Rescuing Interfaces: A Multi-Year Study of Human-Robot Interaction at the AAAI Robot Rescue Competition

Holly A. Yanco and Jill L. Drury

University of Massachusetts Lowell
Computer Science Department
One University Avenue, Olsen Hall
Lowell, MA 01854
holly@cs.uml.edu

The MITRE Corporation
Mail Stop B320
202 Burlington Road
Bedford, MA 01730
jldrury@mitre.org

Abstract

This paper presents results from three years of studying human-robot interaction in the context of the AAAI Robot Rescue Competition. We discuss our study methodology, the competitors' systems and performance, and suggest ways to improve human-robot interaction in the urban search and rescue as well as other remote robot operations.

1. Introduction

In 2000, the Robot Rescue Competition was added to the annual AAAI Robot Competition and Exhibition. In 2002, we started collaborating with the National Institute of Standards and Technology (NIST) to study human-robot interaction in the rescue competition; competitions in 2002, 2003 and 2004 were studied. In 2002, eight teams competed in the event,[1] with four teams making it to the final rounds. In 2003, six teams competed and four teams were in the finals. In 2004, seven teams competed and five teams were in the finals.

In the rescue competition, one or more robots are placed at the entrance to a simulated disaster zone and operated by a person at a remote station. The human-robot interaction in this task is interesting because urban search and rescue is a prime example of a class of safety-critical situations: situations in which a run-time error or failure could result in death, injury, loss of property, or environmental harm [Leveson 1986]. Safety-critical situations, which are usually also time-critical, provide one of the bigger challenges for robot designers due to the vital importance that robots perform exactly as intended and support humans in efficient and error-free operations.

The AAAI rescue competitions have provided fertile ground for studying human-robot interaction (HRI), allowing us to glean insights into how taking different HRI approaches may affect competition performance. Perhaps their greatest advantage for our research is that they have allowed us to see a number of different interface approaches being used to perform the same tasks under the same conditions.[2]

In this paper, we will discuss the conditions under which the competitions took place, followed by the methodology we used to study HRI during and after the competitions. We then present a discussion of the teams' designs, strategies, and performance over the three year study span. We present suggestions for designing robot systems and their interfaces to successfully complete the urban search and rescue task. Finally, we discuss what we can learn about HRI from these competitions.

---

[1] Two additional teams entered as demonstrations only in 2002.
[2] We also studied the Robot Rescue Competitions at RoboCup 2003 and RoboCup 2004. In this paper, we focus on the AAAI competitions.

2. Methodology

2.1 Competition Environment

In all three years, robots competed in the Reference Test Arenas for Autonomous Mobile Robots developed by the National Institute of Standards and Technology (NIST) [Jacoff et al, 2000; Jacoff et al, 2001]. The arena consists of three sections that vary in difficulty, from yellow (the easiest to navigate) to orange and red (the most difficult to traverse). The yellow section emulates an office environment that has been slightly damaged (e.g., there are fallen blinds and overturned tables and chairs). The orange section contains more challenging terrain such as variable floorings, a second story accessible by stairs or a ramp, and holes in the second story flooring. The red section is an unstructured environment containing a simulated pancake building collapse, piles of debris, unstable platforms to simulate a secondary collapse, and other hazardous obstacles such as rebar and loose cinder blocks. Figure 1 shows the NIST arena floor plan.

During each of the AAAI competitions, simulated victims (represented by mannequins) were scattered throughout the arena. Some mannequins were equipped with heating pads to radiate body heat, motors to create movement in the fingers and arms, and/or tape recorders to play recordings of people calling for help. In some cases, partial mannequins were used to simulate trapped victims (e.g., a mannequin torso only might be shown above ground level) or "entombed" victims (e.g., only a mannequin's arm showing coming out of a pile of debris). The difficulty of the search task was varied depending upon the number of victims in the open versus those that are trapped or entombed.

Figure 1. The NIST Reference Test Arena, used in the AAAI Robot Rescue Competitions. Shown is an overview photo of the arena from AAAI-2004. The orange area is in the foreground with the yellow area to the left and red area to the back right.

Between rounds, the victim locations were changed to prevent knowledge gained during earlier rounds from easing the search in later rounds. The walls of the arena were also modified to create new internal floor layouts from one round to another. Repeatedly reconfiguring the arena also prevented operators from having prior knowledge of the arena layout.

The number of victims changed from year to year. In 2002, there were a fixed number of victims in each of the three arena areas. In 2003 and 2004, the number of victims was reduced as the rounds progressed. The difficulty of the victim placement was also increased between rounds in 2003 and 2004, meaning that more surface victims were present in earlier rounds than later ones, while there would be more entombed victims in later rounds.

The arena size also changed from year to year, reflecting the space available in the convention center. In 2002, the total area was 2608 square feet (yellow: 1456 sq ft, orange: 672 sq ft, and red: 480 sq ft). In 2003, the total area was 1824 square feet (yellow: 672 sq ft, orange: 672 sq ft, and red: 480 sq ft). In 2004, the total area was 2240 square feet (yellow: 784 sq ft, orange: 784 sq ft, and red: 672 sq ft). In the discussions of team performance below, we account for the varying size of the arenas over the years.

The number of runs also changed over the years. In 2002, all teams had three runs in the arena. In 2003, teams started with three runs, then the top teams moved to two runs in the final rounds. In 2004, teams still qualified with three runs, but the top teams completed three runs in the final rounds.

While the finer points of the scoring algorithm evolved over the three years, in each case teams were awarded points based on the number of victims found and penalized for allowing robots to bump into obstacles or victims. The judges made a distinction between minor or major bumps of victims, and minor or major damage done to the arena environment. We use these three metrics to assess the performance of teams over the years with respect to one another.

One change to the rules that affected team strategies occurred between 2002 and 2003. In 2002, the teams received extra points for having different victims found by different robots. The rationale for this scoring was to encourage teams to field multiple robots that could spread out to search larger areas than a single robot could cover in the limited run time. However, teams that fielded multiple robots ended up operating them serially, sending one robot in to find a victim, then switching to a second robot to find the next victim. Since the rule did not bring about the desired result, it was dropped in the 2003 version of the rules. Below, we discuss the impact that this change had upon the strategies of the competitors.

2.2 Studying HRI During the Competitions

In all cases, teams voluntarily registered for the competitions. We asked each of them to participate in our study, but made it clear that study participation was not a requirement for competition participation. The majority of teams agreed to participate in all three years; only one team declined over the course of the study.

Prior to the competitions, we sent a questionnaire to teams participating in our study. Questions concerned the robot hardware being used, the type of data provided to the human operator, the level of robot autonomy, the maturity of the robot design, and whether the interface was based on a custom (bespoke) or commercial product.

Once the competitions began, we observed the operator of each team's robot(s) during their competition runs. We videotaped the interface screen(s) and, in 2002, the operators. We also videotaped the robots as they moved through the arena by having a camera operator follow the robots at a short distance.

Because our study could not impact the competition outcome, we were silent observers, not asking the operators to do anything differently during the competition runs. After each run, our observer performed a quick debriefing of the operator via a short post-run interview to obtain the operator's assessment of the robot's performance.

In addition, we were given the scoring materials from the competition judges that indicated where victims were found and penalties that were assessed. We also created maps by hand that showed the approximate paths that the robots took and marked the critical incidents that occurred during the runs. Critical incidents were defined as anomalous situations in which the operator or robot encounters a problem such as hitting an object or victim, or causing damage to the environment. These critical incidents were reflected in the major and minor penalties assessed by the judges.

2.3 Analysis After the Competitions

Our analysis techniques evolved over the three competitions. In 2002, we devised a detailed coding scheme to capture the number and duration of occurrences of various types of activities observed [Yanco, Drury, Scholtz 2004]. Our scheme consisted of a two-level hierarchy of codes inspired by the Natural Goals, Operators, Methods and Selection Rules (GOMS) Language (NGOMSL) analysis technique [Kieras 1988], with header codes capturing the high-level events and primitive codes capturing low-level activities. Every second of the competition runs was coded and we examined the percentage of time that operators spent navigating or monitoring navigation, finding victims, handling failures, or robot logistics (e.g., undocking small robots from a larger robot). We tailored Scholtz' [2002] evaluation guidelines and evaluated the robots' performance against those guidelines.

Although we had coded data in great detail, we chiefly dwelled on the times when things went wrong: the critical incidents. Thus, starting in 2003, we specifically focused on coding critical incident data, starting with the RoboCup 2003 study. We tied each critical incident to a loss of awareness of the robots' status, location, or activities; and used our fine-grained definition of HRI awareness [Drury, Scholtz, and Yanco 2003] to help understand and describe those incidents.

After studying the AAAI 2002 and RoboCup 2003 data, we conjectured that interfaces which emphasized video, sensor maps, geo-spatial maps, and non-occluded windows would tend to result in better competition performance. Accordingly, we performed an analysis of the three years of AAAI data (presented below) by examining competition performance in conjunction with characteristics of the teams' interaction design. Specifically, we examined the percentage of screen real estate devoted to the presentation of video, other sensor, and map data; whether windows were occluded, and the number of separate displays used. We juxtaposed this information with the number of victims found, the percentage of the arena covered, and the bumping penalties accrued.

