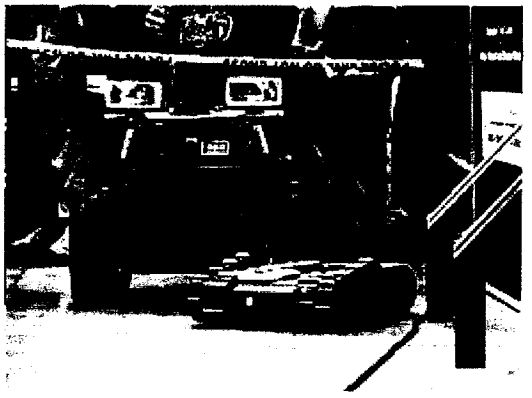

Figure 2: The USF USAR robot team, Fontana (ATRV) and Klink (Urban) (named after two women Star Trek writers).

In addition to participating in the competition (both a preliminary and a final round), the USF team hosted three complete exhibition runs as part of the AAAI Robot Exhibition Program and did numerous other partial exhibitions for the news media at the request of AAAI. The other teams did not exhibit. As such, the USF team had the most experience with the most difficult sections of the test bed and can claim to represent user expertise.

This paper discusses the NIST test bed from the USF experience, and makes recommendations on scoring, improving the test bed, and staging a more USAR-relevant event at RoboCup Rescue in 2001.

Assessment of the Three Sections

The NIST test bed is an excellent step between a research laboratory and the rigors of the field. For example, USF has a USAR test bed (Fig. 3), but it is somewhere between the Yellow and Orange sections in difficulty and cannot provide the large scale of the NIST test bed. One advantage is that the test bed sections can be made harder as needed. An important contribution that should not be overlooked is that the test bed appeared to motivate researchers we talked to: it was neither too hard nor too trivial. This is quite an accomplishment in itself.

Yellow

The USF team did not compete or exhibit in the Yellow Section, entering only for about 1 hour of practicing collaborative teleoperation. Our assessment was that the section was far too much of an office navigation domain- the over-turned chair in one of the rooms was the only real surprise. Only one room had a door and neither Swarthmore nor Arkansas reached it. The arena was at about the level of complexity seen in the Office Navigation Event thread of the AAAI Robot Competition in the mid-1990's.

Figure 3: The USF USAR testbed, a mock-up of a destroyed bedroom.

Orange

The Orange Section consisted of a maze plus a second story connected by a ramp and stairs. Unlike the Yellow Section, the doorways into the Orange and Red Sections had cross-members crowning the doorway at about 4 feet high. This added some feel of confined space. The USF robots entered a very confined maze of corridors to find a surface victim. The Urban served as point man, exploring first, then guiding the ATRV if it found something requiring confirmation or IR sensing. The maze had hanging Venetian blinds in the passage way, and the Urban almost got the cord tangled in her flipper.

The Orange Section also had two forms of entry in the main search area after the robots had navigated the maze. One entry was through the X made by cross-bracing the second story. The Urban could navigate under the cross-bracing, but the ATRV could not. The second form of entry was through a door on hinges. The Urban pushed the door open for the ATRV to enter the main search space (Fig. 4). The Urban attempted to climb the stairs, but the first step was too high for the design. (A Matilda style robot also attempted to climb the stairs but could not either.) It went to the ramp and climbed to the second story.

Figure 4: The Urban holds the door for the ATRV in the Orange Section.

The USF robot was able to avoid negative obstacles (a stairwell and uncovered HVAC ducting in the floor of the second story) to find victims on the second story (Fig. 5). The modified Urban actually flipped its upper camera onto the HVAC hole and peered inside the duct. This shows the utility of having multiple sensors and in different locations.

The Orange Section is also to be commended for providing some attributes of 3D or volumetric search. For example, an arm was dangling down from the second story and should have been visible from the first floor. Note that the dangling arm posed a classic challenge between navigation and mission. The mission motivates the robot or rescuer to attempt to get closer and triage the victim, while the navigational layout prevents the rescuer from approaching without significantly altering course, and even backtracking to find a path.

