Study of Human-Robot Interaction in the AAAI-2002 Rescue Competition

Holly A. Yanco
University of Massachusetts Lowell

Jill L. Drury
The MITRE Corporation

Jean Scholtz
National Institute of Standards and Technology

RUNNING HEAD: STUDY OF HRI FOR SEARCH AND RESCUE

Corresponding Author’s Contact Information:
Holly A. Yanco
Computer Science Department
University of Massachusetts Lowell
One University Avenue, Olsen Hall
Lowell, MA 01854
E-mail: holly@cs.uml.edu
Phone: (978) 934–3642
Fax: (978) 934–3551

Brief Authors’ Biographies:

Dr. Holly A. Yanco is a roboticist with an interest in navigation in unstructured environments, assistive technology, and methods for improving shared human-robot control; she is an Assistant Professor of Computer Science at the University of Massachusetts Lowell.

Dr. Jill L. Drury is a usability engineer and researcher with an interest in evaluating collaborative computing systems; she is an Associate Section Leader in the Systems Engineering Directorate of The MITRE Corporation and an Adjunct Assistant Professor in the Computer Science Department of the University of Massachusetts Lowell.

Dr. Jean Scholtz is a computer scientist with an interest in evaluation of interactive systems and human-robot interaction; she is a research scientist in the Information Access Division of the Information Technology Laboratory at the National Institute of Standards and Technology.
ABSTRACT

Human-robot interaction (HRI) is a relatively new field of study. To date, most of the effort in robotics has been spent in developing hardware and software that expands the range of robot functionality and autonomy. In contrast, little effort has been spent so far to ensure that the robotic displays and interaction controls are intuitive for humans. By bringing together both robotics and human-computer interaction (HCI) expertise, our goal was to apply HCI evaluation techniques to HRI to gain experience with the use of HCI evaluation techniques in the robotics domain and to begin develop a set of design guidelines for HRI. We used as our case study four different robotics systems that competed in the 2002 American Association of Artificial Intelligence (AAAI) Robot Rescue Competition. These systems completed urban search and rescue tasks in a controlled environment with pre-determined scoring rules that provided objective measures of success. We analyzed pre-evaluation questionnaires; videotapes of the robots, interface, and operators; maps of the robots’ 22 paths through the competition arena; post-evaluation debriefings; and critical incidents (e.g., when the robots damaged the test arena). As a result, we developed guidelines for HRI such as: provide fused sensor information to lower cognitive load on user and provide user interfaces that support multiple robots in a single display.
1. INTRODUCTION

Human-robot interaction (HRI) is a relatively new field of study. To date, most of the effort in robotics has been spent in developing hardware and software that expands the range of robot functionality and autonomy. In contrast, little effort has been spent so far in ensuring that the robotic displays and interaction controls are intuitive for humans. While much work has been done in the fields of human-computer interaction (HCI) and computer-supported collaborative work (CSCW) to determine the usability of interfaces, little of this work has been applied specifically to the robotics domain. By bringing together both robotics and human-computer interaction (HCI) expertise, our goal was to apply HCI evaluation techniques to HRI to gain experience with the use of HCI evaluation techniques in the robotics domain and to begin develop a set of design guidelines for HRI.

We chose urban search and rescue tasks for our case study because they constitute prime examples of safety-critical situations (defined as a situation where a run-time error or failure could result in death, injury, loss of property, or environmental harm [Leveson 1986]). Safety-criticality imposes a requirement for error-free operation and is also often time-critical, resulting in a special need for efficient, intuitive HRI.

Specifically, we studied the HRI at the 2002 American Association for Artificial Intelligence (AAAI) Robot Rescue Competition. We focused on the effectiveness of techniques for making human operators aware of pertinent information regarding the robot and its environment.

The study had two parts, centering on the performance of four teams in the AAAI-2002 competition and on the use of two of the teams’ systems by a domain expert. The competition provided a unique opportunity to correlate objective performance (e.g., number of victims found, number of penalties assessed, percentage of competition arena area traversed) with user interface design approaches (e.g., degree of information fusion, presence or absence of a computer-generated map display, etc.). The juxtaposition of the team runs with the domain expert use of interfaces allowed us to compare expert (system developer) versus novice (domain expert) use of the interfaces.

Our contributions include:

- A study of how operators performed using the human-robot interfaces of four different systems in a physical simulation of a search and rescue experience.
- The first use of Scholtz’ [2002] evaluation issues as evaluation criteria for robotics.
- A coding scheme for classifying operator interactions.
- A correlation of performance with interface design.
- Guidelines for interface design for human-robot interaction based upon study results.
2 RELATED WORK FOR EVALUATION OF HRI

Before any interface (robotic or otherwise) can be evaluated, it is necessary to understand the users’ relevant skills and mental models and to develop evaluation criteria with those users in mind. Unlike HCI, where there are empirically validated sets of heuristics that have gained acceptance as evaluation criteria (e.g., [Nielsen 1994]), there is no generally accepted set of evaluation criteria for HRI.

Messina et al. [2001] proposed some criteria in the intelligent systems literature, but they are qualitative criteria that apply to the performance of the robot only, as opposed to the robot(s) and the user(s) acting as a cooperating system. An example criterion is “The system … ought to have the capability to interpret incomplete commands, understand higher level, more abstract commands, and to supplement the given command with additional information that helps to generate more specific plans internally.”

In contrast, Scholtz [2002] proposed six evaluation “issues” that can be used as high-level evaluation criteria. Scholtz raised these issues “to determine what information the user needs to understand what the intelligent system is doing and when intervention is necessary, and what information is needed to make any intervention as effective as possible.” Scholtz’ six issues are as follows:

- Is the necessary information present for the human to be able to determine that an intervention is needed?
- Is the information presented in an appropriate form?
- Is the interaction language efficient for both the human and the intelligent system?
- Are interactions handled efficiently and effectively – both from the user and the system perspective?
- Does the interaction architecture scale to multiple platforms and interactions?
- Does the interaction architecture support evolution of platforms?

Usability evaluations use effectiveness, efficiency, and user satisfaction as metrics for evaluation of user interfaces. Effectiveness metrics evaluate the performance of tasks through the user interface. In HRI, operators’ tasks are to monitor the behavior of robots (if the system has some level of autonomy), to intervene when necessary and to control navigation either by assigning waypoints, issuing a command such as “back-up,” or teleoperating the robot if necessary. Additionally, in search and rescue, operators have the task of identifying victims and their location. The performance in the competition was an overall measure of effectiveness. We also looked at the mistakes or penalties assigned to teams as another way of assessing the effectiveness of the user interface.