2.4 Methodology for Scoring Team Performance

Since the scoring algorithm changed over the three competition years, we have developed a simplified scoring algorithm to normalize and compare performance across the three years. In all three years, the scoring algorithm utilized the number of victims found. The scoring scheme penalized teams for allowing robots to bump into obstacles or victims. The judges recorded a minor penalty for bumping into a victim or causing damage to the environment (subtracting .25 from the number of victims found) and a major penalty was scored for an event such as causing a pancake layer to collapse (subtracting 1).

Our scoring formula is as follows:

V = number of victims found
P = penalties (.25 for each minor penalty, 1 for each major penalty)

Performance Score = (V – P)

3. Competitors

In our data analysis, we only analyzed teams that made it to the finals, as the teams that did not make it often had hardware problems that prevented them from finding any victims or covering much of the arena area. Following is a description of the systems of the teams that made it to the finals in one or more years with a summary of the robot capabilities in the final subsection.

3.1 Team A: 2002

For AAAI 2002, Team A developed a heterogeneous team of five robots, one iRobot ATRV-Mini and four Sony AIBOs. All of the robots were teleoperated serially. The AIBOs were mounted on a rack at the back of the ATRV-Mini, and were undocked to use them independently and redocked after they were used if the operator wanted to continue to take them with the larger robot. Team A developed two custom user interfaces that ran on separate computers: one for the ATRV-Mini (top of figure 2) and another for the AIBOs (bottom of figure 2).

The user interface for the ATRV-Mini had multiple windows. In the upper left corner was a video image taken by the robot, updated once or twice each second. In the lower left corner was a map constructed by the robot using the SICK laser scanner and odometry. In the lower right corner, the raw laser scan information was presented as lines showing distance from the robot. The upper right corner had a window with eight radio buttons labeled 1 to 8 to allow the user to switch camera views. The operator drove the robot using keys on the keyboard to move forward, backward, right and left.

The user interface for the AIBOs had a window with the video image sent from the robot. The operator controlled the robots either using buttons on the GUI or by using the keyboard.

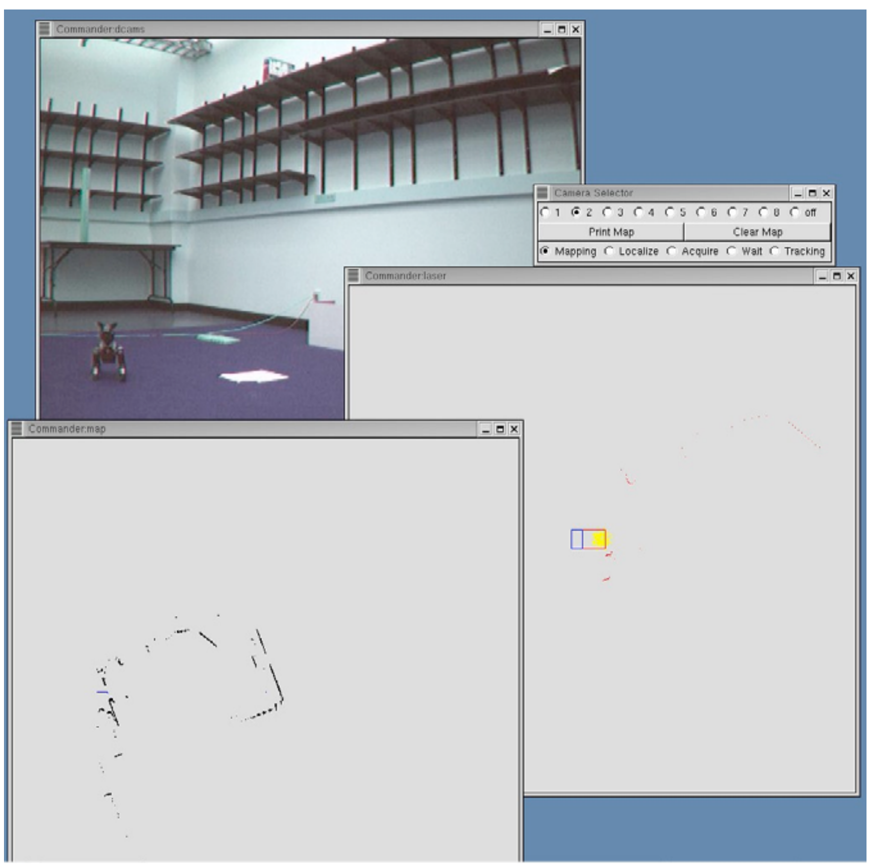

Figure 2. Team A's interface for the ATRV-Mini

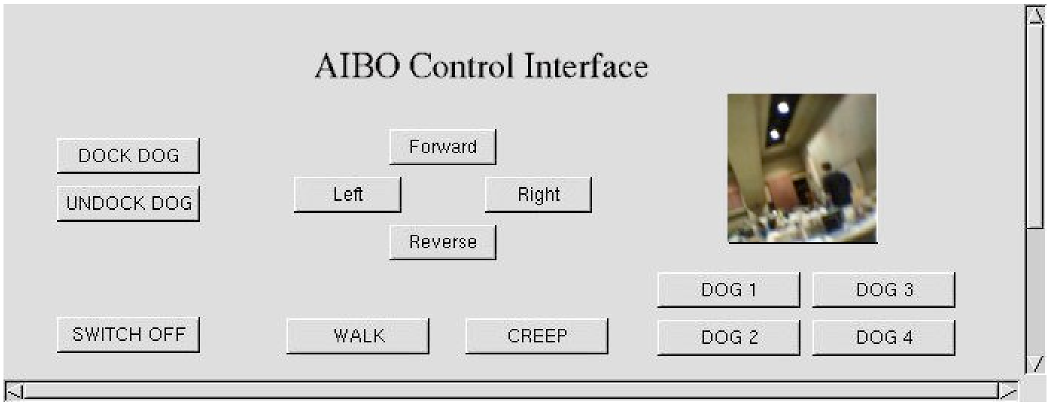

Figure 3. Team A's interface for the AIBOs.

3.2 Team B: 2002, 2003

Team B's robot was an iRobot ATRV-JR. In 2002 and 2003, Team B's custom user interface was displayed on a touch screen. The upper left corner of the interface contained the video feed from the robot. Tapping the sides of the window moved the camera left, right, up or down. Tapping the center of the window re-centered the camera. During the 2002 competition, the video window had not yet been finished, so the video was displayed on a separate monitor; however, the blank window was still tapped to move the camera (figure 3 shows the 2002 interface). The robot was equipped with two types of cameras that the operator could switch between: a color video camera and a thermal camera. By 2003, the video was incorporated into the main screen (figure 4 shows the 2003 interface). Pan, tilt and zoom indicators were also added in 2003; during the 2002 competition, the robot had been driven for several minutes with its camera off-center without the operator seeing an indication of this status on the interface.

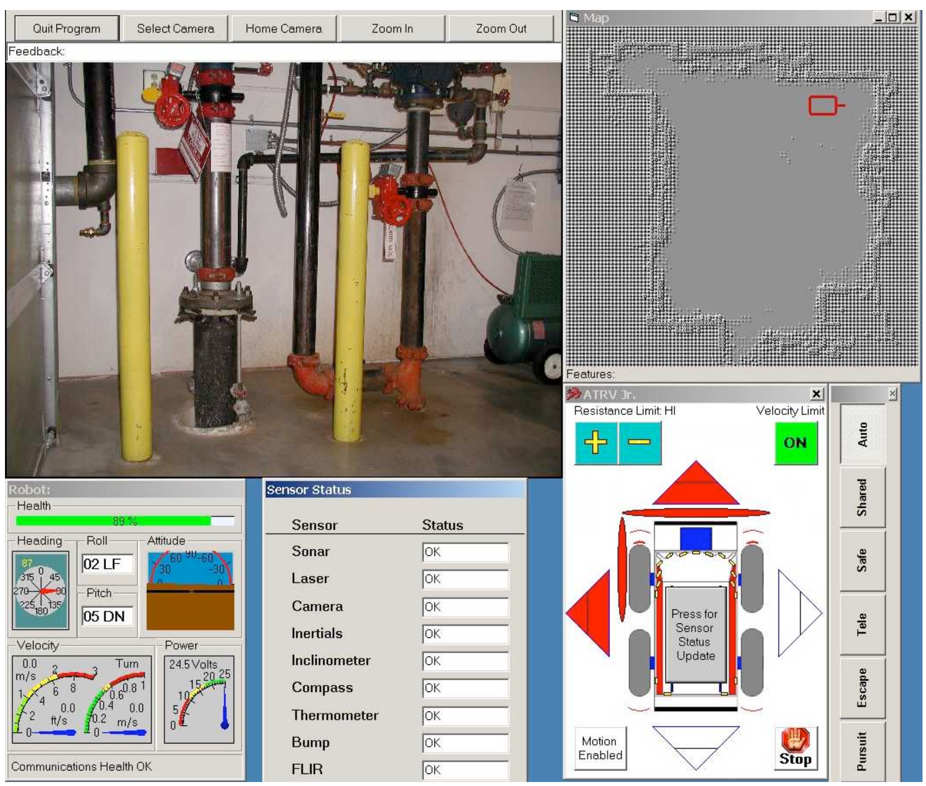

Figure 3. Team B's interface in 2002. At the competition, video was displayed on a separate monitor.

The lower left corner contained a window displaying sensor information such as battery level, heading, and tilt of the robot. In the lower right corner, a sensor map was displayed, showing filled red areas to indicate blocked directions. Space was left for a map in the upper right hand corner during the 2002 competition, as the software for building and displaying maps had not yet been created; in later years the map data was implemented and shown in the middle of the lower row of the display. In 2003, the upper right corner showed the status of the sensor modalities on the robot and allowed the operator to turn different sensor groups on or off.

