"Turn Off the Television!": Real-World Robotic Exploration Experiments with a Virtual 3-D Display

David J. Bruemmer, Douglas A. Few, Miles C. Walton, Ronald L. Boring, Julie L. Marble
Human, Robotic, and Remote Systems Department
Idaho National Engineering and Environmental Laboratory
{bruedj, fewda, mwalton, borirl, marbjl}@inel.gov

Curtis W. Nielsen
Brigham Young University
curtisn@cs.byu.edu

Jim Garner
Washington University at St. Louis
g6@cse.wustl.edu

Abstract

In order to apply mobile robots to a new range of applications, we require control architectures and interfaces that support symbiotic interaction. Remote deployment of mobile robots offers one of the most compelling opportunities to merge human intelligence with machine proficiency. This paper discusses a mixed-initiative control strategy based not on video, but on an abstracted, collaborative workspace -- a 3-D, video-game representation constructed on-the-fly -- that promotes situation-awareness and efficient tasking. The new interface requires orders of magnitude less bandwidth than teleoperation and permits transmission ranges of thousands of miles. Unlike video, which offers only a 1st person, local environment perspective, the 3-D interface changes perspective to support changing levels of operator involvement and robot autonomy. The human-participant study presented evaluates the effectiveness of this interaction substrate on a remote exploration task. Results indicate that this new tool for interfacing humans and intelligent robots can reduce human error, support changing levels of human workload, promote human trust, and enable a spectrum of remote robotic applications which have never before been possible.

1. Introduction

For several decades it has been assumed that humans should interact with robots primarily through a master-slave relationship based on streaming video sent from the robot to the human operator. Humans are visually centric and generally prefer pictures and diagrams to communicate. It makes good sense to utilize video when appropriate, especially if the application requires visual identification of targets or areas of interest. The question is just how useful streaming video really is? Do human operators really prefer and demand streaming video or might some new strategy not be possible that would appeal to the human operator and provide a more effective means to represent the environment and communicate about the task?

It has been well-recognized that dependence on continuous, streaming video is inherently limiting [1], [2]. Video demands high-bandwidth, reliable, continuous communication and is therefore often impossible. Except for short ranges (< 100 meters), transmission of high-bandwidth video is only possible when line of sight can be maintained either with a satellite or another radio antenna. For instance, high-bandwidth video cannot be transmitted through layers of concrete and rebar, making it inappropriate for urban terrain. Likewise, forest and jungle canopy precludes reliable, long-range transmission of high-bandwidth video. It has long been assumed that advances in communication will one day alleviate these technical limitations, but at the present time, reliable transmission of streaming video remains an elusive goal.

Even if it were theoretically possible to transmit unlimited visual data, would video be the optimal method to provide situation awareness and communicate about

the environment? Many environments do not provide useful visual cues. In many military and search and rescue scenarios, the visual scene is often occluded by smoke or dust, giving operators the illusion that the robot is traveling through a hazy world of grey pixels. Many environments lack ambient light such as covert military operations, caves, mines, pipelines, etc.. Even when light is available, video rarely shows the user what they need to see. Often the very features of the environment that are most critical for navigation such as door frames or sudden drop-offs are beyond the visual field presented to the user. Although cameras can be well-placed to provide multiple-perspectives, the inundation of video data can prove debilitating for an operator who generally can focus on only one perspective at a time. Ultimately, video provides only a first-person view of the local environment – this perspective is only appropriate for a master-slave relationship between the human and robot and is poorly suited to support changing levels of robot autonomy.

Most of the recent work in simultaneous localization and mapping has used an occupancy grid approach to build a 2-D representation of the world [3,4,5,6]. This approach is well-suited to the scanning lasers and ultrasonic sensors available for use on mobile robots. The problem with 2-D maps is that they too offer only one perspective – a god's eye view of the world – which, like video, cannot scale to support different modes of robot autonomy and different levels of operator involvement.

If we really want to engender dynamic cooperation where the human and robot can work together as peers, then we must have a new form of shared representation that is meaningful to both the robot and the human. Despite a progress in the area of computer vision, streaming video is for the most part unintelligible to the robot and does not provide a meaningful form of representation to the robot. On the other hand, a representation that is built up from the robot's own sense of the world (i.e. range sensing) allows the robot to link semantic abstractions to real-world locations and entities.