Not only must the necessary information be present, it must also be presented in such a way as to maximize its utility. Information can be present but in separated areas of the interface. This results in users having to manipulate windows to gain an overall picture of system state. This takes time and can result in an event not being noticed for some
Information fusion is another aspect of presentation. Time delays and errors occur when users need to fuse a number of different pieces of information.

As robots become more useful in various applications, we think in terms of using multiple robots. Therefore the user interfaces and the interaction architectures must scale to support operators controlling more than one robot.

Robot platforms have made amazing progress in the last decade and will continue to progress. Rather than continually developing new user interaction schemes, is it possible to design interaction architectures and user interfaces to support hardware evolution? Can new sensors, new types of mobility, and additional levels of autonomy be easily incorporated into an existing user interface?

We use Scholtz’s issues as an organizing theme for our analysis, tailoring them to be specific to the urban search and rescue environment (see the methodology section). We focused primarily on the first four evaluation criteria in this study but we also discuss the latter two based on the four systems we analyzed.

2.1 Evaluation of HRI

Evaluation methods from the HCI and CSCW worlds can be adapted to use in HRI as long as they take into account the complex, dynamic, and autonomous nature of robots. The HCI community often speaks of three major classes of evaluation methods: inspection methods (evaluation by user interface experts), empirical methods (evaluation involving users), and formal methods (evaluation focusing on analytical approaches). Robot competitions lend themselves to empirical evaluation because they involve users performing typical tasks in as realistic an environment as possible (see [Yanco 2001] for a description of some robot competitions). Unfortunately (from the viewpoint of performing formal usability testing), robot competitions normally involve the robot developers, not the intended users of the robots, operating the robots during the competition. The performance attained by robot developers, however, can be construed as an “upper bound” for the performance of more typical users. Specifically, if the robot developers have difficulty using aspects of the interface, then typical users will likely experience even more difficulty. Additionally, robot competitions afford an interesting opportunity (one not attained so far in formal usability testing of HRI) to correlate HRI performance under controlled conditions to HRI design approaches.

Previous studies of HRI have also used different types of empirical methods. Fong [2001] used a contextual inquiry approach [Holtzblatt and Jones 1993] to evaluate a “collaborative control” approach to sharing control between the user and the robot, but it was not clear that the subjects of the contextual inquiry drew from the pool of typical users. In contrast, Casper and Murphy [2002, also Casper 2002] performed a study of real operators directing robots through the rubble of the World Trade Center towers. They did not call their data collection methodology contextual inquiry, but the work contained some of the features of contextual inquiry because robot usage was examined in the context of a real search and rescue situation.
While the AAAI Robot Competition provided us with an opportunity to observe users performing search and rescue tasks, there were two limitations. First, we were not able to converse with the operators due to the time constraints they were under, which eliminated the possibility of conducting talk-aloud protocols and also eliminate our ability to have operators perform specific tasks. Secondly, the competition simulated a rescue environment. Many of the hazards and stress-inducing aspects of an actual search and rescue environment were missing. Nonetheless, this environment was probably the closest we could use in studying search and rescue tasks due to safety and time constraints in actual search and rescue missions.

Two patterns were observed in previous HRI empirical testing efforts that limit the insights obtained to date. The first, as mentioned above, is a tendency for robot performance to be evaluated using atypical users. For example, Yanco [2000] used a version of a usability test as part of an evaluation of a robotic wheelchair system, but did not involve the intended users operating the wheelchair (the wheelchair was observed operating with able-bodied occupants). We have started to break this pattern by also analyzing the use of two urban search and rescue robot systems by a Fire Chief, a more typical user, after the competition runs were completed.

The second pattern that limits HRI empirical testing effectiveness is the tendency to conduct such tests very informally. For example, Draper et al. [1999] tested the Next Generation Munitions Handler/Advanced Technology Demonstrator (NGMH/ATTD), which involves a robot that re-arms military tactical fighters. While experienced munitions loaders were used as test subjects, testing sessions were actually hybrid testing/training sessions and test parameters were not held constant during the course of the experiment. Data analysis was primarily confined to noting test subjects’ comments such as “I liked it when I got used to it.” Our study took advantage of the structure inherent in the conduct of the AAAI Robot Competition to keep variables such as environment, tasks, and time allowed to complete tasks constant.

3 METHODOLOGY

The two portions of the study consisted of evaluating the interfaces as their developers competed and when the domain expert performed four tasks with each of the interfaces. We correlate competition performance with various features in the interface design; therefore, we describe the competition scoring methodology after the subsection on our evaluation of the competition. We coded the resulting videotapes of both portions of the study using the same coding scheme, which we describe at the end of this section.

3.1 Methodology for Assessing Interaction Design

An accepted evaluation methodology in human-computer interaction is to take a general set of principles and tailor them for use in evaluating a specific application (e.g., see [Nielsen 1993]). We took Scholtz’ evaluation criteria [Scholtz 2002] and tailored them as follows to be more specific to the case of HRI in an urban search and rescue context.
“Is the necessary information for the human to be able to determine that an intervention is needed present?” becomes *Is sufficient status and robot location information available so that the operator knows the robot is operating correctly and avoiding obstacles?* “Necessary information” is very broad. In the case of urban search and rescue robots, operators always need information regarding the robot’s health if it is not operating correctly. Another critical piece of information operators need is the robot’s location relative to obstacles, regardless of whether the robot is operating in an autonomous or teleoperated mode. In either case, if the robot is not operating correctly or is about to collide with an obstacle, the operator will need to take corrective action.

“Is the information presented in an appropriate form?” becomes *Is the information coming from the robots presented in a manner that minimizes operator memory load, including the amount of information fusion that needs to be performed in the operators’ heads?* Robotic systems are very complex. If pieces of information that are normally considered in tandem (e.g., a robot’s video and laser ranging sensor information) are presented in different parts of the interface, the operator will need to switch his attention back and forth, remembering what he saw in a previous window in order to fuse the information mentally. Operators can be assisted by information presentation that minimizes memory load and maximizes information fusion.

“Is the interaction language efficient for both the human and the intelligent system? Are interactions handled efficiently and effectively – both from the user and the system perspective?” Combining these two, they become *Are the means of interaction provided by the interface efficient and effective for the human and the robot (e.g., are shortcuts provided for the human)?* We consider these two criteria together because there is little language per se in these interfaces; rather, the more important question is whether the interactions minimize the operator’s workload and result in the intended effects.

“Does the interaction architecture scale to multiple platforms and interactions?” becomes *Does the interface support the operator directing the actions of more than one robot simultaneously?* A goal in the robotics community is for a single operator to be able to direct the activities of more than one robot at a time. Multiple robots can allow more area to be covered, can allow for different types of sensing and mobility, or can allow for the team to continue operating after an individual robot has failed. Obviously, if this is to be the case, the interface needs to enable the operator to switch his or her attention among robots successfully.