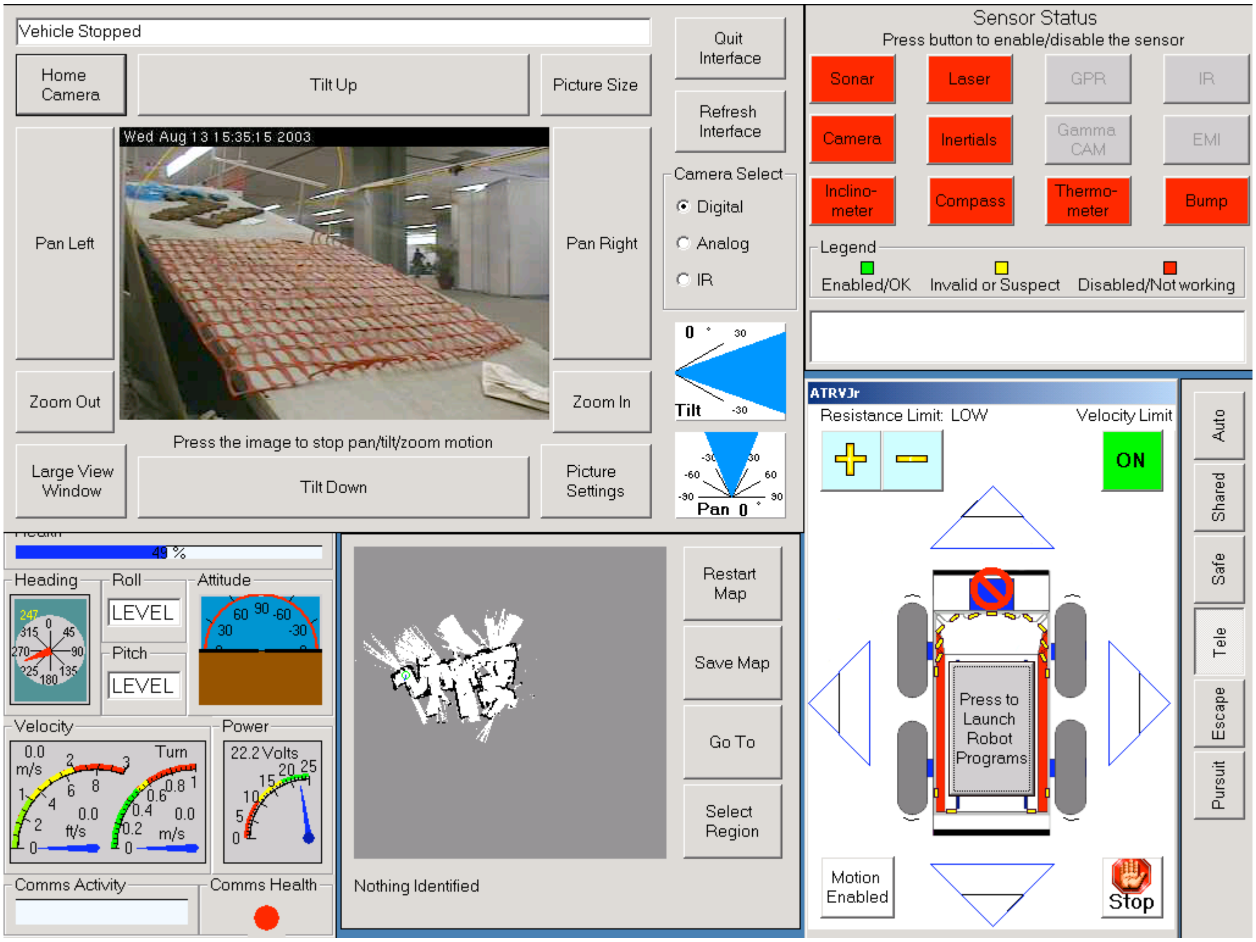

Figure 4. Team B's interface in 2003.

The robot was controlled through a combination of a joystick and the touch screen. To the right of the sensor map, there were six mode buttons: Auto (autonomous mode), Shared (shared mode, a semi-autonomous mode in which the operator can "guide" the robot in a direction but the robot does the navigation and obstacle avoidance), Safe (safe mode, in which the user controls the navigation of the robot, but the robot uses its sensors to prevent the user from driving into obstacles), Tele (teleoperation mode, in which the human controller is totally responsible for directing the robot), Escape (a mode not used in the competition) and Pursuit (also not used in the competition). Typically, the operator would click on one of the four mode buttons, then start to use the joystick to drive the robot. When the operator wished to take a closer look at something, perhaps a victim or an obstacle, he would stop driving and click on the video window to pan the camera. For victim identification, the operator would switch to the thermal camera for verification.

3.3 Team C: 2002, 2003, 2004

Team C developed their system on two identical RWI Magellan Pro robots, which are round robots equipped with a pan-tilt-zoom camera, sonar and infrared sensors. In 2002, the robots had a mixed level of autonomy: they could be fully teleoperated or the robots could provide obstacle avoidance while achieving a specified goal. The robots could run simultaneously, but were operated serially in practice. Waypoints were used to generate maps from the robot's current location to the starting point. The operator gave the robot relative coordinates to move towards in order to command a robot. The robot then autonomously moved to that location using reactive obstacle avoidance. The robot’s ability to carry out a command without assistance allowed for the perception that the operator moved both robots “at once,” even though he was controlling them serially. It was the operator’s trust in the robots’ autonomy that allowed this type of operation; the operator did not need to monitor the progress of one robot while commanding the other.

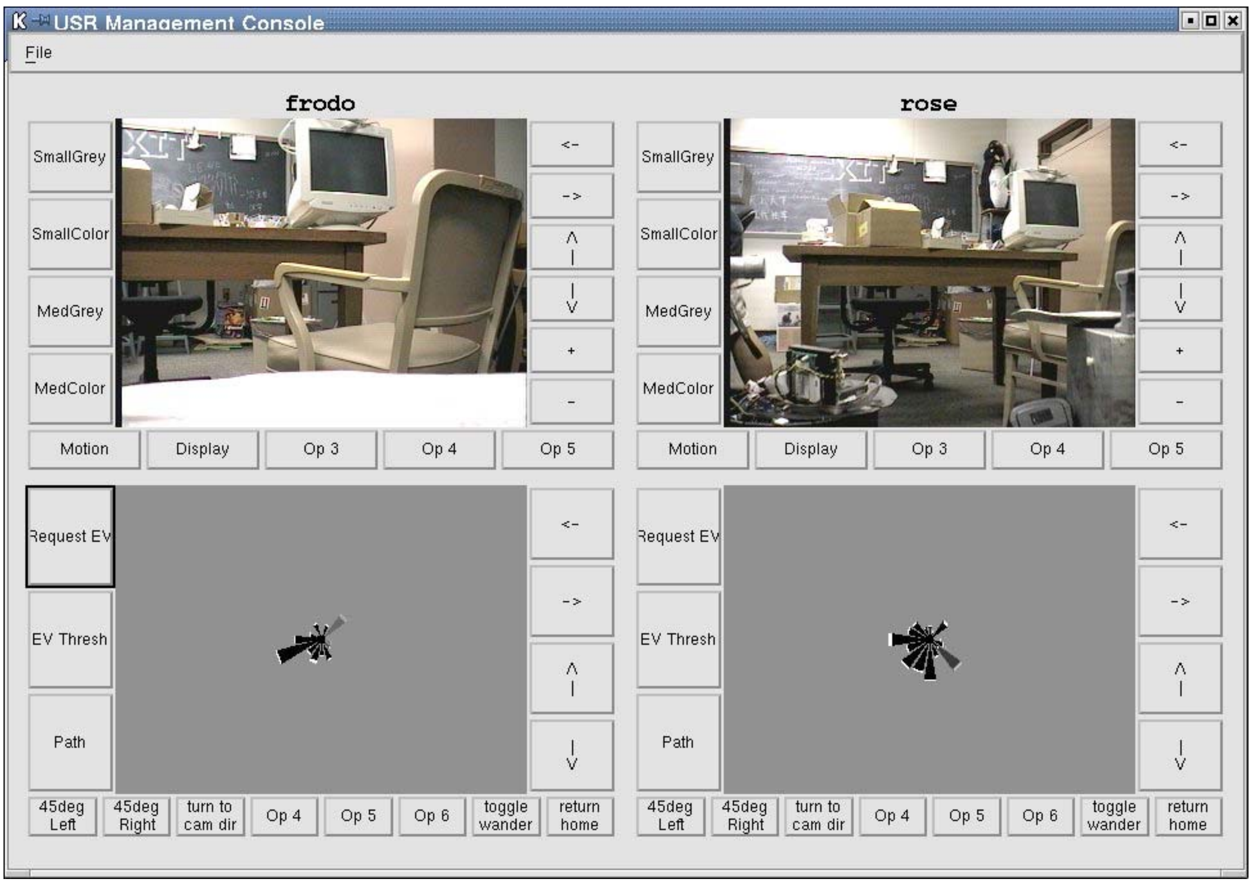

Figure 5. Team C’s interface in 2002.

