Industrial Robot

An International Journal

Performance evaluation of autonomous mobile robots

Adam Jacoff
Based at Intelligent Systems Division, National Institute of Standards and Technology, Gaithersburg, Maryland, USA
Elena Messina
Based at Intelligent Systems Division, National Institute of Standards and Technology, Gaithersburg, Maryland, USA
John Evans
Based at Intelligent Systems Division, National Institute of Standards and Technology, Gaithersburg, Maryland, USA

Emerald

Industrial Robot: An International Journal, Vol. 29 No. 3, 2002,

Research article

Performance evaluation of autonomous mobile robots

Adam Jacoff
Elena Messina and
John Evans

The authors

Adam Jacoff, Elena Messina and **John Evans** are at Intelligent Systems Division, National Institute of Standards and Technology, Gaithersburg, Maryland, USA.

Keywords

Performance, Robots, Perception, Evaluation

Abstract

The National Institute of Standards and Technology (NIST) has initiated a program to develop quantitative metrics for machine intelligence. One of the possible approaches to evaluating machine intelligence is task-based performance testing, like a mouse in a maze. A series of application-specific testbeds is envisioned. NIST has created a set of reference test arenas for evaluating the performance of autonomous mobile robots performing urban search and rescue tasks. Robots must explore the maze-like test course, map the environment, find the simulated victims, and then report back their findings. In this paper, we describe our approach toward developing performance metrics for autonomous mobile robots through standardized testing within representative environments and objective performance evaluations. Our intent is to help accelerate the robotic research communities' advancement of mobile robot capabilities, thereby improving the effectiveness of robots performing within industrial settings, hazardous environments, and in exploration.

Electronic access

The research register for this journal is available at
http://www.emeraldinsight.com/researchregisters

The current issue and full text archive of this journal is available at
http://www.emeraldinsight.com/0143-991X.htm

Industrial Robot: An International Journal
Volume 29 · Number 3 · 2002 · pp. 259–267
© MCB UP Limited · ISSN 0143-991X
DOI 10.1108/01439910210425568

1. Introduction

The National Institute of Standards and Technology (NIST) has been collaborating with other government agencies and university researchers to develop methods of evaluating and measuring the performance of robotic and other intelligent systems. The community agrees that it would benefit from having uniform, reproducible means of measuring capabilities of their systems to evaluate which approaches are superior under which circumstances, and to help communicate results. One of the efforts in the performance metrics program at NIST is the creation of reference test arenas for mobile autonomous robots. The first set of arenas was targeted at the Urban Search and Rescue (USAR) application and was designed to represent, at varying degrees of verisimilitude, a collapsed building. This is a domain that is very dangerous for human rescuers, poses an almost infinitely difficult spectrum of challenges, and yet provides an opportunity for robots to play a pivotal support role in helping to save lives (Blitch, 1996). Developed with sponsorship by the Defense Advanced Research Projects Agency (DARPA), the NIST arenas were first deployed at the American Association for Artificial Intelligence (AAAI) Mobile Robot Competition in 2000. In 2001, the arenas were again used at the International Joint Conference on Artificial Intelligence (IJCAI). The RoboCupRescue League, emerging under the RoboCup umbrella (Kitano *et al.*, 1997), will henceforth be using the arenas to host their physical robot rescue league competitions. A discussion of the details of competitions is contained in Section 3 of this paper.

There are three main sets of customers for the arenas. The first are researchers, who need testing opportunities. The repeatable obstacles (sensory and physical) that are focused toward mobile robotic perception and intelligent behavior provide them with challenges for their robots. The second are the sponsors of research. They can use the arenas for validation exercises to objectively evaluate robots in structured, repeatable, representative environments. The arenas can be used to validate robotic purchases, identify strengths and weaknesses in systems, and compare the cost effectiveness of different approaches. Finally, the end users of the robots can benefit from the resulting performance metrics. Our eventual goal is to develop standard performance metrics through our experiences with the arenas, which can be used by mobile robot purchasers to specify thresholds of performance required to meet their application needs.

There were several motivating factors for taking a testbed approach to measuring performance. The first was the desire to be able to compare "apples to apples." When researchers publish results, they describe performance of their systems in their laboratory, making it impossible to compare and contrast with others' results. Isolating tests for sensing, behaviors, and other robotic capabilities and making these tests reproducible allows the community to make meaningful comparisons of algorithms, sensors, platforms, and other independent items. A standardization of the challenges, through use of the arenas, enables a direct comparison of approaches.