“Does the interaction architecture support evolution of platforms?” becomes *Will the interface design allow for adding more sensors and more autonomy?* A robotic system that currently includes a small number of sensors is likely to add more sensors as they become available. In addition, robots will become more autonomous and the interaction architecture will need to support this type of interaction. If the interaction architecture has not been designed with these possibilities in mind, it may not support growth.
3.2 Methodology for Studying Team Performance

Teams voluntarily registered for the competition. We asked them to participate in our study, but made it clear that study participation was not a requirement for competition participation. The incentive to participate in the study was the chance to have their robot system used by a domain expert in the second part of the study.

Teams were asked to fill out a questionnaire before the start of the competition. The questions inquired about the robot hardware being used, the type of data provided to the human operator, the level of autonomy achieved by the robot, the maturity of the robot design, and whether the interface is based on a custom (bespoke) or commercial product. This information was used to inform our study regarding the information that is presented to the users through each interface.

Once the competition began, we observed the operator of each team’s robot(s) during the three 15 minute runs of the competition. The operator and the interface screen were videotaped. The robots were also videotaped in the arena; cameras were placed in various locations around the arena in an attempt to keep the robot constantly within sight.

We were silent observers, not asking the operators to do anything differently during the competition than they would have already done. This was an important point: our study could not impact upon the competition outcome. Note that this meant we could not ask the participants to use the “thinking aloud” protocol [Ericsson and Simon 1980], although one participant who was eager to obtain feedback on his interface voluntarily voiced his thoughts as he worked with his robot during the competition. At the conclusion of 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 such as hitting objects or victims that occurred during the runs.

3.3 Methodology for Scoring Team Performance

The 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]. There are three regions in the arena, which vary in difficulty. The yellow region is the easiest to traverse, with flat floors. The orange region is more difficult to traverse with variable floorings, stairs, a ramp, and holes. The red region is the most difficult to traverse, with a very unstructured environment consisting of simulated rubble piles, pancake layers, and hazardous junk such as rebar and wire cages.

The scoring algorithm utilizes the number of victims found and the accuracy of reporting the location of the victims. The scoring scheme penalizes teams for allowing robots to
bump into obstacles or victims.\textsuperscript{1} The judges recorded a minor victim penalty for bumping into a victim (subtracting .25 from the number of victims found) and a major victim penalty was scored for an event such as causing a pancake layer to collapse on a victim (subtracting 1). Minor damage to the environment, such as moving a wall a small amount, was marked as a minor environment penalty (subtracting .25) while major environment damage, such as causing a secondary collapse, was considered a major penalty (subtracting .75).

The scoring formula is as follows: 
\[ V = \text{number of victims found} \]
\[ P = \text{penalties} \]
\[ A = \text{accuracy} = 1 \text{ if map produced by system; 0.6 if good quality hand drawn map produced; 0.4 if poor hand drawn map produced (the accuracy score was determined by the competition judges).} \]

\[
\text{Performance Score} = (V - P) \times A
\]

3.4 Methodology for Studying Domain Expert Performance

After the competition, we had access to a search and rescue domain expert: a special operations fire chief who has participated in training sessions with robots for search and rescue. The goals of the evaluation were to assess ease of learning as well as ease of use. To evaluate ease of learning, the domain expert was asked to explore the interface for five minutes to determine what information is available in the interface. Then the domain expert was given about five minutes of training, which would be a realistic amount of training in the field in an emergency condition if the primary (more thoroughly-trained) operator suddenly becomes unavailable [Murphy 2002]. After the training, the domain expert was asked to describe the information available in the interface that he didn’t see during the initial exploration period. Finally, the domain expert was asked to navigate the robot through the arena.

The domain expert was able to verbalize his thoughts as he navigated the robots. He produced a combination of talk-aloud and think-aloud protocols [Ericsson and Simon 1993]. In general, as he was navigating through the arena, he used the talk-aloud protocol. However, there were a number of times when we experienced technical difficulty, and the chief had to wait for a resolution to the problem before he could proceed. During these times, his verbalizations were more introspective.

3.5 Methodology for Coding Team and Domain Expert Sessions

Our data consisted of videotapes, competition scoring sheets, maps of robot paths, questionnaire and debriefing information, and researcher observation notes. The richest

\textsuperscript{1} The scoring algorithm used for comparing teams in this study differs from the official scoring algorithm used in the competition [AAAI/RoboCup 2002]. We factored out measures that were unrelated to the interface, such as a measure for calculating a bonus when unique victims were found by different robots.
source of information was the videotapes. In most cases, we had videotapes of the robots moving through the arena, the user interfaces, and videos of the operators themselves.

To make the most of the videotaped information, we developed a coding scheme to capture the number and duration of occurrences of various types of activities observed. Our scheme consists of a two-level hierarchy of codes: header codes capture the high-level events and primitive codes capture low-level activities. The following header codes were defined: identifying a victim, robot logistics (e.g., undocking smaller robots from a larger robot), failures (hardware, software, or communications), and navigation/monitoring navigation (directing the robot or observing its autonomous motion). Three primitive codes were defined: monitoring (watching the robot when it is in an autonomous mode), teleoperation (“driving” the robot), and user interface manipulation (switching among windows, selecting menu items, working with dialog boxes, typing commands, etc.).

Our coding scheme was inspired by the structure of the Natural Goals, Operators, Methods, and Selection rules Language (NGOMSL) used to model user interface interaction [Kieras 1988]. NGOMSL models consist of a top-down, breadth-first expansion the user’s top-level goals into “methods,” and the methods contain only primitive operations (“operators”), branch statements, and calls to other NGOMSL methods. Our top-level header codes can be thought of as NGOMSL goal-oriented methods for identifying a victim, navigation/monitoring, or handling robot logistics or failures. While our primitives are not always true primitives (e.g., an activity such as teleoperation can be usually broken down into finer-grained motor control actions), they are at the lowest level it makes sense to analyze and thus are analogous to NGOMSL primitives.

The coding was done by two sets of researchers. To obtain inter-coder reliability, both sets coded the same run and compared results. The Kappa computed for agreement was .72 after chance has been excluded. We then discussed the disagreements and, based on a better understanding, we coded the remaining runs. We easily agreed upon the coding for the events that were observable, but noted that the timing of those events could only be determined within a few seconds. Unfortunately, we could not see the robot when it was in a covered area or when it was in the small portions of the arena that the cameras did not cover. We are looking for ways to overcome these limitations for subsequent competitions.

4 DESCRIPTIONS OF SYSTEMS STUDIED

Eight teams entered the competition. However, we only investigated the HRI of the four teams who found victims during their runs; these teams were also the top-ranked teams. Teams that were unable to find victims most often had hardware failures and no significant amount of HRI to study. In this section, we describe each of the four systems in our study, including the user interface and the robot hardware.