In 2002, Team C used a custom interface that was developed for a “sophisticated user” (according to the developers); by 2004 they had changed the interface significantly. In 2002, the screen was split down the middle; each side was an interface to one of the two robots (see figure 5). The top window for each robot displayed a current video image from the robot and the bottom window displayed map information. In a large proportion of the 2002 competition, the team used a text-based user interface because of latency problems with the GUI. The alternative interface included 14 text windows and 4 small, graphic windows (half for each of the robots), with robot control achieved via keystrokes.

In 2003, the GUI was modified to include three windows per each of the two robots: one for the robot's sensor view, one for the video stream and one for the map. Robots could be controlled with the mouse or with key bindings. The keybindings were set to be similar to the standard keyboard controlled video game.

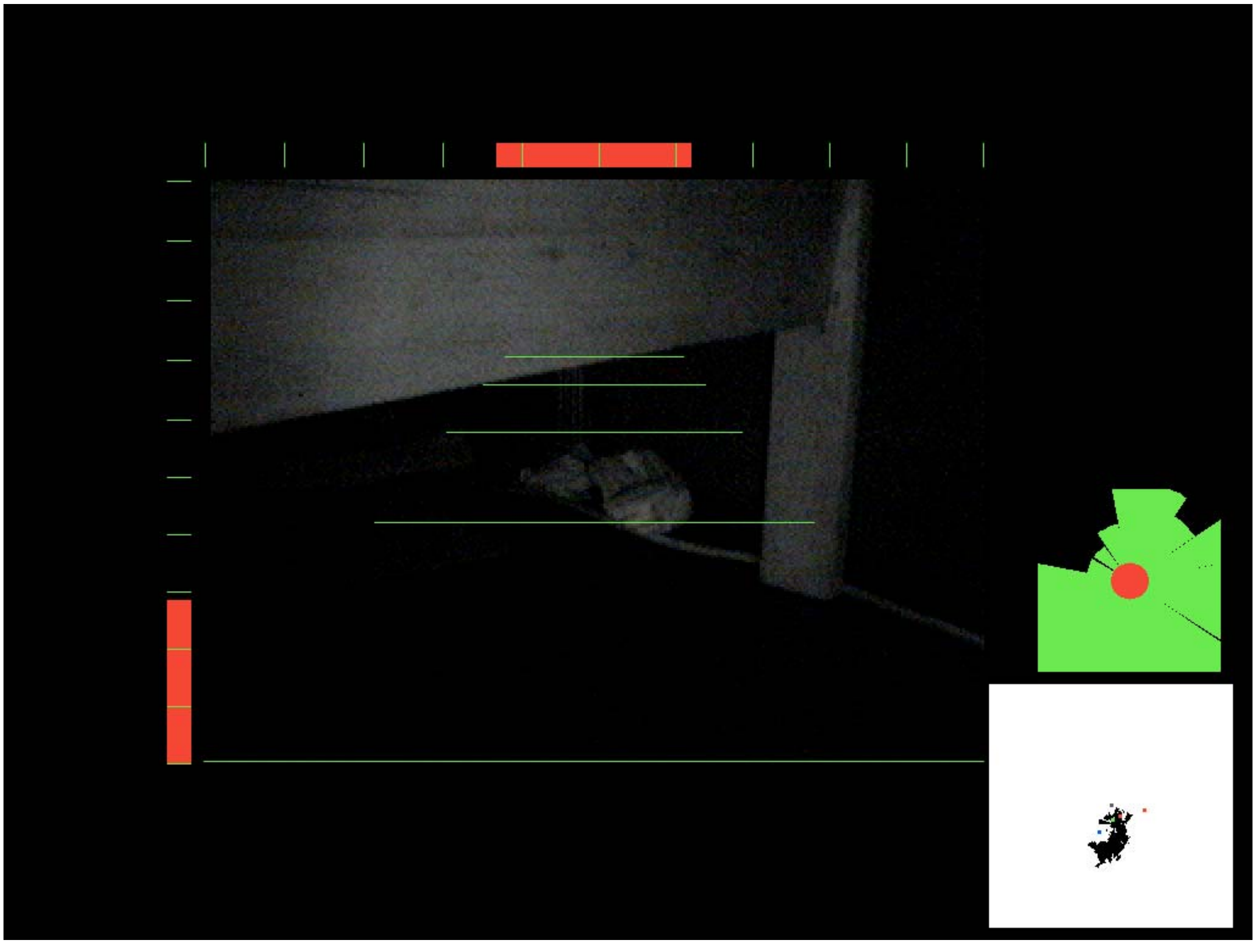

Figure 6. Team C's interface in 2004.

By 2004, the video game analogy was carried out even further (see figure 6). A switch was made to driving the robot using a joystick. The GUI was changed to make it full screen, with a single window. The size of the video window was increased and pan/tilt indicators were added to the top and left of the video window. The GUI also included the robot's current sensor data on the right and the map created by the robot just under that. Audio was also adding to the robot's sensing capabilities through the use of a wireless microphone.

With the move to the single window, the team lost the ability to see both of its robots' operations at the same time. However, with the change to the rules made in 2003, it was less important to focus on operating multiple robots and the second robot became a backup. The focus also shifted from more autonomous operation in 2002 to more teleoperation in 2003 and 2004.

3.4 Team D: 2002

Team D had custom built robots, one wheeled and one tracked, both with the same sensing and operating capabilities. The robots were teleoperated serially. A wireless modem was used to communicate between the user interface and the robots.

Team D developed a custom user interface on two screens. One monitor displayed the video feed from the robot that was currently being operated. The other monitor had a pre-entered map of the arena, on which the operator would place marks to represent the locations of victims that were found. The robots were driven with keyboard controls.

3.5 Team E: 2003, 2004

In 2003, Team E's entry consisted of four AIBO robots (dogs) and a blimp. The AIBOs were equipped with a camera and infrared sensors. The blimp was equipped with a camera. Robots were controlled using a mix of teleoperation and autonomy; the only autonomy was setting a hot spot to move to.

The interface consisted of four monitors for robot control and mapping purposes: two computers were used to control two AIBOs each, 1 used to control the blimp and the last one showed a shared map. Operators built the shared map using Unreal Tournament (shown in figure 7) based upon the information they received from the robot, marking verified victims as purple and potential victims as green. The barrels shown are obstacles or arena walls.

The team used three operators, one per two dogs and a third to control the blimp.

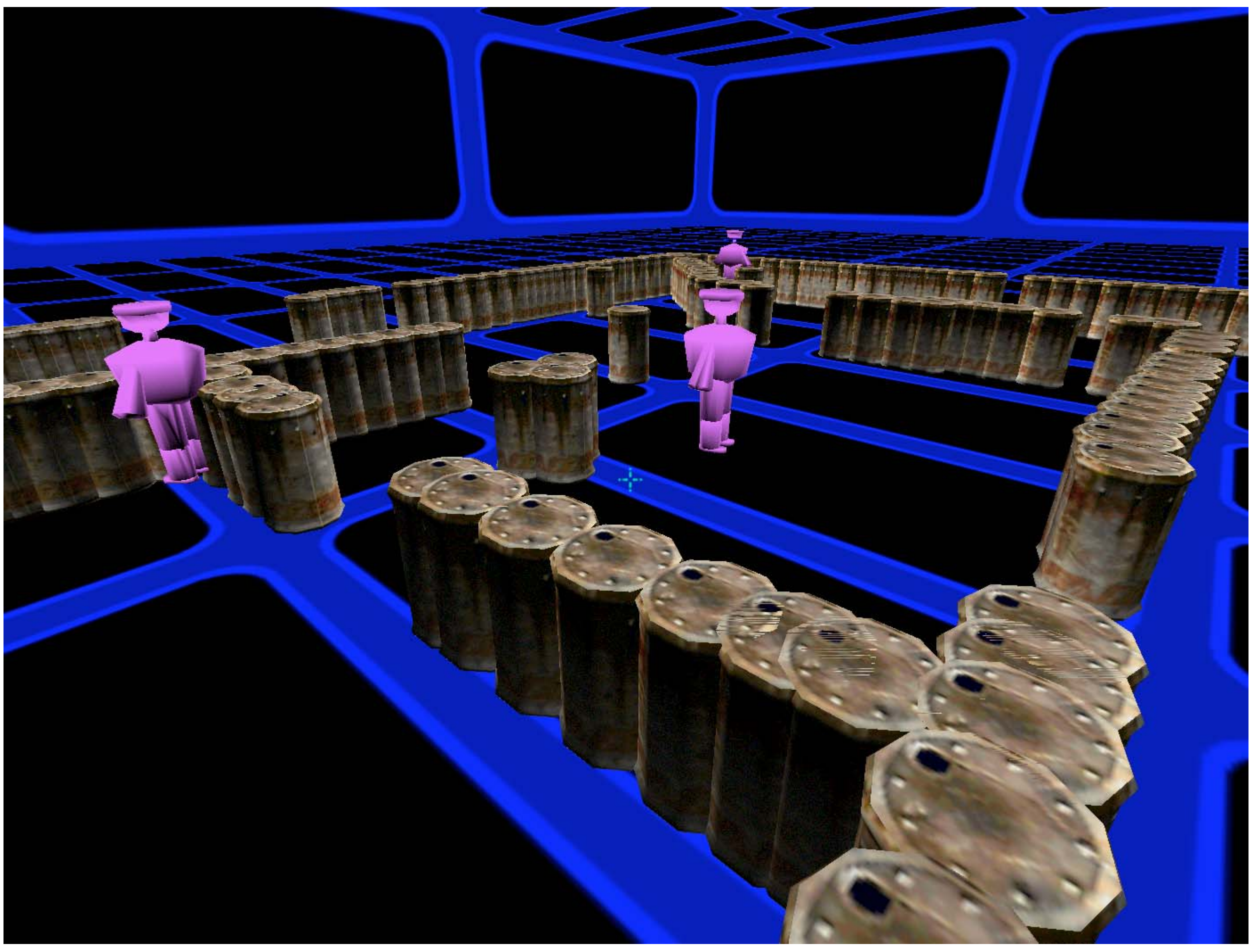

Figure 7. The shared map from Team E's interface in 2003.