Within the world of computer gaming, we found exactly what we were looking for: the ability to turn the problem of remote robotic deployment into a video game. Once the environment, robot and task are encoded into the representation, it is possible to alter the perspective at will to support changing levels of robot autonomy. Most importantly, the abstracted data necessary to map the robot's real-world sensor data to a video game display can be sent over an low-bandwidth data transmission such as a single cell-phone or a long range radio. Whereas video requires at least 3,000,000 bits / second, the game interface requires only 64,000 bits / second. This is a savings of almost 5000%. Due to the limited bandwidth, we can limit our transmission speed to 9600 baud, which allows us to transmit many miles through thick concrete, canopy and even the ground itself, enabling a new realm of possible applications.

2. Mixed-Initiative Control

Teleoperated systems have often failed to address the limitations of telepresence inherent to current communication technologies. On the other hand, attempts to build and use autonomous systems have failed to acknowledge the inevitable boundaries to what the robot can perceive, understand, and decide apart from human input. Both approaches have failed to build upon the strengths of the robot and the human working as a cohesive unit [7]. Alternatively, mixed-initiative systems can support a spectrum of control levels. Mixed-Initiative robots should possess intrinsic intelligence, knowledge and agency; protect humans, environment and self; dynamically shift levels of initiative to accept different levels and frequencies of intervention, and recognize when help is needed (from human or machine).

Towards these aims, research efforts at the INEEL have developed a novel robotic system that can leverage its own intelligence to support a spectrum of control levels. We submit that rather than conceive of machines as mere tools or, on the other hand, as totally autonomous entities that act without human intervention, it is more effective to consider the machine as part of a dynamic human-machine team. Within this schema, each member has equal responsibility for performance of the task, but responsibility and authority for particular task elements shifts to the most appropriate member, be it human or machine. For instance, in a remote situation, the robot may be in a much better position than the human to react to the local environment, and consequently, the robot may take the leadership role regarding navigation. As leader, the robot can drive autonomously or can "veto" dangerous human commands to avoid running into obstacles or tipping itself over.

The resulting robotics system including hardware, software, and interface components, is designed to support changing levels of operator involvement. The ability of the robot to change its level of autonomy on the fly supports changing communication, cognitive, perceptual and action capabilities of the user and robot. With the new system, communications dropouts no longer result in the robot stopping dead in its tracks or, worse, continuing rampant until it has recognized that communications have failed.

3. System Design

3.1 Robot Implementation

To give a robot this capability is no easy task. If we are to someday collaborate with robots as peers, we must, first develop more trustworthy robot platforms and behaviors. Since no one platform is appropriate for all tasks, the INEEL has developed a behavior architecture that can port seamlessly to a variety of robot geometries and sensor suites including those shown below. For the study reported in this paper, the "ATRV mini" robot (shown to the far left in the Figure below) was utilized.

Figure 1: The family of robots on which the INL dynamic autonomy control architecture resides

INEEL has worked for some time to provide robust mechanisms that allow the robot to protect itself and the environment. To do so, we fuse a variety of range sensor information including inertial sensors, compass, wheel encoders, laser range finders, computer vision, thermal camera, infrared break beams, tilt sensors, bump sensors, sonar, and others. The robot does not assume that these sensors are working correctly, but rather continuously evaluates its own perceptual capabilities and behavior. Novel sensor-suites and fusion algorithms enhance capabilities for sensing, interpreting, and "understanding" environmental features. With the new system we are not limited to visual feedback. Instead, the robot is able to abstract information about the environment at many levels including terse textual descriptions of the robot's local surroundings.

In terms of the study discussed in this paper, the video capabilities were of paramount importance. For the study we used a standard off-the-shelf Sony pan-tilt-zoom camera with auto focus and auto iris capabilities. The most critical decision was how to transmit this data. Due to the distance and physical occlusions separating the control station from the actual robot environment, analog video was not an option. By exploiting the ethernet infrastructure already in place throughout the building we were able to use wireless ethernet to transmit from the robot to a wireless access point, which was connected to the building's network. This then allowed us to provide continuous, realiable video, which far exceeded the performance of anything which could be accomplished in a purely wireless fashion. Also, to provide a fast update rate, we used an AXIS 2001 video compression box (www.axis.com) that digitizes the analog video and uses an MPEG format to efficiently send out the video data over an IP-based network (i.e. wireless LAN).