2 When chance was not factored out the agreement was .8.
<table>
<thead>
<tr>
<th><strong>Taxonomy Classification</strong></th>
<th><strong>Description</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Autonomy level</td>
<td>Percentage of time that a robot is carrying out a task on its own.</td>
</tr>
<tr>
<td>Amount of intervention</td>
<td>Percentage of time that a human operator must control a robot.</td>
</tr>
<tr>
<td>Human-robot ratio</td>
<td>The ratio of operators to robots; the operating ratio is the ratio of people to number of robots that can be controlled simultaneously.</td>
</tr>
<tr>
<td>Level of shared interaction</td>
<td>Various combinations of whether the humans and robots act independently or as part of team(s).</td>
</tr>
<tr>
<td>Composition of robot teams</td>
<td>Whether teams of multiple robots are homogeneous or heterogeneous.</td>
</tr>
<tr>
<td>Available sensors</td>
<td>A list of sensor types available on the robot platform.</td>
</tr>
<tr>
<td>Provided sensors</td>
<td>A subset of the list of available sensors; the sensor data that is actually provided in the interface.</td>
</tr>
<tr>
<td>Sensor fusion</td>
<td>A list of functions mapping the provided sensor data to the fused output (e.g., {(sonar, lidar) \rightarrow \text{map}}).</td>
</tr>
<tr>
<td>Criticality</td>
<td>The importance of getting the task done correctly in terms of its negative effects should problems occur.</td>
</tr>
<tr>
<td>Time</td>
<td>Whether the humans and robots work together in the same time (synchronously) or different times (asynchronously).</td>
</tr>
<tr>
<td>Space</td>
<td>Whether the humans and robots work together in the same place (collocated) or in different places (non-collocated).</td>
</tr>
</tbody>
</table>

**Table 1: HRI Taxonomy [Yanco and Drury 2002]**

We developed a taxonomy for HRI systems [Yanco and Drury 2002], which we use in table 2 to describe the systems. A short description of each of the taxonomy classifications is contained in table 1. Since the task is the same for all of the systems, the following categories have the same values across all systems: Criticality = High, Time = Synchronous, and Space = Non-collocated.

### 4.1 Team A

Team A developed a heterogeneous team of five robots: one iRobot ATRV-Mini and four Sony AIBOs, for the primary purpose of research in computer vision and multi-agent systems. They spent three months developing their system for the rescue competition. All robots were teleoperated serially. The AIBOs were mounted on a rack at the back of the ATRV-Mini. The AIBOs needed to be undocked to start their usage 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, which were created for use by the developers and were not tested with other users before the competition. There was one user interface for the ATRV-Mini and another for the AIBOs. The UIs ran on separate computers. Communication between the UI and the robots was accomplished using a wireless modem (802.11b).

The user interface for the ATRV-Mini, shown in figure 1, 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
<table>
<thead>
<tr>
<th>Taxonomy element</th>
<th>Team A</th>
<th>Team B</th>
<th>Team C</th>
<th>Team D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autonomy</td>
<td>0%</td>
<td>0 – 100% (level selected by operator)</td>
<td>50%</td>
<td>0%</td>
</tr>
<tr>
<td>Intervention</td>
<td>100%</td>
<td>0 – 100% (level selected by operator)</td>
<td>50%</td>
<td>100%</td>
</tr>
<tr>
<td>Human-Robot Ratio</td>
<td>1/5 (operating ratio 1/1)</td>
<td>1/1</td>
<td>1/2 (operation level usually 1/1)</td>
<td>1/2 (operation level usually 1/1)</td>
</tr>
<tr>
<td>Interaction</td>
<td>One human, multiple robots</td>
<td>One human, one robot</td>
<td>One human, multiple robots</td>
<td>One human, multiple robots</td>
</tr>
<tr>
<td>Robot-team-composition</td>
<td>Heterogeneous (ATRV-Mini =R1; AIBO = R2)</td>
<td>Homogeneous</td>
<td>Homogeneous</td>
<td>Heterogeneous (wheeled = R1; tracked = R2)</td>
</tr>
<tr>
<td>Available-sensors (R1)</td>
<td>Video (circle of 8 cameras), sonar (unused), SICK laser scanner, odometry</td>
<td>Video, sonar, infrared, SICK laser scanner, thermal camera</td>
<td>Video, sonar, infrared, bump</td>
<td>Video</td>
</tr>
<tr>
<td>Provided-sensors (R1)</td>
<td>Video, current laser scan (raw and processed for map)</td>
<td>Video, thermal imaging (raw), infrared, bump, laser scan, and sonar</td>
<td>Video, sonar, infrared</td>
<td>Video</td>
</tr>
<tr>
<td>Sensor-fusion (R1)</td>
<td>{Odometry, laser scanner} (\rightarrow) map</td>
<td>{Infrared, bump, laser scan, sonar} (\rightarrow) sensor map</td>
<td>{Sonar, infrared} (\rightarrow) overhead map (evidence grid)</td>
<td>None</td>
</tr>
<tr>
<td>Available-sensors (R2)</td>
<td>Video</td>
<td>N/A</td>
<td>N/A</td>
<td>Video</td>
</tr>
<tr>
<td>Provided-sensors (R2)</td>
<td>Video</td>
<td>N/A</td>
<td>N/A</td>
<td>Video</td>
</tr>
<tr>
<td>Sensor-fusion (R2)</td>
<td>None</td>
<td>N/A</td>
<td>N/A</td>
<td>None</td>
</tr>
</tbody>
</table>

Table 2: Classification of the studied systems using the HRI taxonomy

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, shown in figure 2, 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. (The domain expert controlled the AIBOs using the keyboard because of a problem with the GUI at that time.)
Figure 1: Team A’s interface for the iRobot ATRV-Mini

Figure 2: Team A’s interface for the Sony AIBOs
4.2 Team B

Team B has been developing their robot system for use in hazardous environments for less than one year. The robot is an iRobot ATRV-Jr. Communication is achieved through a proprietary, low-bandwidth communication protocol over 900 MHz radio.

The custom user interface, shown in figure 3, was developed for expert users and tested with novice users and real operators. The interface is displayed on a touch screen. In the upper left corner of the interface is the video feed from the robot. Tapping the sides of the window moves the camera left, right, up or down. Tapping the center of the window recenters the camera. (During the 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.) The robot was equipped with two types of cameras that the operator could switch between: a color video camera and a thermal camera.

In the lower left corner is a window displaying sensor information such as battery level, heading, and tilt of the robot. In the lower right corner, a sensor map is displayed, showing filled red areas to indicate blocked directions. In the picture of the interface
above, a map of the environment is displayed in the upper right corner. Although this space was left for a map during the competition, the software for building and displaying maps had not yet been created, so no maps were provided to the operator.