In 2004, the team used an entirely different robot: an Inuktun VGTV Extreme with a single operator. The Inuktun robot is a tracked robot that has the ability to change its shape from a low-profile up to a higher-profile triangle. The robot has a tether to connect it to the operator control unit

(OCU) and power; for the rounds, a second person was required to act as a tether manager to prevent it from getting caught on obstacles in the arena. The robot was teleoperated using a joystick. It was equipped with a video camera and a forward-looking infrared (FLIR) camera.

The interface was displayed on two screens. The OCU screen displayed the video feed from the robot. The second screen was used to display a map of the environment, which was built by the operator throughout the run.

3.6 Team F: 2003[3]

Team F entered the 2003 competition with an ActivMedia Pioneer 2AT. The robot was equipped with a camera and sonar sensors. The robot was teleoperated via a joystick by a single operator. The operator created the map of the arena using the robot's sonar readings. The robot had a fixed camera.

The interface was on one monitor with a maximum of three windows on the interface. One window displayed the video from the robot, one window displayed the map built by the robot and the third was for control. The map from the interface is shown in figure 8.

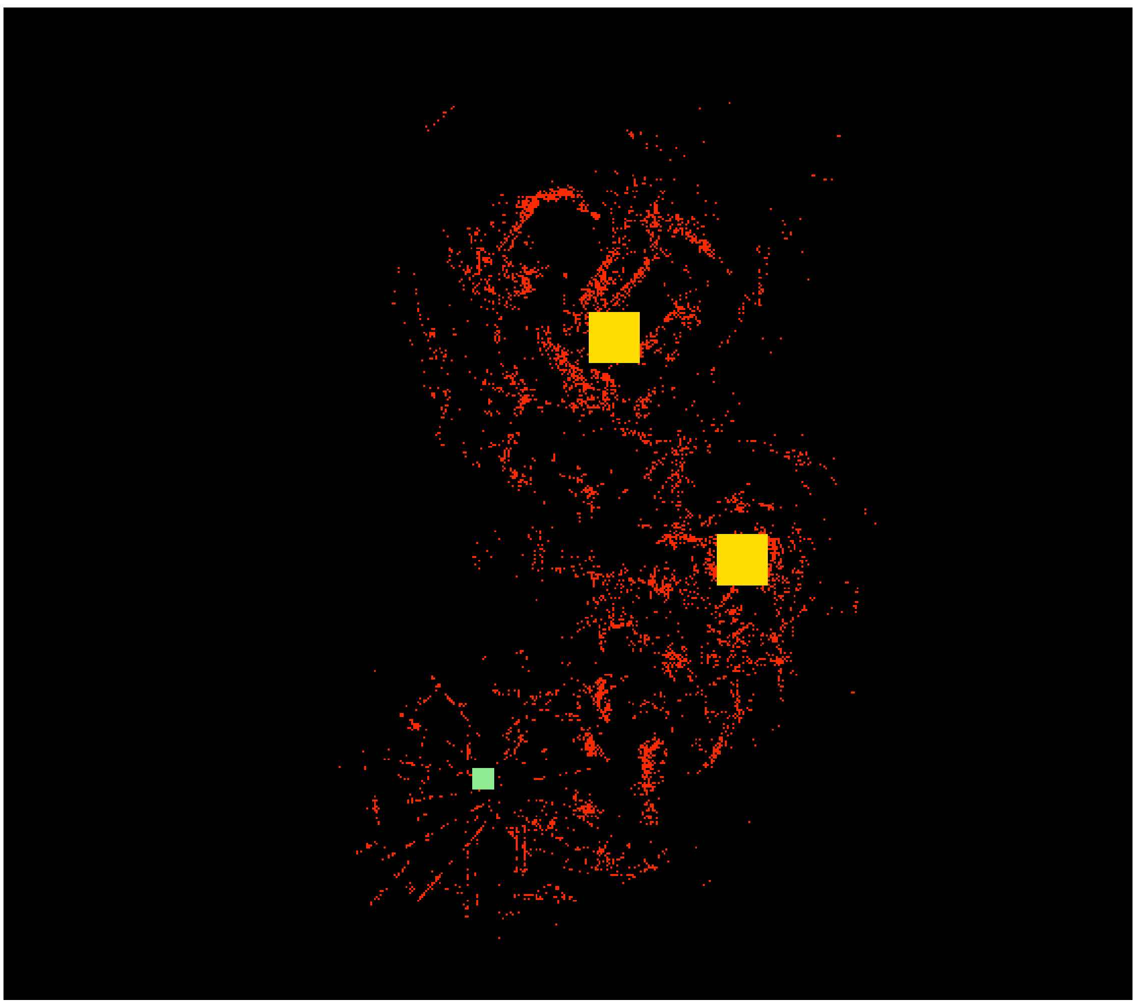

Figure 8. The map from Team F's interface in 2003.

[3] Team F entered in 2002, but did not make the finals, so we discuss 2003 only.

3.7 Team G: 2004[4]

Team G had a robot team comprised of three ActivMedia Pioneer 3-DX robots. Each robot was equipped with a camera, sonar sensors, infrared sensors used as virtual bumpers, laser ranging, a scanning pyro sensor to detect body heat, contact bumpers, and wheel encoders. The robots were controlled using a mix of autonomy and teleoperation.

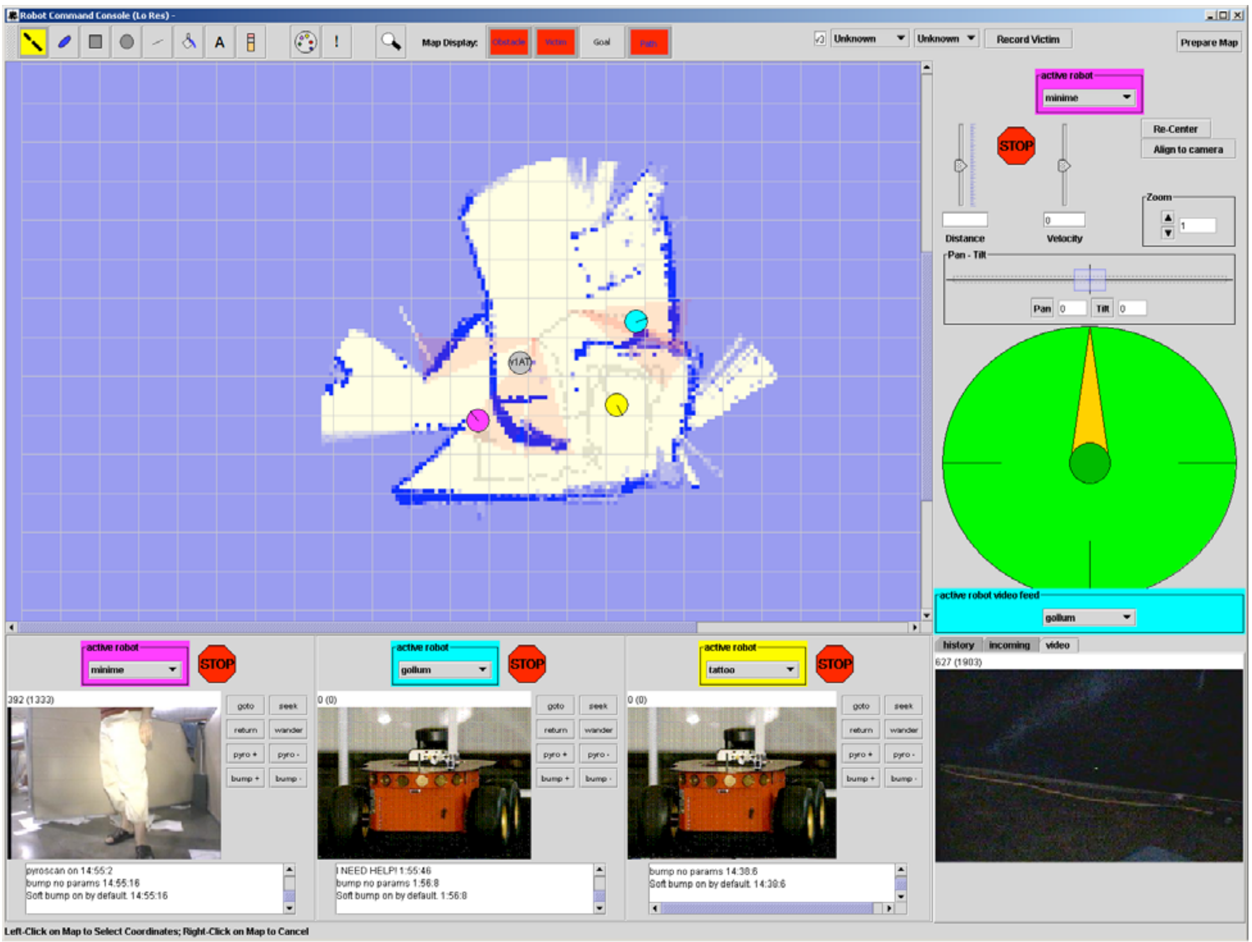

Figure 9. Team G's interface in 2004.