3.2 Operator Control Unit

The screen shot below shows a full view of the standard interface with the video interface module visible.

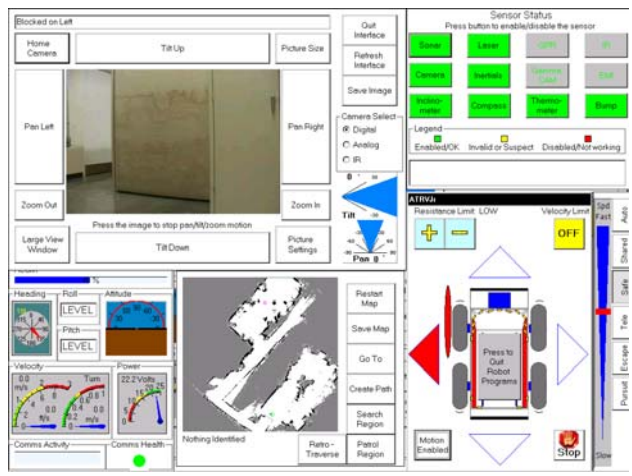

Figure 2: The standard configuration of the INL control interface

The robotic interface is the culmination of iterative usability testing and redesign [5]. In designing the interface, we attempted to strike a balance between ease of robotic control and the rich information display necessary for monitoring hazardous environments or conducting search and rescue. The interface consists of a single touch screen display containing five sizeable windows (see Figure 2). The upper left-hand window on the screen contains a video feed from the robot as well as controls for adjusting the camera. The upper right-hand window contains status indicators and controls for the robot's sensors. The lower right-hand window features movement
status indicators and controls as well as a mode selector for different levels of robot autonomy. The lower central window provides an emerging map of the environment as determined by a simultaneous mapping and localization algorithm discussed below. The lower left-hand window contains information about the robot's operational status.

Control of the robot can be achieved by touching appropriate areas of the display. The effect of these touches depends on the mode of autonomy. When in direct control, operators primarily give directional commands using the joystick. Depending on which interface configuration we use, operators may also pan, tilt and zoom the camera by using another, 3-axis joystick on the interface console or can use a "joystick hat" that consists of a mini-joystick and buttons placed on top of the main joystick. However, for the present study we wanted to insure that operators did not waste time trying to adjust the field of view. For a search and detection task, the camera manipulation capabilities can be very useful. However, for this experiment, which was focused on map building and exploration, we believed that allowing users to devote time to panning, tilting and zooming could bias the experiment towards the game interface. Consequently, we told participants not to utilize the pan, tilt or zoom capabilities unless they believed it was necessary for navigation.

3.3 Mapping and Localization

One of the reasons why intelligent indoor robots have not infiltrated our lives *en masse* is that if we are not willing to instrument the environment, robots inevitably become disoriented in unstructured domains. As we sought to create a viable representation of the environment, we quickly learned that the most important concern is the accuracy of the map and the ability of the robot to precisely locate itself within this representation. There have been many approaches to the problem of how to simultaneously build and localize within a map while moving through a new environment.

Within the INEEL control architecture, the ability of the robot to automatically generate the 3D virtual map representation is based on a technique developed at SRI called Consistent Pose Estimation (CPE) that allows for efficient incorporation of new laser scan information into a growing map [8]. Within this framework, SRI has found a solution to the challenging problem of loop closure: how to optimally register laser information when the robot returns to an area previously explored (and 'recognize' that it was there previously) [9]. CPE is one method for performing Simultaneous Localization and Mapping (SLAM). It is based on original work by Lu and Milios, who showed that information from the robot's wheel encoders and laser sensors could be represented as a network of probabilistic constraints linking the successive positions (poses) of the robot [10]. CPE provides an efficient means of generating a near-optimal solution to the constraint network, and yields high-quality metric maps (see figure below). Using this algorithm, the robot can not only build a representation of its environment on-the-fly, but can maintain an accurate estimation of where it is within this map to within +/- 5cm.