The robot is controlled through a combination of a joystick and the touch screen. To the right of the sensor map, there are six mode buttons. From top to bottom, they are Auto (autonomous mode), Shared (shared mode, a semiautonomous 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 over to the thermal camera for verification.

4.3 Team C

Team C has been developing their robots for less than 2 years as a research platform for vision algorithms and robot architectures. They use two identical robots, RWI Magellan Pros. Communication between the user interface and robots is achieved with an RF modem.

The robots have a mixed level of autonomy: they can be fully teleoperated or the robots can provide obstacle avoidance while achieving a specified goal. The robots could run simultaneously, but were operated serially. Waypoints were used to generate maps from the robot's current location to the starting point. Command of a robot was achieved by giving it relative coordinates to move towards. The robot then autonomously moved to that location using reactive obstacle avoidance. This allows for the perception that the operator is moving both robots “at once,” even though he is controlling them serially. It is the operator’s trust in the robots’ autonomy that allows this type of operation; the operator does not need to monitor the progress of one robot while commanding the other.

A custom interface has been developed for a “sophisticated user” (according to the developers). Team C started run 1 using a graphical user interface, but switched to a text-based interface when there were command latency problems with the GUI. In the GUI, the screen was split down the middle; each side was an interface to one of the two robots. The top window for each robot displays a current video image from the robot and the bottom window displays map information.

In the alternative text-based interface, the screen had 14 text windows and 4 graphic windows, half for each of the robots. The 7 text windows are for the following: the IPC (interprocess communication) server, the navigation module, the vision module, the mapping module, the navigation command line, a window for starting and monitoring the
visual display, and a window for starting and monitoring the map display. The two graphic windows are for displaying the camera image and the map image. The computer was running an enlarged desktop during the competition, and the operator sometimes needed to switch to another part of the desktop (effectively switching to another screen) for other pieces of the interface. The robots were controlled with keystrokes. During run 3, the operator used the text-based interface; it was already set up at the start of the run.

![Figure 4: Team C’s graphical interface](image)

### 4.4 Team D

Unlike the other three systems, Team D did develop their robots for search and rescue over the past year. They 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 4 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.

### 5 RESULTS

We present two types of results for the teams: the objective measures from the competition and the results of our coding. We also present the coded results of the
domain expert’s performance along with his talk-aloud and think-aloud protocols. Finally, we analyze performance using the tailored Scholtz criteria from section 3.1.

5.1 Team Runs

Each team had three 15-minute runs during the competition. We only coded runs 1 and 3 due to the failure of the video data capture equipment during run 2. The total times are in some cases less than the allotted 15 minutes; it was sometimes difficult to discern the actual starting time for the competition to coordinate the start of data capture.

<table>
<thead>
<tr>
<th>Team</th>
<th>Run</th>
<th>Total Time</th>
<th>% Time Navigation/ Monitoring Navigation</th>
<th>% Time Victim ID</th>
<th>% Time Failure</th>
<th>% Time Logistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Team A</td>
<td>1</td>
<td>10:39</td>
<td>46%</td>
<td>51%</td>
<td>0%</td>
<td>3%</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>14:45</td>
<td>62%</td>
<td>18%</td>
<td>19%</td>
<td>1%</td>
</tr>
<tr>
<td>Team B</td>
<td>1</td>
<td>14:33</td>
<td>81%</td>
<td>19%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>16:42</td>
<td>77%</td>
<td>23%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Team C</td>
<td>1</td>
<td>13:26</td>
<td>59%</td>
<td>23%</td>
<td>17%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>14:39</td>
<td>69%</td>
<td>12%</td>
<td>18%</td>
<td>0%</td>
</tr>
<tr>
<td>Team D</td>
<td>1</td>
<td>15:12</td>
<td>55%</td>
<td>32%</td>
<td>0%</td>
<td>12%</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>13:30</td>
<td>87%</td>
<td>4%</td>
<td>0%</td>
<td>9%</td>
</tr>
</tbody>
</table>

Table 3: Time spent in each of the primary header codes for competition runs.

<table>
<thead>
<tr>
<th>Team</th>
<th>Run</th>
<th># Victims</th>
<th>Penalties</th>
<th>Accuracy</th>
<th>Score</th>
<th>Team Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Team A</td>
<td>1</td>
<td>4</td>
<td>8 * 0.25</td>
<td>1.0</td>
<td>2.0</td>
<td>3.5</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3</td>
<td>6 * 0.25</td>
<td>1.0</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>Team B</td>
<td>1</td>
<td>3</td>
<td>5 * 0.25</td>
<td>0.6</td>
<td>1.05</td>
<td>&lt; 1.05</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>1 * 0.25 + 3 * 0.75</td>
<td>NA</td>
<td>1.2</td>
<td>3.15</td>
</tr>
<tr>
<td>Team C</td>
<td>1</td>
<td>3</td>
<td>0 * 3 * 0.25</td>
<td>0.4</td>
<td>1.2</td>
<td>3.15</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4</td>
<td>9 * 0.25</td>
<td>0.6</td>
<td>1.95</td>
<td></td>
</tr>
<tr>
<td>Team D</td>
<td>1</td>
<td>6</td>
<td>4 * 0.25</td>
<td>0.5^10</td>
<td>1.875</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3</td>
<td></td>
<td>0.6</td>
<td>1.2</td>
<td>3.075</td>
</tr>
</tbody>
</table>

Table 4: Team scores, computed using the algorithm in section 3.3.

3 Includes navigation to get a new angle for victim id after a judge said that the first image was unclear.
4 Wireless modem failures.
5 Additionally, about 25% of the victim identification time was spent in logistics while deploying AIBOs.
6 The operator spent 90% of this navigation time in a confused state. However, the equipment had not malfunctioned, so this was not coded as a failure.
7 GUI latency, panoramic image failure
8 Panoramic image failure, vision system on one robot failed midway through run
9 Logistic time for Team D was for switching between their two robots
10 This number reflects a penalized accuracy score, as determined by the judges. There was some question as to whether advance knowledge of the arena layout had been obtained.
Table 3 shows the percentage of time spent in each of the primary header code activities. The majority of time for most runs is spent navigating, followed by identifying victims. Time spent in logistics or failures was time taken away from looking for victims.

Team scores are shown in table 4. Since we did not analyze the HRI in run 2, we only consider runs 1 and 3 in the scoring. Using the scoring algorithm in section 3.3, the rankings for the two rounds would be as follows: 1st Place: Team A, 2nd Place: Team B, 3rd Place: Team C, and 4th Place: Team D.11

5.2 Domain Expert Runs

5.2.1 Ease of Learning

The domain expert, a special operations fire chief trained in search and rescue and with experience using robots, used two systems: Team A (teleoperated) and Team B (different autonomy modes). We started each session with a short amount of time for the Chief to explore the interface without instruction. After this period, we asked him to state what he could figure out about the interface. Then the system developers explained the interface to him, and we asked the Chief what was in the interface that he had not seen before.