The interface, shown in figure 9 is a single window on the full screen that consists of several panes. The left side of the interface has the map of the environment, built by contributions from all of the robots. Under the map is a control pane for each of the three robots; this pane contains commands for the robot as well as the robot's video stream. At the top right are velocity controls with driving controls directly below it. Beneath that is the area to select the active robot. At the bottom is the option area, which can display messages to the robot or from the robot or can be used as an additional video area (it is slightly larger than the other video windows).

[4] Team G entered in 2002 as a demonstration only, so that year is not discussed.

3.8 Team H: 2004

Team H used custom-built robots for their entry. Two of the robots were modular snakes that could be outfitted with legs or wheels. The snake had video, audio, thermal sensing and a CO2 sensor and ran tethered for longer run times. The other robot was a modified RC car that deployed sensor motes in the arena as it drove forward (the motes came off the back of the car).

The interface, shown in figure 10, consisted of three monitors, two showing the video from the snakes and the other showing the video from the RC car. The snakes were driven using joysticks and the RC car had a two stick drive mechanism. The set up was spread across two tables. It was originally designed for two operators (one operating one snake and the other operating the other snake and the RC car), but the team decided to use a single operator to remove the two operator penalty (the score was divided by the number of operators that a team had). This change required the single operator to get out of his seat and move to the other station to operate the other robot(s).

Figure 10. Team H's interface set up in 2004. To the left is the station for operating one of the snakes. The top monitor shows the video from the snake and the two computers are used for control. To the right is the projected video from the other snake, with the RC car's video shown on the computer screen below that. To the right of the computer is the joystick for driving the second snake and the controller for the RC car is to the right of the joystick. The operator would move back and forth between the two stations.

3.9 Team J: 2004[5]

Team J had two custom platforms. The first had differential drive using two modified RC servos and used a PDA as its processor. The second robot was a modified RC car with a 533 MHz mini-ITX. Both robots were fully autonomous, using computer vision as the only sensor.

The interface for the modified RC car is shown in figure 11. It displayed a pair of stereo images from the robot on the top of the screen, side by side. The left bottom corner showed the pan of the camera and the right bottom showed the map created by the robot.

3.10 Summary Comparison of Robots

Most teams entered commercially built robots into the competition. The exceptions were Team D in 2002 (two custom-built robots, one with wheels and one with treads), Team H in 2004 (custom-built snake robots and a modified RC car) and Team J in 2004 (modified RC car). The commercially built robots were wheeled (ATRV, Magellan and Pioneer models) or treaded (Inuktun VGTV). Some of the wheeled platforms had all-terrain capability (ATRVs, Pioneer ATs), but the

[4] Team J also entered in 2002 and 2003, but did not make the finals those years, so we discuss 2004 only.

Figure 11. Team J's interface in 2004.

robots' sizes prevented them from effectively navigating the pancake layer structure of the red arena; the weight of the robots would collapse the pancake layer. The Inuktun was light enough to traverse the pancake levels, but its tether had a tendency to get caught on things in the arenas.

Of the platforms used, only Team E's 2004 entry and Team D's 2002 entry had treads. In general, the robots used by the competitors did not have the all-terrain capability to enter the red arena, resulting in most of the arena coverage occurring in the yellow arena, followed by the orange arena. In general, competitors in the AAAI competition have not focused on the mobility of robots, but have instead concentrated on the control algorithms and interface designs. The RoboCup competitions have seen a much higher percentage of entered robots that can climb over obstacles.

Most of the robot systems were teleoperated during the runs. In fact, Team C had much more autonomy in 2002 than the system did in 2003 and 2004. Most teams found it more expedient to teleoperate robots through the environment and there was nothing in the rules to encourage autonomy over teleoperation. One possible change to the rules to encourage more autonomy could be an imposed period of "radio drop out" where the operator would not be able to send commands to the robot. An autonomous robot could run during this time, but a teleoperated robot would not; a rule like this would allow autonomous robots to have more time to navigate the arena to search for victims.

Not all robots were teleoperated. Team J's 2004 entry was completely autonomous; it was started at the door of the arena and ran on its own for the full time. Team B had several autonomy modes in their system and did use a mix of these modes. Teams C, E (2003) and H had the ability to set goal points for the system to travel to in an autonomous mode. However, most competitors found full teleoperation to be the preferred control method for the competition task.

It is also possible that teleoperation is preferable because it requires the operator to participate in the full operation of the robot. Since the runs were short (20 minutes maximum), the operator could build his or her own mental model of the environment, reducing the requirement on a system generated map. We have found in usability studies that users lose situation awareness after extended periods of autonomy [Yanco and Drury 2004].

4. Results and Discussion

We first examine the scoring metrics used in the competitions that remained constant over the three years: the number of victims found, the number of minor penalties assessed and the number of major penalties assessed. In different years, scoring also depended upon the quality of the map produced, the number of robots used during the run that found unique victims, which of the three areas (yellow, orange or red) the victims were found in, and the number of operators controlling the robots. Since these measures were not the same in all three years, we have eliminated them for the purposes of this analysis. Additionally, in 2004, there was a required start between the orange and red arenas, ensuring that teams had to enter one of these two arenas that year.

Table 1. Scoring metrics for the teams in all three years.

	Run	Victims Found	Minor Penalties	Major Penalties	Victims – .25*Minor – 1*Major	Competition Rank or Award (if any)
Team A: 2002	1	4.0	8.0	0.0	2.0	
	2	4.0	5.0	1.0	1.8	3rd
	3	3.0	6.0	0.0	1.5	(2002)
	Average	**3.7**	**6.3**	**0.3**	**1.8**	
Team B: 2002	1	3.0	5.0	0.0	1.8	
	2	3.0	3.0	6.0	-3.8	Technical
	3	0.0	1.0	3.0	-3.3	Award (2002)
	Average	**2.0**	**3.0**	**3.0**	**-1.8**	
Team B: 2003	1	2.0	0.0	0.0	2.0	
	2	3.0	11.0	0.0	0.3	1st
	3	5.0	1.0	0.0	4.8	(2003)
	4	4.0	1.0	1.0	2.8	
	5	3.0	3.0	1.0	1.3	
	Average	**3.4**	**3.2**	**0.4**	**2.2**	
Team C: 2002	1	3.0	0.0	0.0	3.0	
	2	1.0	3.0	1.0	-0.8	2nd
	3	4.0	3.0	0.0	3.3	(2002)
	Average	**2.7**	**2.0**	**0.3**	**1.8**	
Team C: 2003	1	4.0	0.0	0.0	4.0	
	2	4.0	0.0	0.0	4.0	2nd
	3	0.0	0.0	0.0	0.0	(2003)
	4	1.0	1.0	1.0	-0.3	
	5	3.0	0.0	0.0	3.0	
	Average	**2.4**	**0.2**	**0.2**	**2.2**	
Team C: 2004	1	3.0	1.0	0.0	2.8	
	2	4.0	0.0	0.0	4.0	1st
	3	3.0	0.0	0.0	3.0	(2004)
	4	1.0	0.0	0.0	1.0	
	5	4.0	0.0	0.0	4.0	
	6	3.0	0.0	0.0	3.0	
	Average	**3.0**	**0.2**	**0.0**	**3.0**	
Team D: 2002	1	6.0	9.0	0.0	3.8	
	2	7.0	17.0	0.0	2.8	1st
	3	3.0	4.0	0.0	2.0	(2002)
	Average	**5.3**	**10.0**	**0.0**	**2.8**	
Team E: 2003	1	1.0	0.0	0.0	1.0	
	2	2.0	0.0	0.0	2.0	Technical
	3	3.0	0.0	0.0	3.0	Award
	4	3.0	0.0	0.0	3.0	(2003)
	5	4.0	0.0	0.0	4.0	
	Average	**2.6**	**0.0**	**0.0**	**2.6**	

Team E: 2004	1	2.0	0.0	2.0	2.0	
	2	2.0	4.0	1.0	1.0	
	3	3.0	0.0	1.0	1.0	
	4	2.0	0.0	2.0	2.0	
	5	2.0	2.0	1.5	1.5	
	6	2.0	0.0	2.0	2.0	
	Average	**2.2**	**1.0**	**1.6**	**1.6**	
Team F: 2003	1	0.0	0.0	0.0	0.0	
	2	3.0	1.0	0.0	2.8	
	3	3.0	3.0	3.0	-0.8	
	4	3.0	2.0	1.0	1.5	
	5	0.0	2.0	1.0	-1.5	
	Average	**1.8**	**1.6**	**1.0**	**0.4**	
Team G: 2004	1	1.0	0.0	0.0	1.0	
	2	0.0	1.0	0.0	-0.3	2nd
	3	1.0	0.0	0.0	1.0	(2004)
	4	1.0	0.0	0.0	1.0	
	5	1.0	1.0	0.0	0.8	
	6	0.0	0.0	0.0	0.0	
	Average	**0.7**	**0.3**	**0.0**	**0.6**	
Team H: 2004	1	1.0	4.0	0.0	0.0	
	2	4.0	2.0	0.0	3.5	3rd
	3	4.0	5.0	1.0	1.8	(2004)
	4	1.0	0.0	0.0	1.0	
	5	1.0	1.0	0.0	0.8	
	6	3.0	0.0	0.0	3.0	
	Average	**2.3**	**2.0**	**0.2**	**1.7**	
Team J: 2004	1	0.0	0.0	0.0	0.0	
	2	1.0	1.0	0.0	0.8	
	3	2.0	0.0	0.0	2.0	
	4	2.0	0.0	0.0	2.0	
	5	0.0	0.0	0.0	0.0	
	6	3.0	0.0	0.0	3.0	
	Average	**1.3**	**0.2**	**0.0**	**1.3**	

We have observed that returning teams have improved their performance using our scoring metric of the number of victims found minus .25 for each minor penalty and 1 for each major penalty. Three of the teams (B, C, and E) were in the finals for more than one of the three years.