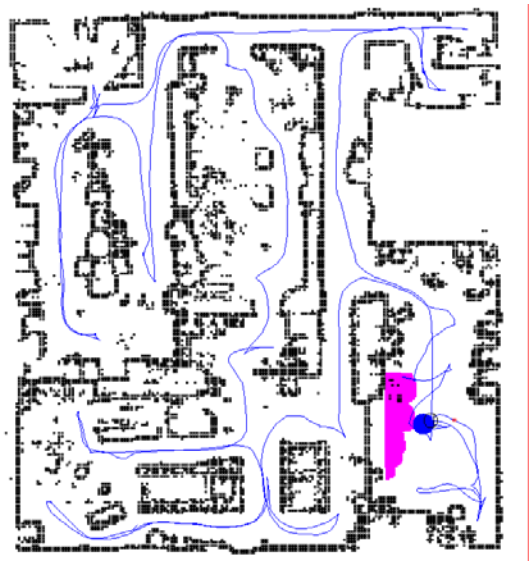

Figure 3: A map of the INL robotics building built using SRI International's mapping and localization software.

3.3 Virtual 3D Display

The major innovation explored in this paper is the development and evaluation of the interface component shown below which allows the robot's view of the world to be communicated efficiently to the user. It also provides a means to store and fuse a variety of task-centric information as semantic entities within the 3-D world, reducing the mental workload of the user. In particular, the ability to reduce the environment down to a semantic map can greatly help users perform tasks that demand spatial reasoning and memory such as a map building or maze exploration task. Most importantly, the interface allows users to slide seamlessly between an egocentric display and exocentric display to provide different levels of intervention. Adjustable perspectives enable the user to interact with the robot efficiently regardless of the task at hand.

This component has been developed by melding technologies from Idaho National Lab (INL) [11], Brigham Young University (BYU) [12], and Stanford Research Institute (SRI) International [8]. The 3-D virtual display is not based on 3-D range sensing, but rather on the 2-D mapping and localization algorithm described above. The processing constraints on the robot negate the possibility of handling 3 dimensional range data even if a suitable sensor was available. Instead, we have found it sufficient to transform the 2-D map into a 3-D map

shown in figures 3 and 4. The benefit of the 3-Dimensional interface is that the perspective can be scaled in all directions. Like the other interface modules discussed above, this component can be scaled to any size and is intended to be used together with the other interface modules as needed.

Note that in the figure above, the collaborative workspace includes not only obstacles, but also other semantic entities that are of significance to the operator. The operator may choose entities from a drop down menu or may actually choose to insert translucent still images excerpted from the robot video. In this manner, the workspace can be used to help the user remember not only what has been seen, but where. In this manner, the workspace can support both virtual and real elements. Within the workspace, the dark rectangles represent walls or objects identified by the mapping algorithm and the robot model is drawn at its current localized position with respect to the discovered map. The robot model is also scaled to match the size of the actual environment, thereby enabling the user to comprehend the relative position of the robot in the real environment. By changing the zoom, pitch, and yaw of the field of view, it is possible to move from a more egocentric perspective where the user is actually looking out from the robot all the way to a fully exocentric view where the entire environment can be seen at once. As Scholtz points out, the roles of human operators do not remain static and interfaces should be designed to adapt accordingly. [13] Figure 4 shows a perspective somewhere in between a fully egocentric and exocentric display. This perspective can be used to support tasks that demand navigation and spatial reasoning simultaneously.

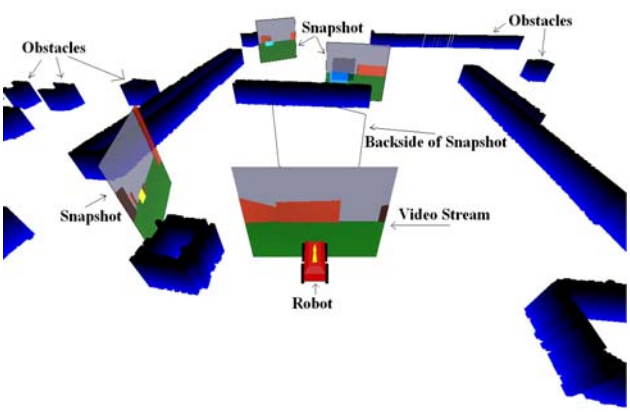

Figure 4: The virtual 3-D display which includes obstacles as well as task-centric semantic entities