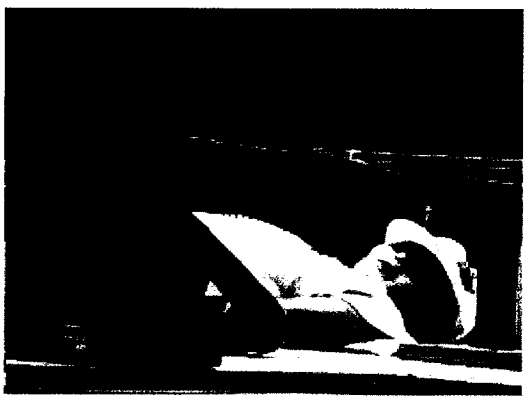

Figure 5: Close up of victim lying on the second floor of the Orange Section.

Red

The Red Section at first appeared harder (Fig. 6), however, in practice it was easier for the ATRV than the Orange Section due to more open space. The floor was made up of tarps and rocks on plywood. The ATRV and Urbans were built for such terrain. The Red Section contained two layers of pancaking, with significant rubble, chicken wire, pallets, and pipes creating navigation hazards for the Urban. Only about 30% of the area was not accessible to the larger ATRV due to the large open space.

One nice attribute of the Red Section is that it lends itself to booby-traps. The pancake layers were easily modified between runs to create a secondary collapse when the Urban climbed to the top. Using current technology, the Urban operator and/or software agents could not see any signs that the structure was unstable.

Recommendations on Scoring

The AAAI Competition did not use any metric scores for their USAR event, relying entirely on a panel of four judges, none of whom had any USAR experience. The AAAI Competition had published metrics prior to the competition that were to be used in scoring,(Meeden 2000) but did not use

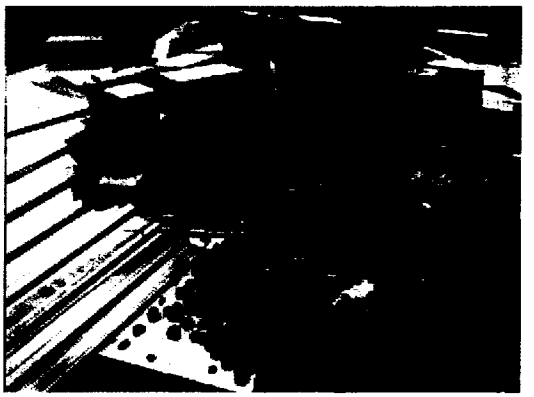

Figure 6: Overview of the Red Section.

those metrics on site and the scoring was subjective. The published metrics appeared to be a good first start (with our reservations given below) and no reason was given why AAAI abandoned them.

1. Use quantitative scoring, at least as a basis for the competition. The scores might be modified by a qualitative assessment of the AI involved, but there should be a significant numerical aspect to the scoring.
2. Distribute victims in same proportions as FEMA statistics given in FEMA publication USFA/NFA-RS1-SM1993 (FEMA 1993) and award points accordingly. Detecting a surface victim and an entombed victim require much different sensing and intelligence.

Surface	50%
Lightly trapped	30%
Trapped in void spaces	15%
Entombed	5%

3. Have a mechanism for unambiguously confirming that the victims identified were identified. It was not clear to the audience when a victim had been correctly detected or when the robot had reported a false positive. Perhaps an electronic scoreboard showing the number of false positives and false negatives (missed victims) could be displayed and updated during the competition. (Swarthmore used beeping and USF flashed the headlights. The judges appeared to accept that if there was a victim in the general direction of the robot's sensors at the time of the announced detection that a victim had been found. In the case of USF, only one judge took time during the competition look at the technical rescue display workstation, which provided both the sensor data and the fused display, to confirm what the robot was seeing.)
4. Points for the detection of a victim should also depend on the time at which the technical rescue crew is informed of the discovery and the accuracy of the location, either in terms of absolute location or a navigable path for workers to reach the victim. Robots which overcome inevitable communications problems by creating a relay of "commsbots" or returning to locations where broadcasting worked are to be rewarded. (The Swarthmore robot beeped when

it thought it found a victim, but in terms of truly communicating that information to rescue workers, it stored the location of all suspected victims until the competition was ended. In practice, if the robot had been damaged, the data would have been lost. Also, the map was not compared quantitatively to the ground truth.)