For Team B, the Chief said that there was no real time video (the team was having trouble with their video link at this time, so they were only sending about one frame per second). He noted that there were sensors around the robot, pointing to the sensor map in the lower right hand corner, and that the map appeared to be displaying proximity information. After the Chief talked with Team B’s developer, he stated that he had learned about the control modes for the robot.

For Team A, the Chief said that he saw a laser map on the lower right, a video display on the upper right, an ultrasonic map on the left and a data window under that. He could not see how to drive, but thought he’d do it using the arrow and the mouse. After the developer’s explanation, the Chief learned that the window on the left did not have an ultrasonic map, but was instead displaying a map created as he drove using the laser scan and odometry. He also learned how to control the robot and that there was a ring of cameras on top of the robot for the video window. A window with radio buttons labeled 1 to 8 was used to switch from one camera view to another in the video window.

5.2.2 Performance

The Chief had been a judge for the competition so he was more familiar with the arena at the time of his runs than any of the competitors had been. Table 5 shows the amount of time the domain expert spent in each of the primary header codes. The times shown in the table include the time that the expert was using the systems, not any time that he was speaking to the system developer or the researchers.

11 The actual rankings in the competition, which included run 2 as well as other measures such as in which part of the arena victims were found, were as follows: 1st Place: Team D, 2nd Place: Team C, 3rd Place: Team A, and 4th Place: Team B.
We observed the Chief relying heavily upon the live video for navigation. He would drive, change camera angles, then resume driving. We will discuss this further in section 5.3.1.

### 5.3 Evaluation Using Tailored Scholtz Criteria

We use the performance of the teams and of the domain expert, the results of coding activities of the operators during the competition runs, and an examination of critical incidents to discuss Scholtz’s criteria as tailored in section 3.1.

#### 5.3.1 Is sufficient status and robot location information available so that the operator knows the robot is operating correctly and avoiding obstacles?

<table>
<thead>
<tr>
<th>Run</th>
<th>Arena Penalties</th>
<th>Victim Penalties</th>
<th>Rank(^{14})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Team A (ATRV-Mini)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1 minor(^{15})</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4 minor</td>
<td>0</td>
<td>–</td>
</tr>
<tr>
<td>Team A (AIBOs)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>7 minor(^{16})</td>
<td>0</td>
<td>–</td>
</tr>
<tr>
<td>3</td>
<td>2 minor</td>
<td>0</td>
<td>–</td>
</tr>
<tr>
<td>Team B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3 minor</td>
<td>2 minor</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>1 minor, 3 major</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Team C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3 minor</td>
<td>0</td>
<td>–</td>
</tr>
<tr>
<td>Team D</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>9 minor</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>4 minor</td>
<td>0</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 6: Number of penalties incurred by the teams

---

12 Victims were removed from the arena during the chief’s runs.
13 Communication failure: video signal was not updating.
14 1 is the best, 4 is the worst bumping record overall based on numbers of bumps. AIBOs are not ranked because they were used for only short periods of time.
15 The ATRV-Mini was used for approximately 12 minutes during each of runs 1 and 3. Normalizing to 15 minutes would result in 1.25 and 5 minor arena penalties, which does not affect Team A’s overall ranking.
16 The AIBOs were used for approximately 3 minutes during runs 1 and 3.
The number of penalties for each team is shown in table 6. Note that Team A’s two different types of robots are listed separately, because the ATRV-Mini has dramatically different sensor capabilities than the AIBOs. Although another team, Team D, also fielded robots of different types, their robots differed only in their navigation properties (one type was tracked and the other was wheeled); otherwise, their sensor suites and operational capabilities were identical.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team</th>
<th>Sensor Types Provided</th>
<th>Sensor Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>C</td>
<td>Video, sonar, infrared</td>
<td>Sonar, infrared fused into overhead map</td>
</tr>
<tr>
<td>2</td>
<td>A(^{17})</td>
<td>Video, laser ranging</td>
<td>Odometry, laser ranging fused into laser map</td>
</tr>
<tr>
<td>3</td>
<td>B</td>
<td>Video, raw thermal imaging, sonar, infrared, bump, laser ranging</td>
<td>Infrared, bump, laser ranging, and sonar fused into sensor map</td>
</tr>
<tr>
<td>4</td>
<td>D</td>
<td>Video</td>
<td>None</td>
</tr>
</tbody>
</table>

**Table 7: Summary of sensor types provided and sensor fusion for each team**

We had thought that Team B would fare slightly better than they did. Team B’s operator experienced serious confusion when he forgot that his robot’s video camera was pointing in a direction other than straight ahead. This resulted in more than half of Team B’s first run (8 1/2 minutes) being wasted. The interface did not provide any reminders that the video camera was pointing off-center, so this lack of awareness of robot state (rather than a paucity of sensor data) caused him to run into more obstacles and find fewer victims than he might have otherwise.\(^{18}\) We are unsure why he also had a poor run 3. During this run, he operator was frustrated that his robot was “too big.” In fact, he did have the largest robot in the competition.

We saw several specific instances where operators were unaware of robot location/surroundings. In several cases (e.g., Team D during run 1), there was not enough awareness of the area immediately behind the robot, causing the robot to bump obstacles when backing up. Even when moving forward, several operators (e.g., Team B during run 3) hit walls and were not aware of doing so. One of Team A’s robots was trapped under fallen Plexiglas but the operator was never aware of this situation. Since they didn’t have precise awareness of the area immediately around the robot, operators (e.g., Team B during run 3) had a difficult time maneuvering the robots in tight spaces.

One of the debriefing questions we asked after each run was how the operator perceived the performance of the run. Surprisingly, Team B’s operator stated after run 3 that he had had not hit anything during the run. However, his perceptions did not correspond with reality; he had incurred 1 minor and 3 major arena penalties during this run. Clearly, the operator did not have sufficient awareness of the robot, its surroundings and its activities.

\(^{17}\) ATRV-Mini only.

\(^{18}\) This problem was corrected by the developers before other runs by changing the program to recenter the camera.
The Chief’s bumping performance is shown in Table 8. While he was not scored, we marked the times that he hit objects just as was done for the teams. These penalties were marked over the full length of the Chief’s runs, which were about 10 minutes longer than an average team run.

While using Team A’s system, the Chief asks twice if someone is watching in the arena. The first time he says he isn’t sure if the robot is clear of a wall. The second instance occurs when he thinks the robot might be caught on a cable but he is told that the robot is clear. To resolve this, he deploys an AIBO from the ATRV-Mini and positions the camera on the ATRV-Mini so that he can view the AIBO while he is teleoperating it.