4.0 Experiment

We needed some means to empirically evaluate the effectiveness of the virtual 3-D display. How well would this new interface strategy compare when placed head to head against a reliable streaming video display? Although the robot permits several different modes of autonomy, we ran all trials using safe mode which allows the human to drive the robot by using the joystick, but allows the robot to take initiative to protect itself and the environment from collision. Although our other studies have focused on performance across different modes of autonomy [7], [12], the purpose of this experiment was to provide a fair comparison of video versus the virtual 3-D interface. We decided that at least for this experiment, multiple modes of autonomy might complicate the analysis.

4.1 Participants

The experiment was performed over a seven day period within the St. Louis Science Center and utilized 64 volunteers between the ages of 12 and 60 who happened to be at the St. Louis Science Center. We believe that the participants in this study represent a random sampling of the population. In any case, a previous study with our control architecture indicates that there are no statistically significant effects for age or gender pertaining to users of our system [14].

4.2 Test Arena

We chose to base the experiment around a true remote deployment such that the operator control station was located several stories above and several hundred feet to the side of the arena where the robot operated. This arena was built by the production staff of the Science Center and utilized rocks, trees and manequins as well as plywood dividers to create a maze environment. The maze (see Figure 5) was confusing for operators driving the robot and despite the presence of visual features within the environment, participants (as well as the researchers) found it extremely difficult to spatially reason about the environment without the map.

To make the experiment as compelling as possible we chose to compare our game interface to the best possible video we could produce. In our first trial runs we realized that video participants would be at a significant disadvantage simply because the ambient lighting, although seemingly normal, cast shadows that made it very difficult to navigate with video. Although this is often the case in real-world deployments, we wanted to give the participants using video a fair shot and consequently, the production staff augmented the ceiling lighting quite significantly.

Figure 5: A partial view of the arena built at the St. Louis Science Center (Note the red robot)

4.3 Procedure

Each participant was given basic instructions on how to use the interface, but no participant was permitted to drive the robot until the start of the trial run. Each trial run was exactly 3 minutes. We have done previous experiments [7], [12] which required participants to devote between 1 minute to 1 hour operating the robot. We found that too little time may confound the data because first time operators can often be confused at first. On the other hand, too much time leads to boredom, which can also mask the differences between participants and interfaces. We believe that for this task the time allotted was appropriate. It permitted the majority of participants to explore over 50% of the total environment while only one person was able to build the entire map in 3 minutes.

For each new trial, the robot and human began with no map. Each participant was told to direct the robot around the environment in order to build as big a map as possible. This task involves spatial reasoning because the operator must be able to perceive the frontiers of the map and direct the robot to them in an optimal fashion. A day of preliminary runs proved that users with no map representation at all were dead in the water. To make the evaluation of our virtual 3-D display more rigorous, we gave both the participants access to the 2-D map. We presented exactly one half of the participants with the interface as depicted in figure 2. These participants were able to use both the 2-D map and the video module. For the other half we occluded the video module entirely with the virtual 3-D interface module.

During each trial, the interface stored a variety of useful information about the run including the joystick bandwidth used, the number of messages sent from the interface to the robot and the number of times that the robot was forced to take initiative to prevent a collision. As the human drives the robot, the interface indicates physical blockages that impede motion in a given direction as red ovals next to the iconographic representation of the robot (lower left of figure 2). When the robot takes initiative to stop, the user should immediately be able to discern that the robot is indeed blocked based on these indicates. However, previous experiments taught us that not every operator is able or willing to attend to these visual indications. As a result, we also employed a force feedback joystick that resists motion in the blocked direction. Thus when the human, fails to understand the situation and continues to try to advance the robot into an obstacle, the joystick vibrates and emits a loud noise. It is specifically these instances which the system automatically logs. Consequently, the robot initiative metric indicates not only human error, but also human confusion.