5. Contact with the victims should be prohibited unless the robot is carrying a biometric device that requires contact with the victim. In that case, the robot should penalized or eliminated from competition if contact is too hard or otherwise uncontrolled. (The Arkansas robots repeatedly struck the surface victim it had detected.)
6. Fewer points should be awarded for finding a disconnected body part (and identifying it as such) than for finding a survivor.
7. Require the robots to exit from the same entry void that they used for entry. This is a strict requirement for human rescue workers in the US, intended to facilitate accounting for all resources. (The AAAI Competition permitted exiting from anywhere on the grounds that the robot may need to find a clear spot to communicate its results.)
8. Have all competitors start in the same place in the warm zone, and do not permit them to be carried by human operators inside the hot zone. The exception is if the robot has to be carried and inserted in an above grade void from the outside. (Swarthmore and Arkansas manually placed their robots in the yellow section, with Arkansas actually placing their robots within specific rooms in the yellow section.)
9. Do not permit human operators to enter the hot zone and reset or move robots during the competition. (Arkansas team members repeatedly entered the hot zone to reboot errant robots and to physically move robots to new rooms to explore.)
10. Have multiple runs, perhaps a best of three rounds approach used by AUVSI. (NIST "booby-trapped" the Red Section after the AAAI Preliminary Round, making it extremely easy to create a secondary collapse. This was done to illustrate the dangers and difficulties of USAR. However, if the AAAI rules had been followed, this would have resulted in a significant deduction of points from the USF team, and quite a different score between runs. The difficulty of the courses should be fixed for the competition events, and changed perhaps only for any exhibitions.)

It should be clear from the above recommendations that a quantitative scoring system which truly provides a "level playing field" is going to be hard to construct. Unlike RoboCup, where the domain is a game with accepted rules and scoring mechanisms, USAR is more open. In order to facilitate the relevance of the competition to the USAR community, we recommend that scoring mechanisms be derived in conjunction with USAR professionals outside of the robotics community and with roboticists who are trained in USAR. We propose that a rules committee for RoboCup Rescue physical agent be established and include at least one representative from NIST, NIUSR, and one member of the research community who had worked and published in USAR.

Recommendations for Improving the NIST Testbed

The NIST testbed was intended to be an intermediate step between a research laboratory and a real collapsed building. The three sections appeared to be partitioned based on navigability, rather than as representative cases of severity of building collapses or perceptual challenges. For example, the basic motivation for the Yellow versus the Orange and Red Sections appeared to be to engage researchers with traditional indoor robot platforms (e.g., Nomads, RWI B series, Pioneers, and so on). An alternative strategy might be to consider each section more realistically, where the Yellow Section would be a structurally unsound, but largely navigable, apartment building, the Orange Section might be an office building in mixed mode collapse such as many of the buildings in the 1995 Hanshin-Awaji earthquake, and the Red Section might be a pancake collapse such as seen in the front of the Murrow building at the Oklahoma City bombing. This approach would permit traditional indoor robot platforms to navigate, but require advances in detection of unfriendly terrain such as throw rugs or carpet, doors, etc.

For All Sections

In addition to the suggestions made above, we offer some possible improvements to the test bed:

1. Create void spaces in each section more typical of USAR (Fig. 7). In particular, there were no lean-to and V void spaces (USFA 1996; NFPA 1999) in any of the 3 sections. The red section did have some light pancaking. Victims in even the Yellow Section should be placed behind furniture and occluded by fallen furniture or even sheet-rock or portions of the ceiling.

Figure 7: Infrared images of a lightly trapped, void trapped, and entombed victim.

2. Put tarps and high powered lights ("beams of sunlight") over portions of all courses to create significant changes in lighting conditions, most especially darkness. As it stands now, the testbed is a poor test of the utility of infra-red.
3. Entries were all doors at grade. Many voids are actually above grade, irregular, and have been knocked in the wall, even in buildings that have not collapsed. Each section should have one or more above grade entry voids from the "outside". This will support the testing of concepts for automating the reconnaissance and deciding how to deploy resources, as per the rescue and recovery of lightly trapped victims, use of reconnaissance results to

locate lightly trapped victims, and searching void spaces after hazard removal phases of a structural collapse rescue (Casper, Micire, & Murphy 2000).

4. Each section should contain more human effects. For example, the Yellow and Orange Sections should have throw rugs on the floors, fallen debris such as magazines, books, bills, toys, etc. Otherwise, the Yellow Section is actually easier than the Office Navigation thread in the AAAI competitions during the mid-1990's.