The Chief had begun to experiment with Team B’s system earlier and stopped due to wireless interference. He did not feel comfortable relying on the sensor display, the still updates, and various modes of autonomy for navigation. In this early run, the Chief was using the safe mode of navigation and was unable to understand why he couldn’t navigate through a perceived opening. He put the robot in teleoperation and discovered that the “opening” was covered with Plexiglas, but only when people called from the arena area to state that the robot had charged through a panel.

<table>
<thead>
<tr>
<th>Arena Penalties</th>
<th>Victim Penalties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Team A (ATRV Jr.)</td>
<td>0</td>
</tr>
<tr>
<td>Team A (AIBOs)</td>
<td>0</td>
</tr>
<tr>
<td>Team B</td>
<td>2 minor, 6 major</td>
</tr>
</tbody>
</table>

Table 8: “Penalties” incurred by the Chief during his runs

When using both systems, the Chief adjusted the camera views frequently but even then he had difficulty knowing where the robot was. The team operators using these systems relied far less on moving the cameras around to acquire awareness than the Chief did. Team B’s operator relied on the sensor data and used various modes of autonomy. He used the camera views when he was identifying a victim. However, he also had imperfect awareness; there were a number of instances when he bumped into obstacles and was penalized in the scoring but never noticed this during the run. Team A’s operator used the dynamically created map and the laser scanning data for navigating, but he also had sub-optimal awareness. One of the AIBOs fell off the ATRV-Mini, but the operator was completely unaware of this.

The developers seem to feel more comfortable relying on sensor data other than video. This may be a false sense of security as their penalty scores reflect their lack of awareness. The Chief, a novice user, was more cautious and, even though he commented about the usefulness of the sensor data, he still relied heavily on live video feeds. Again, this proved to be problematic as well. Further, not all of the necessary information is presented to users; more information is needed regarding the awareness of the relationship of the robot to its environment, as evidenced by a number of bumping incidents.
5.3.2 Is the information coming from the robots presented in a manner that minimizes operator memory load, including the amount of information fusion that needs to be performed in the operators’ heads?

System D was the only system in the competition that had no information fusion in the system, using only video. System A was the only system that presented a map in the display that included the walls of the arena. This map allowed the operator to see where he had been so that he could hopefully avoid covering the same territory numerous times. Table 9 shows that Team A had better coverage than all with the exception of Team D for run 1.\(^{19}\)

<table>
<thead>
<tr>
<th>Run</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Team A 50% yellow</td>
</tr>
<tr>
<td>3</td>
<td>Team A 35% yellow</td>
</tr>
<tr>
<td>1</td>
<td>Team B 20% yellow, 5% orange</td>
</tr>
<tr>
<td>3</td>
<td>Team B 35% yellow, 10% orange</td>
</tr>
<tr>
<td>1</td>
<td>Team C 30% yellow</td>
</tr>
<tr>
<td>3</td>
<td>Team C 35% yellow</td>
</tr>
<tr>
<td>1</td>
<td>Team D 80% yellow</td>
</tr>
<tr>
<td>3</td>
<td>Team D 15% yellow, 10% orange, 5% red</td>
</tr>
</tbody>
</table>

Table 9: Amount of the arena covered

Team C also had a map in their interface, but it presented only the sonar readings of the robot as it moved through the arena. No corrections were made for dead reckoning errors. While Team A’s map looked like a floor plan, Team C’s map looked like a fat line composed of black triangles. However, we see in table 7 that Team C’s coverage of the arena ranks second, above the two teams (C and D) with no mapping capabilities.

Although additional sensor information should provide additional awareness as a general rule, this rule does not hold true if more information is provided but the information is not integrated into the displays in a way that an operator can use. In general, lack of data fusion, other than that contained in maps, hindered operators’ ability to quickly obtain an understanding of the robot's status and location. For example, in systems A and B, the video image was presented separately from the sonar or laser ranging sensor data, often in opposite corners of the display screen. Such separation causes the operator to mentally synthesize the data as opposed to having the interface provide a combined picture.

Presenting related data in opposite corners of the display is an example of how the displays were not laid out for maximum efficiency nor memory load minimization. Evidence of this trend can be seen by the fact that operators spent a large percentage of the time in user interface manipulation. Various types of information were, in general, presented in separate windows so that operators spent significant time periods moving between windows. Operators then had to remember what was in one window and combine it with information in other windows. Some operators needed to constantly

\(^{19}\) There was some question as to whether Team D had prior knowledge of the arena for run 1.
glance between video and other data, or move between windows on the display, while mentally fusing the various pieces of information.

When using System A, the Chief initially noted that he relied primarily on the laser for navigation. However, his primary navigation method was to stop teleoperating the robot and to change the view of the camera to look around. This was also the case when the Chief used Team B’s system. He relied heavily on live video and commented when the reception was particularly bad. The Chief had some experience teleoperating robots prior to our study. However, video can miss some types of obstacles, as evidenced by the fact that the Chief drove through a Plexiglas panel.

During the first run, the Team B operator moved the robot’s video camera off-center to look at a victim for identification, and also switched to his thermal camera to verify that it was a live victim. After the victim identification, the operator switched to shared mode to allow the robot to get out of a tight space with less operator intervention. At this point, the operator forgot that he had turned his camera to the left. When he switched back to safe mode, he found that the results of his actions did not correspond to the video image he saw. This confusion resulted in the operator accidentally driving the robot out of the arena into the crowd, and bumping into a wall trying to get back into the arena. It also resulted in substantial operator confusion (we recorded quotes such as, “it’s really, really hard,” “I got disoriented,” “hmmmn,” and “oh, no!”). During the last run, Team B’s operator did not have good visibility into the areas behind the robot, making it difficult for him to maneuver it out of narrow spaces (“this is very difficult”). After the last run, Team B’s operator commented that he had not bumped anything, yet four bumping penalties were assessed by the judges.

Team C started a run using a GUI, but within two minutes, the operator determined that there was too much lag time between command issuance and response. As a result, he shut down the GUI and brought up seven windows that formed an earlier version of the interface (the debugging version). It took a little over a minute and a half for the operator to shut down the GUI and bring up all the windows for the earlier interface version. In this interface, the operator needed to shuffle through the seven windows to view different types of information and entered commands in several of the windows.

5.3.3 Are the means of interaction provided by the interface efficient and effective for the human and the robot (e.g., are shortcuts provided for the human)?