For each trial we saved the map produced by the robot. In order to determine performance, we ran each of the maps through a software algorithm that calculates the percentage of the real world that was explored. We believe that this approach, although not without its problems, provides a reasonable, objective assessment and is much more relevant than a measure of distance traveled.

Figure 6: A near-complete map built up by one of the participants.

4.4 Results

In contrast to our previous experiments [7, 14] which were designed to determine the benefits of different levels of robot initiative, this experiment focused on a quantitative analysis of performance, workload, error, and feeling of control collected during the exploration task. Based on our previous experiment with novice users [14] it was assumed no statistical performance differences exist across age or gender differences.

To metric task performance we calculated percent coverage by comparing the maps generated during the exploration task with an a priori map of the task environment. This comparison showed no significant statistical difference between the use of the video interface module and the virtual 3D map module, *M* .71, *M* .61, respectively, $F(1,31)=0.572$ $p = 0.079$. This indicates that sacrificing video does not, as is almost always assumed, necessarily result in significant performance degradation.

Using joystick bandwidth as a metric for human workload and robot initiative as a metric for human error, we found that operators using the virtual 3-D display worked less and demonstrated fewer instances of navigational error. On average the joystick bandwidth for participants using the 3DI was 1057.50 messages from the interface to the robot compared to 1229.07 for operators using video feed from the robot, $F(1,31)=2.024$, $p < 0.05$ while the robot initiative for participants using the 3DI averaged 11.00 to the 14.29 average or the video participants, : $F(1,31) = 0.399$, $p < 0.05$.

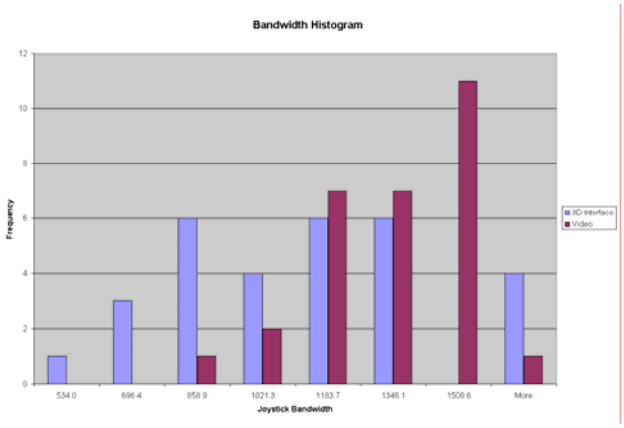

Figure 7: Joystick bandwidth histogram

In addition to reduced workload and fewer errors the 3DI operators enjoyed an elevated feeling of control while operating the robot. The average feeling of control for the 3DI operator was 7.219 compared with the 7.059 average of the video operators, $F(1.31)=0.497$, $p < 0.05$. In summary, with no penalty to task performance, operators enjoyed a reduced workload, fewer errors, and a heightened sense of control while remotely operating a robotic system without the use of video.

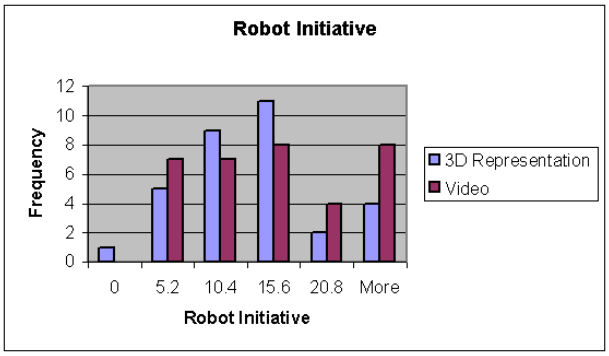

Figure 8: Robot Initiative Histogram

5. Conclusions

The experiment discussed in this paper provides compelling evidence that we can support the visually centric needs of human operators without necessarily resorting to video. Across a variety of homeland defense, military, department of energy, space exploration and industrial contexts, we can begin to apply this new interaction method to a broad range of tasks and applications where continuous video was not possible**.** Already, the United States Joint Robotics Program responsible for robots used across the Department of Defense, is investigating the possibility of using this system on small mobile robots to be sent into caves, underground bunkers and large engineered structures.