5. Each section should contain real doors with door knobs or at least the commercial code handles for disabled access. The doors in the Yellow and Orange section were both easily opened panels. (USF was able to easily identify the swinging door in the Orange Section and use the Urban to open the door for the ATRV to pass through. None of the other teams got to the room with the door in the Yellow Section). All rooms in any section should have doors and some of those doors should be off their hinges or locked. This will test the advances in object recognition, reasoning, and manipulation.

6. If possible, victims should produce a more realistic heat profile than a heating pad. This is needed for detection and to test advances in assessment of the context of the victim (how much they are covered, etc.).

For the Orange and Red Sections

1. Cover everything with dust to simulate the cinder block and sheet-rock dust that commonly covers everything in a building collapse. Victims who are alive often move enough to inadvertently shake off some of this dust, making color detection a very important component of victim detection. (USF used a "distinctive color detector" as one of their four vision agents. The distinctive color agent looked for regions of color that were different than the average value. This appeared to work during the competition for the Red Section, which was less colorful (no wallpaper, etc.), but there wasn't enough data to draw any statistical conclusions.)

2. Make the surfaces uneven. All the surfaces were level in their major axis; even the ramp in the Orange Section was flat, not canted to one side.

3. Use real cinder blocks. The USF Urban was able to move the faux cinder blocks on the ramp in the Orange Section rather than navigate around (Fig. 8).

4. Make a "box maze" for entry to introduce more confined space. Rescue workers who are certified for confined space rescue use a series of plywood boxes which can be connected together to form long, dark, confined mazes. The mazes are easily reconfigured. A similar box maze could be constructed from the lightweight paneling material.

5. The terrain of both sections was still fairly easy compared to the field, and dry. Perhaps as robot platforms evolve, the courses should contain water.

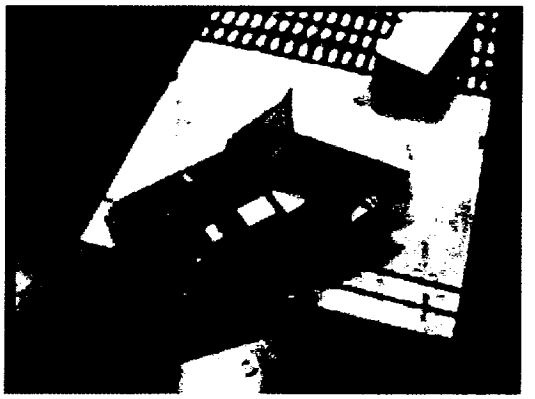

Figure 8: The Urban has pushed the cinder block around rather than traversed over it.

Other Suggestions

The testbed is primarily intended to be a standard course for experimentation. The AAAI Competition did not especially further experimentation, as that the competition judges collected no metric data. However, the AAAI Competition performed a valuable service by illustrating the potential conflict between science and exhibitions. The public viewing interfered with testing and validating aspects of AI in two different ways. Public viewing may also lead to a tendency towards "cuteness" at the expense of showing direct relevance to the USAR community.

Viewing versus Validation

The conflict between spectator viewing and validation is best seen by the following example. One of the USF vision agents identified large regions of heat using a FLIR, then fused that data with regions of motion, skin color, and distinctive color extracted by software agents operating on video data. If there was a sufficiently strong correlation, the operator interface began beeping to draw the operator's attention to the possibility of a survivor. (The RWI supplied user interface for the Urban requires almost full attention just to navigate, detection assistance is a practical necessity.)

Unfortunately, the test bed has Plexiglas panels to facilitate judge and spectator viewing. AAAI permitted spectators to ring the sections during the competition. Between the low height of walls and the Plexiglas, these spectators were visible and produced color, motion, and IR signatures even when the USF robots were facing interior walls due to views of exterior walls in other sections. As a result, USF had to turn off automatic victim notification through audio and rely strictly on color highlighting in the display windows.

A long-term solution is to insert cameras into the testbed to record, map, and time robot activity as well as broadcast the event to a remote audience. The competition chair stated that the audience should be allowed viewing access on the grounds that rescue workers would be visible in a real site. We note that at a "real site", access to the hot zone is strictly controlled and very few, certified technical rescue workers are permitted in the hot and warm zones. The rest must wait in the cold zone at least 250 feet from the hot zone (Casper,

Micire, & Murphy 2000). Also, at a real site, walls would have blocked views of people versus the half height panels.