We saw evidence of inefficient interaction mechanisms that resulted in the user having to switch windows or modes frequently, primarily because the output of each sensor seemed to be provided in a different window. Further, we noted instances where interactions were not effective. The prime example of an ineffective interaction was the case where the operator’s efforts to navigate the robot through the arena were unsuccessful due to the fact that he had forgotten that he had previously changed the pointing angle of the video camera from a straight-ahead orientation. Since the interface provided no clues to remind the user that he had forgotten to restore the video camera angle, he persisted in navigating in the wrong directions.
5.3.4 Does the interface support the operator directing the actions of more than one robot simultaneously?

The amount of work an operator needed to do to use a robot (via the user interfaces in the competition) was sufficiently high so that it was unrealistic to expect to operate multiple robots simultaneously. Even though the systems of Teams A, C, and D were designed to operate with more than one robot simultaneously, in practice, the robots were controlled serially. (Recall, however, that Team C was able to have more than one robot navigating at a time due to the autonomy of the systems. However, the operator could only focus on one robot at a time due to the split interface.) Facilitating additional autonomy would help workload but some amount of monitoring would still be necessary.

With the teams’ current user interface designs, virtually all of the operators’ attention was needed to run one robot at a time due to several reasons. First, as mentioned above, operators were busy integrating information from the video and the other portions of the interface (e.g., the map showing the current location of the robot in x-y space, thermal images, and video images). Second, there was a high overhead cost to switch from operating one robot to another. All of the windows were duplicated for each robot, rather than having the information integrated into one set of windows. In fact, our coding revealed that Team C, the only team to field two robots simultaneously, spent 7% of their navigation time giving commands to the user interface in run 1, in which one robot was used. During run 3, when two robots were operated, 13% of the navigation time was spent issuing commands to the robots. Doubling the number of robots doubled the number of commands. Clearly, there will be problems when scaling up, even when robots have some autonomy.

Three of the four competition teams fielded more than one robot. The fact that only one (Team C, Run 3) of the teams operated more than one robot at a time is indicative that their interaction architectures are not appropriately scaled to handle interactions with multiple robots simultaneously. The approach taken to adding multiple robots seems to be to add another set of windows, where many of the windows display only one type of sensor data. With this approach, the user quickly runs out of screen real estate and the cognitive power to mentally fuse the appropriate information for each robot.

Further, each of the interfaces examined makes the user completely responsible for gathering awareness of the robot's state and location by means of moving the video camera around.20 Hence we saw many short periods of navigation with lots of gathering awareness in between, where the robot stops moving as the operator manipulates the cameras. Such an approach is difficult to do for more than one robot simultaneously.

5.3.5 Will the interface design allow for adding more sensors and more autonomy?

The interaction architectures we studied do not support robot evolution. Robot evolution usually involves additional sensors and more autonomy; more sensors will require more

---

20 No other sensors could be manipulated by the interface; if the user wanted to get a different view using non-video sensors, the robot would need to be moved.
windows if the current interaction architectures are extended. While one robot user interface we examined does support various modes of autonomy that could ease operator workload, it currently falls to the operator to determine which mode should be used and to switch the robot as necessary. Table 10 shows the percentage of navigation spent in each of three autonomy modes\textsuperscript{21} for Team B, as well as the number of mode switches made during the run.

<table>
<thead>
<tr>
<th>Run</th>
<th>% Time Safe</th>
<th>% Time Shared</th>
<th>% Time Autonomous</th>
<th># of Mode Switches</th>
<th>Time Spent Navigating</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>63</td>
<td>34</td>
<td>3</td>
<td>20</td>
<td>11:47</td>
</tr>
<tr>
<td>3</td>
<td>56</td>
<td>8</td>
<td>17\textsuperscript{22}</td>
<td>19</td>
<td>12:50</td>
</tr>
</tbody>
</table>

Table 10: Percentages of time spent in different autonomy modes by Team B’s operator

We see that Team B’s operator made 20 mode switches in run 1 and 19 mode switches in run 2. The Chief changed modes 12 times during his run, with the majority of the switches occurring at the end of his time with the robot.

It would be more helpful if the robot could determine the necessary mode based on sensor information and suggest it to the operator, rather than relying on the operator to constantly revisit the decision regarding the optimal mode.

6 DISCUSSION

Our study of the operator role looked at systems ranging from complete teleoperation to systems allowing some degree of autonomy. We looked at systems with sensory input ranging from video only to robots with sensor suites that included laser ranging, sonar, infrared, thermal cameras and video cameras. We found that more sensor types do not necessarily increase awareness, especially if the sensor data is not well fused into information for the operator.

We present initial guidelines for designing interfaces for human-robot interaction, based upon our observations in the study:

- Provide a map of where the robot has been. As we saw in section 5.3.2, operators using systems with maps were more successful in navigating arena area. Without a map, the operator must try to track the robot’s path in his head.
- Provide fused sensor information to lower the cognitive load on user. In the three interfaces with multiple data types (systems A, B, and C), all required the user to mentally fuse video with other sensor streams.
- Provide user interfaces that support multiple robots in a single display. We saw in section 5.3.4 that the number of commands doubled when two robots were used instead of one. These commands needed to be entered in two separate windows.

\textsuperscript{21} The operator never used teleoperation, which did not provide any sensor mitigation.

\textsuperscript{22} The percentages for run 3 do not sum to 100\% for the three modes; 19\% of the navigation time was spent in unknown modes, due to an inability to see which mode was selected by the user in the videotapes.
• Minimize the use of multiple windows. With additional sensor fusion, more information could be displayed in a single window.
• Provide more spatial information about the robot in the environment. This could take the form of a map, discussed above, or some other method. At the very least, operators must be aware of their immediate surroundings to avoid bumping into obstacles or victims.
• Provide robot help in deciding which level of autonomy is most useful. System B had four levels of autonomy available, and the operator needed to select the method appropriate for the current situation. The sensor data on the robot could be processed to assist with this decision. For example, we noticed that the operator of system B changed to autonomous mode whenever he felt that he was in a very tight situation; the robot could easily automate this switch or the suggestion of this switch.

This paper contains frameworks for helping to organize results of future evaluations: the taxonomy of HRI, evaluation issues and coding methods. We encourage other researchers in the HRI field to utilize and extend these frameworks to maximize our ability to learn from future studies and to be able to quickly transfer results into practice.

ACKNOWLEDGEMENTS

Many people assisted with the study, including Brian Antonishek (NIST), Mike Baker (UMass Lowell), Jenn Casper (American Standard Robotics), Adam Jacoff (NIST), Elena Messina (NIST), Mark Micire (American Standard Robotics), Jesse Miyoshi (NIST), Justin Osborn (NIST), Craig Schlenoff (NIST), Phil Thoren (UMass Lowell), Brian Weiss (NIST), the participating teams, and the domain expert. Thanks to the editors and reviewers for their suggestions to improve this paper. This work was supported in part by NSF IIS-0308186.

REFERENCES


Murphy, R. (2002). Personal communication.