Team B went from an average score of –1.8 in 2002 to an average score of 2.2 in 2003. While the layout of their interface was largely the same in 2003 as it had been in 2002, the mapping software was completed and a robot-generated map was displayed on the screen. The interface also moved from two monitors in 2002 (the video could not yet be displayed on the interface) to a single screen in 2003 when the video was displayed on the interface.

Team C had an average score of 1.8 in 2002, 2.2 in 2003 and 3.0 in 2004. Over the years, their interfaces moved towards greater fusion in the display. In 2002, most runs were completed using a text-based interface with 14 windows. In 2003, the interface had 6 graphical windows and keyboard-based control. In 2004, the interface was in a single window on the interface that had three panes (video, sensor map and world map) and the robot was controlled using the joystick.

Much of Team C's improvement came from a reduction in the number of penalties accrued rather than an increased number of victims found. Team B also saw a reduction in assessed penalties, but still had more than Team C. The robots' morphologies likely contribute to this difference. Team C's robots are round and smaller than Team B's rectangular robot; being able to turn in place reduces the possibility for hitting things on the side of the robot. Team B did find more victims in 2003 than 2002. Team E found fewer victims and incurred more penalties in 2004 than they did in 2003; as pointed out previously, they used a different robot system in each of the years.

While we observed improvements in Teams B and C over the competition years, we did not see the same improvement in Team E, who had an average score of 2.6 in 2003 and 1.6 in 2004. However, the team's entry was completely different these two years, changing both the type of robots used and their interfaces. As such, these two entries do not show any iterative improvements; they could be considered two separate entries, although entered by the same school.

4.1 Navigating the Arena

Table 2 shows the amount of the arena that each team covered in each of the runs. Since the arena sizes varied from year to year (as described above in section 2.1), the table gives total square footage covered. It also gives the percentages of each arena area (yellow, orange and red) covered. During the competitions, a bonus was given for victims found in the areas that were more difficult to navigate; a victim in the red area was worth more than a victim in the orange area, which in turn was worth more than a victim in the yellow area.

Table 2. Arena coverage for the teams in all three years.

	Run	**Sq Ft Covered**	**% of Each Arena Covered**		
			Yellow	**Orange**	**Red**
Team A: 2002	1	728	50	0	0
	2	509.6	35	0	0
	3	509.6	35	0	0
	Average	**582.4**			
Team B: 2002	1	324.8	20	5	0
	2	537.6	30	15	0
	3	576.8	35	10	0
	Average	**479.7**			
Team B: 2003	1	201.6	30	0	0
	2	268.8	25	15	0
	3	302.4	5	40	0
	4	336	50	0	0
	5	302.4	0	45	0
	Average	**282.24**			
Team C: 2002	1	436.8	30	0	0
	2	291.2	20	0	0
	3	509.6	35	0	0
	Average	**412.5**			
Team C: 2003	1	403.2	60	0	0
	2	403.2	60	0	0
	3	57.6	0	5	5
	4	336	50	0	0
	5	403.2	60	0	0
	Average	**320.64**			
Team C: 2004	1	627.2	80	0	0
	2	470.4	35	25	0
	3	470.4	60	0	0
	4	666.4	85	0	0
	5	431.2	30	25	0
	6	509.6	40	25	0
	Average	**529.2**			

Team D: 2002	1	1164.8	80[6]	0	0
	2	1237.6	85	0	0
	3	309.6	15	10	5
	Average	**904**			
Team E: 2003	1	134.4	20	0	0
	2	201.6	0	30	0
	3	268.8	40	0	0
	4	336	50	0	0
	5	436.8	65	0	0
	Average	**275.5**			
Team E: 2004	1	313.6	40	0	0
	2	392	0	50	0
	3	548.8	40	30	0
	4	268.8	0	0	40
	5	470.4	20	40	0
	6	100.8	0	0	15
	Average	**349.1**			
Team F: 2003	1	–[7]	–	–	–
	2	235.2	35	0	0
	3	268.8	40	0	0
	4	336	50	0	0
	5	235.2	0	35	0
	Average	**215.0**			
Team G: 2004	1	235.2	30	0	0
	2	196	0	25	0
	3	78.4	10	0	0
	4	156.8	20	0	0
	5	156.8	0	20	0
	6	235.2	30	0	0
	Average	**176.4**			
Team H: 2004	1	313.6	40	0	0
	2	–[8]	–	–	–
	3	313.6	0	40	0
	4	67.2	0	0	10
	5	196	0	25	0
	6	336	0	0	50
	Average	**245.3**			
Team J: 2004	1	235.2	0	30	0
	2	196	25	0	0
	3	313.6	40	0	0
	4	313.6	40	0	0
	5	134.4	0	0	20
	6	431.2	40	15	0
	Average	**270.7**			

We do not see a steady increase in the amount of arena covered by repeating teams or by new teams in later years. In each successive year, victims became harder to locate (fewer of them as well as more hidden), which may account for less area being covered. Additionally, after 2002, teams were required to identify victims with as many modalities as possible, showing the judge at the operator table the different markers. This change may have slowed down the progress in the arena. However, it should be noted that while repeating teams did not cover more area in later years, they did score higher, which means that they found more victims and had fewer penalties.

[6] In this round, the judges penalized the team for looking at the arena in advance, probably accounting for their large coverage of the yellow arena.
[7] The coverage map was missing for this run.
[8] The coverage map was missing for this run.

There is not a strong correlation between the amount of area covered and the outcome of the competition. Some competitors who covered more area do not end up finding more victims, probably due to the non-uniform distribution of victims. Additionally, when traveling quickly enough to cover more area, there is the potential to miss victims, particularly well-hidden ones.

4.2 Interface Designs

We have analyzed the design of the interfaces used to control the robots and present the data in table 3. Of primary importance are the displays of the video, the map and any other sensor data, as these provide information that the robot's operator needs to complete the task. We have measured the percentage of the interface devoted to these three areas. Other areas of the interface are devoted to controls (option selections, simulated joysticks and the like) or are unused real estate on the screen.

We have also examined whether the interface has window occlusion, where some of the interface windows block others. Window occlusion can bury important information beneath another window, causing it to be missed or requiring that the operator bring the proper window to the foreground when it is needed. This switching takes time away from the run, resulting in less area covered and fewer victims found.

Finally, we look at the number of displays used by a team. Since operators tend to focus on the video window, any information placed on another monitor is unlikely to be checked frequently. Additionally, requiring a change of focus to another monitor can reduce task performance.

Table 3. Interface characteristics of the competitors.

	% Video	**% Map**	**% Sensor Map**	**Window Occlusion?**	**Occlusion Magnitude**	**# Displays**
Team A: 2002	19.2% (Mini), 19% (AIBO)	23.3% (Mini)	0	Possible, but only if debug window needed	Low (Mini), None (AIBO)	2
Team B: 2002	5.2% of display #2, 100% of display #3	38.7% of display #1, but not functioning during competition	19% of display #1	No	None	3
Team B: 2003	9.6%	8.6%	17.5%	No	None	1
Team C: 2002	21% (total: two windows, one for each robot)	30.2% (total: two map windows, one for each robot)	0	Yes	High	1
Team C: 2003	5.8% (Small, one window); 23.2% (large, one window) - can have 2 video windows showing	24.7% (one map window - can have two)	6.5% (one window - can have two) 3.30%	Yes	High	1

Team C: 2004	47%	4.3%	3.3%	No, but a mode switch	None	1
Team D: 2002	50% or 100% of display #2 (depending on mode)	63.3% of display #1 for pre-entered map	0	No	None	2
Team E: 2003	AIBO control: 4% each, 8% total	AIBO control: 44.8%	0	No	None	4 (2 with 2 AIBOs each, 1 for blimp, 1 for shared map)
Team E: 2004	100% of OCU monitor	39.6% of computer monitor	0	No	None	2
Team F: 2003	3.6%	7%	0	No	None	1
Team G: 2004	3.4% each, up to 3 total, 1 per active robot	47%	0	No	None	1
Team H: 2004	100% of two monitors (snake video)	40% of one monitor (Not working during competition)	0	No	None	5
Team J: 2004	34.4% (for both stereo image)	24.3%	0	Some, when starting windows	Not during run	1

Operators relied primarily on video for victim identification. Other sensors (e.g., thermal or audio) were only used after the initial identification to confirm the presence of a victim by using another sensing modality. The reliance on video places a constant demand on the operators to watch the video for indications of victims.

As such, the size of the video window should be very important when looking at the performance of a system. Looking at Team C, the only three year competitor, we see that their best performance occurred in 2004, when their video window was the largest. Team B also had their best year (2003) with the largest video window. However, there is not a strong correlation between the video window size and any individual team's performance in a given year. Team C also had their best performance when they eliminated window occlusion in 2004, which had been present in both their 2002 and 2003 interfaces.