Moreover, we believe that the virtual 3-D interface will promote dynamic autonomy and allow the potential benefits of mixed-initiative control to be more fully realized. Interfaces built around video are appropriate primarily for a master-slave relationship and are unsuitable for monitoring dynamic autonomy systems that permit different levels of operator involvement. This issue is especially important for multiple robot operations where it becomes impossible for a single operator to monitor or task multiple robots in a teleoperated fashion. Using the virtual 3-D display, it is possible to represent all robots within the same display. In fact, current work at the INEEL is adapting the virtual 3D display for use in a countermine operation where multiple robotic vehicles used for humanitarian and military demining can simultaneously contribute to and be tasked via the same display.

6. Acknowledgement

This work is supported through the INEEL Long-Term Research Initiative Program under DOE Idaho Operations Office Contract DE-AC07-99ID13727 and by DARPA under grant NBCH1929913.

7. References

1. D. J. Bruemmer, J. L. Marble, D. D. Dudenhoeffer, Intelligent Robots for Use in Hazardous Environments, *PerMIS 2002*, Gaithersburg, MD, August 2002.

1. B. Trouvain, H. Wolf, F. Schneider, Impact of Autonomy in Multi-Robot Systems on Teleoperation Performance, *In Proceedings from the 2003 NRL Workshop on Multi-Robot Systems*, Washington, D. C. March, 2003.

2. S. Thrun, "A Probabilistic Online Mapping Algorithm for Teams of Mobile Robots", *Int. Journal of Robotics Research*, 20(5), pp. 335-363.

3. D. Fox, J. Ko, B. Stewart, K. Konolige, and B. Limetkai. "Distributed multi-robot mapping". In A. Schultz, L. Parker, and F. Schneider, editors, *Multi-Robot Systems: From Swarms to Intelligent Automata*, volume II. Kluwer, 2003.

4. F. Dellaert, D. Fox, W. Burgard, and S. Thrun, "Monte Carlo Localization for Mobile Robots," *IEEE International Conference on Robotics and Automation (ICRA99)*, May, 1999.

5. W. Burgard, D. Fox, D. Hennig and T. Schmitdt. "Estimating the absolute position of a mobile robot using position probability grids." *In Proc. of the Thirteenth National Conference on Artificial Intelligence*, pages 896-901, 1996.

6. F. Dellaert, D. Fox, W. Burgard, and S. Thrun, "Monte Carlo Localization for Mobile Robots," *IEEE International Conference on Robotics and Automation (ICRA99)*, May, 1999.

7. Marble, J. L., Bruemmer, D. J., Few, D. A. "Lessons Learning from Usability Tests with a Collaborative Cognitive Workspace for Human-Robot Teams." In Proceedings of the *2003 IEEE International Conference on Systems, Man and Cybernetics*, Washington, D.C., October 5-8, 2003.

8. K. Konolige. Large-scale map-making. *In Proceedings of the National Conference on AI (AAAI)*, San Jose, CA, 2004.

9. Gutman, J.S. and K. Konolige, "Incremental Mapping of Large Cyclic Environments", *CIRCA 99*, Monterey, California, 1999.

10. Lu, F. and E.E. Milios, "Globally Consistent Range Scan Alignment for Environment Mapping", *Autonomous Robots*, 4(4), 1997.

11. D. J. Bruemmer, M.O. Anderson. "Intelligent Autonomy for Remote Characterization of Hazardous Environments," *in Proceedings of the IEEE International Symposium on Intelligent Control*, Houston, TX. October 2003.

12. C. Nielsen, B. Ricks, M. Goodrich. D. Bruemmer, D. Few, M. Walton. "Snapshots for Semantic Maps," In *Proceedings of Systems, Man and Cybernetics 2004*, The Hague, Netherlands, October 10-13, 2004.

13. Scholtz, J., Human-Robot Interactions: Creating Synergistic Cyber Forces, *In Proceedings from the 2002 NRL Workshop on Multi-Robot Systems*, Washington, D. C. March, 2002.

14. D. Bruemmer, R. Boring, D. Few, J. Marble, M. Walton. "I Call Shotgun!: An Evaluation of Mixed-Initiative Control for Novice Users of a Search and Rescue Robot," In *Proceedings of Systems, Man and Cybernetics 2004*, The Hague, Netherlands, October 10-13, 2004.