Second, in order to record and broadcast the event, photographers and cameramen were permitted in the ring during the exhibitions and competition. During the exhibition, a cameraman repeatedly refused to move out of the robots' way. When the robot continued on, it almost collided with the video recorder.

Therefore, we recommend:

1. At least the Red Section should be fitted with walls and ceilings to block the view of non-testbed elements and the audience.
2. The test bed sections should be fitted with cameras and no one should be permitted in the test bed during timed events. If a robot dies (such as the USF Urban due to a faulty power supply or the Arkansas robots due to software failures), the robot should remain there until the session is complete.

Relevance to the USAR Community

In our opinion, the AAAI Competition missed several opportunities to show a clear relevance of the NIST test bed and robots to the USAR community. As discussed earlier, USAR professionals should be involved in setting the rules as well providing realistic scenarios. In general, any further competition venues, such as RoboCup Rescue, should actively discourage anything that might be construed as trivializing the domain. For example, Swarthmore costumed their robot as a Florence Nightingale style nurse, which rescue workers were likely to find offensive. Likewise, a handwritten "save me" sign was placed next to a surface victim.

The test bed may also miss relevance to the USAR field if it focuses only on benchmarking fully autonomous systems rather than on more practicable mixed-initiative (adjustable autonomy) systems. The Urban type of robot in a hardened form capable of operating in collapsed structures must be controlled off-board: they do not have sufficient onboard disk space to store vision and control routines. Therefore, the test bed should measure communications bandwidth, rate, and content in order to categorize the extent of a system's dependency on communications. Also, the test bed should include localized communications disrupters to simulate the effect of building rubble on communications systems.

Conclusions

Based on our five complete runs in the NIST test bed at AAAI and numerous informal publicity demonstrations, the USF team has had the most time running robots in the test bed. We conclude that the NIST test bed is an excellent halfway point between the laboratory and the real world. The test bed can be evolved to increasingly difficult situations. The initial design appears to have focused on providing navigational challenges, and it is hoped that future versions will add perceptual challenges.

Our recommendations fall into four categories. First, scoring or validation will be a critical aspect of the test bed. The AAAI competition did not implement a quantitative scoring system and thus provides no feedback on what are reasonable metrics. We recommend many metrics, but our guiding suggestion is to get knowledgeable representatives from the USAR community involved in setting up scenarios and metrics. In particular, we note that the victims should be distributed in accordance to FEMA statistics for surface, lightly trapped, void trapped, and entombed victims, and then points awarded accordingly. One major issue that arose from the USF team trying to reconstruct its rate of victim detection was that there needs to be an unambiguous method for signaling when a victim has been detected. Another aspect of scoring is to complement the proposed AAAI "black box" (external performance) metrics with a rigorous "white box"(software design and implementation) evaluation. Second, the test bed should be made more representative of collapsed buildings. We believe this can be done without sacrificing the motivation for the different sections. For example, all sections need to have void spaces representative of the three types discussed in the FEMA literature (lean-to, V, and pancake). The Yellow Section can still have a level, smooth ground plane but the perceptual challenges can be more realistic. Third, the test bed should resolve the inherent conflict between spectator viewing and validation. We believe this can be done by inserting cameras into the test sections as well as adding tarps and walls. Finally, we strongly urge the mobile robotics community to concentrate on making the NIST test bed and any competition venue which uses the test bed to be relevant to the USAR community. The community should resist the tendency to "be cute" and instead use the test bed as a means of rating mixed-initiative or adjustable autonomy systems that can be transferred to the field in the near future as well as the utility of fully autonomous systems.

Acknowledgments

Portions of the work reported in this paper were conducted as part of work under the DARPA BAA 9835 program and through a donation from SAIC. The authors would like to thank AAAI, the USF Department of Computer Science and Engineering, and the USF Engineering College for travel support, and Mark Powell and Brian Minten for their help in programming and maintaining the robots.

References

Casper, J.; Micire, M.; and Murphy, R. 2000. Issues in intelligent robots for search and rescue. In *SPIE Ground Vehicle Technology II Conference.*

Federal Emergency Management Agency. 1993. *Rescue Systems I.*

http://www.cs.swarthmore.edu/~meeden/aaai00/contest.html.

National Fire Protection Association. 1999. *Standard on Operations and Trainging for Technical Rescue Incidents.*

United States FIre Administration. 1996. *Technical Rescue Program Development Manual.*