However, the number of monitors does seem to impact performance. All teams who ranked first or second in the competition, except the first place team in 2002, used a single monitor for their interface. Even the first place team in 2002 (Team D) spent most of their control time looking at a single monitor; their second monitor was used only to mark a victim when one was found. Other teams with two or more monitors (Team A 2002, Team B 2002, Team E 2003, Team E 2004 and Team H 2004) finished in third, received a technical award or didn't place.

There are two teams with a single monitor who didn't rank (Team F 2003 and Team J 2004). An explanation for Team F's performance may be their very small video window (3.6% of screen area). Team J's performance was not dependent upon the number of monitors or video size since their robots ran fully autonomously. Without an interface, their result would have been the same; the sole purpose of the interface was to convey information back to someone watching the robot.

4.3 Use of Multiple Robots

In 2002, the rules favored teams with multiple robots. The scoring algorithm gave a bonus for the number of robots that found unique victims; if two robots found the same victim and no other unique victims, the bonus would not be given. The purpose of this scoring rule was to encourage teams to create entries with some level of autonomy, so that multiple robots could be searching the arena at the same time. Additionally, in a USAR situation, there is a benefit to finding as many victims as possible as quickly as possible, in which case multiple robots could be helpful.

However, the scoring rule did not contribute in the way that was expected. Instead, we observed that teams with multiple robots would control the robots serially. In 2002, three of the teams had more than one robot (Teams A, C and D). Teams A and D teleoperated their robots using a single operator (so as not to incur the penalty for having more than one operator), meaning that they would move one robot, then change to control another robot. Team C did have the capability of setting a goal point for one robot, then would set the goal for the other robot while the first was moving. In 2002, the three multi-robot teams ranked in the top three places of the competiton.

Since the rule did not seem to encourage autonomy, the scoring bonus for multiple robots was removed. This did not discourage teams from entering multiple robots (2 of the 4 finalists in 2003 had more than one robot and 4 of the 5 finalists in 2004 had multiple robots). However, we did observe that the robots were used differently. Team C began to use their second robot as a backup, sending it in only if the first robot had problems.

Team G used three robots to build a shared map of the space. The operator controlled the robots individually, but could set waypoints in addition to teleoperating the robots. We observed that the robots were deployed into the arena in a manner that allowed a trailing robot to view the robot in front of it.

Team H explicitly used one of their other robots to view a robot that was having difficultly moving. In order to assess the problem, the operator left the station of the malfunctioning robot and turned to the station for the other. He maneuvered the other robot into a position where he could view the stuck robot and determine how the robot was caught so that it could be freed.

5. Designing for Effective HRI

With experience, competitors have moved to a single interface with a larger percentage of the screen dedicated to video. However, there were new teams that joined the competition each year and returning teams that came the next year with completely different entries, so it is difficult to assess trends across all of the interfaces. There were trends in the two repeating teams that kept the same hardware between years. Our observations of the competitions and the analysis above have shown that there are design guidelines that can be applied to USAR situations:

- *Use a single monitor for the interface*. Requiring the user to switch between multiple monitors seems to degrade performance.
- *Larger video windows assist in the success of the task, or, at least, very small video windows seem to hinder performance*. When teleoperating a robot, operators rely on the video to determine the best way to navigate in the environment. Additionally, since most of the robots used unprocessed video as the primary means for locating victims, the operators need to focus on the video window to look for victims.
- *Window occlusion hinders operation*. The operator is slowed down by needed to find the correct window and switch to it.
- *When multiple robots are available, use one to view another*. The remote view of a robot is often helpful, particularly if the robot is stuck or in a tight situation.
- *Design for the intended user, not the developer*. We have seen teams move from keyboard controls to joystick operation. Requiring users to remember mappings for 10

or more keys can be demanding. Joysticks can often be mapped more naturally into robot movement.

We believe that these design guidelines would hold true for all tasks with remote robots where the robot is not capable of full autonomy.

6. Lessons Learned

As there are no constraints on the hardware or software used in competitions, we were able to see a number of different types of user interfaces and interactions. Competitors perform the same tasks using different systems, so we can see how different user interaction approaches can affect performance. In competitions, performance can be measured objectively, such as via the number of victims found and the number of penalties for bumping arena walls or victims. USAR competitions form an upper bound on how well an interface can support users, in the sense that the people operating robotic systems in USAR competitions are the extremely technically-savvy developers of those systems.

Extensive data collection is possible during USAR competitions. We videotaped the robot(s) in the arena and over the shoulders of the robot operator(s). We captured the operators' interactions with the interface via dynamic screen-capture software that allowed us to play back what the operator saw and did with the interface displays. We mapped where the robots traveled in the arena and marked where bumping penalties occurred. Further, we conducted brief post-run interviews.

The detailed data collection efforts allow us to characterize the operators' use of the interface. For example, in past competitions we have determined the percentage of time users spent in different activities, such as in manipulating the interface versus navigating the robot; also, we examined what parts of the interfaces were most heavily used, and for what purposes.

Studies conducted at USAR competitions also help us learn about operators' situation awareness (SA) strategies. The most generally accepted definition of SA is Endsley's [1988]: the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future. Having good SA is especially important in remote robot operations, meaning the robot operator cannot see the robot from their control terminal; this is the case for USAR competitions. We were able to observe what information operators extracted from the interface to attain or reacquire SA, such as the example in which an operator angled a camera downward to see the robot wheels' in relationship to obstacles and voids.

Even in cases where operators had good SA, they still occasionally damaged the arena or their robots, or took actions that would harm victims (if they were real people rather than mannequins). Thus, USAR competitions afford us the opportunity to perform critical incident analyses. For example, we observed more bumping incidents happening to the rear of robots for those systems that did not have rear-facing cameras. We also investigated whether operators were aware that critical incidents had occurred, for example via post-run interviews. Often the information operators received via the interface was not sufficient to alert them to the fact that a critical incident had occurred; we looked at what information and/or presentation mechanism would have been necessary to provide the missing cues.

It is also worthwhile noting what we could not learn via studies of USAR competitions. Because we could not do anything that would jeopardize a team's competitive chances, we could not ask operators to "think aloud" [Ericsson and Simon 1980] or interrupt them with questions during a competition run. Thus, competitions afford limited insight into operators' mental models or ways of thinking about the interfaces. Further, we could not use anything from the class of "explicit performance" techniques for SA measurement such as the Situation Awareness Global Assessment Technique (SAGAT) [Endsley 1987], because these techniques involve short suspensions of the task during which operators answer questions. Further, we could not ask that the teams employ USAR personnel for competition runs. Thus, we could not normally see how well the interfaces could be used by their intended end users. In one case, however, we were able to invite a Fire Chief attending the competition to use two of the robotic systems in the NIST test arena.

Acknowledgements

This work is sponsored in part by the National Science Foundation (IIS-0415224, IIS-0308186) and the National Institute of Standards and Technology. Our thanks to all of the participating teams and those who assisted with the data collection, particularly Jean Scholtz (NIST), Brian Antonishek (NIST), Jeff Young (NIST), Elena Messina (NIST), Adam Jacoff (NIST), Brian Weiss (NIST), Mike Baker (UML), Bobby Casey (UML), Brenden Keyes (UML), and Phil Thoren (UML).

References

Drury, J.L., J. Scholtz and H.A. Yanco (2003). "Awareness in human-robot interactions." *Proceedings of the IEEE Conference on Systems, Man and Cybernetics*, Washington, DC, October 2003.

Endsley, M. R. (1988). "Design and evaluation for situation awareness enhancement," *Proceedings of the Human Factors Society 32nd Annual Meeting*, Santa Monica, CA.

Endsley, M. R. (1987). "SAGAT: a methodology for the measurement of situation awareness (NOR DOC 87-83)." Hawthorne, CA: Northrup.

Engelmore, R., and A. Morgan, eds. (1986). *Blackboard Systems.* Reading, Mass.: Addison-Wesley.

Ericsson, K. A. and H. A. Simon (1980). "Verbal reports as data." *Psychological Review* 87: 215 - 251.

Jacoff, A., E. Messina and J. Evans (2001). "A reference test course for autonomous mobile robots." *Proceedings of the SPIE-AeroSense Co*nference, Orlando, FL, April.

Jacoff, A., E. Messina and J. Evans (2000). "A standard test course for urban search and rescue robots." *Proceedings of the Performance Metrics for Intelligent Systems Workshop*, August.

Kieras, D. E. (1988). "Towards a practical GOMS model methodology for user interface design." In M. Helander (ed.), *The Handbook of Human-Computer Interaction*, 135 - 157. Amsterdam: North-Holland.

Leveson, N.G. (1986). "Software safety: why, what and how." *ACM Computing Surveys*, 18, 125 – 162, June.

Scholtz, J. (2002). "Evaluation methods for human-system performance of intelligent systems." *Proceedings of the 2002 Performance Metrics for Intelligent Systems (PerMIS) Workshop*.

Yanco, H.A. and J. Drury (2004). "'Where am I?' Acquiring situation awareness using a remote robot platform." *Proceedings of the IEEE Conference on Systems, Man and Cybernetics*, October.

Yanco, H.A., J.L. Drury, and J. Scholtz (2004). "Beyond usability evaluation: Analysis of human-robot interaction at a major robotics competition." *Journal of Human-Computer Interaction*, Volume 19, Numbers 1 and 2, pp. 117-149.