A second desire was being able to "teach to the test." The arenas provide an objective set of measures for evaluating different robotic implementations. The arenas are not idealized "blocks world" tests. They provide some fairly realistic challenges that mobile robots should be able to address. We hasten to add that the USAR domain is extremely challenging. Although the arenas do provide some elements of what may be encountered in a collapsed building, they are not representative of the reality of a disaster scene, but an abstraction thereof.

Another concern of research sponsors and of researchers themselves is the slowing of progress due to re-invention of the wheel. When building a robot, numerous hardware and software subsystems are required and it is very difficult (or not possible) to reuse any work done by other organizations. By highlighting successful approaches and demonstrating clear effectiveness against well-known obstacles, we hope that others will adopt them and expedite their progress in other groundbreaking areas.

Finally, practice makes perfect: arenas that are available to researchers year-round should enable them to repeat experiments and therefore debug and improve their systems. The arenas are set up in Gaithersburg, Maryland (near NIST) and can be used by researchers at any time. Robustness comes through repetition and testing outside perceived limits. The three arenas provide increasing levels of difficulty, so the researchers can move on to new challenges once they master the simpler sections.

2. Design considerations

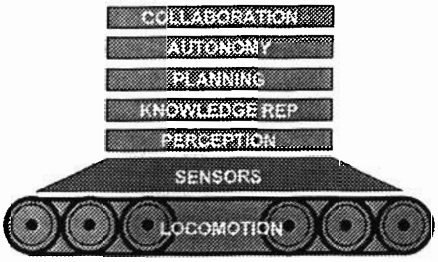
Elements of mobile robot autonomy

The primary goal of the test arenas is to provide reproducible measurements and tests of mobile robot capabilities and behaviors. There are many components that come together to create a fully autonomous mobile robot. Recognizing that there are going to be different degrees of autonomy implemented, the arenas attempt to isolate the various elements of autonomy and test the capabilities that can be demonstrated by any given robot. The targeted elements of robot autonomy are shown graphically in Figure 1. We believe these elements are fundamental for autonomous mobile robots, regardless of the application domain. For a more in-depth discussion of the design considerations for the arenas, see Jacoff *et al.* (2000).

At the lowest level is the locomotion capability of the robot's physical platform. Although two of the three arenas provide challenges for locomotion and agility of the robots, our emphasis (and that of the AAAI competitions) is on algorithms.

The next higher level is sensory perception. The robots have to sense what is in their environment in order to navigate in it, detect hazards, and identify goals such as exits and simulated victims. Sensor fusion is an important capability, as no single sensor will be able to identify or classify all aspects of the arenas. The simulated victims, analogous to the cheese for the mouse in a maze, provide incentive for the robots to investigate every corner of the arena. These simulated victims are represented by a collection of different

Figure 1 Elements of mobile robot autonomy



sensory signatures. They have shape and color characteristics that look like human figures and clothing. They emit a consistent heat signature, just as humans do. Some simulated victims have motions such as waving, and some emit sounds such as low moans, calls for help, or simple tapping. Any or all of these signs of life should be detected, identified, investigated further, and if confirmed as a victim, the location should be mapped.

The arenas are designed to pose challenges to typical robotic navigation sensors. For example, acoustic-absorbing materials confuse sonar sensors. Laser sensors have difficulty with clear materials such as windows. Highly regular striped wallpaper and other types of materials pose challenges to stereo vision algorithms. Compliant objects that may visually look like obstacles require the robots to apply tactile sensors or other means of verifying that they can indeed push them aside (e.g. curtains in doorways).

Localization of the robot through position sensors is also tested. Different flooring materials affect localization schemes based on wheel encoders. Ideally, the robot will use additional cues from the environment to help localize its position and maintain correct maps.

Knowledge representation is the next level. It encompasses the robot's ability to model the world, using both a priori information (such as might be needed to recognize certain objects in an environment) and newly acquired information (obtained through sensing the environment as it explores). In the mobile robot competitions for AAAI and RoboCupRescue, the robots are expected to communicate to humans the location of victims and hazards. Ideally, they would provide humans with a map of the environment they have explored, with the simulated victim and hazard locations clearly identified. The environment that the robots operate in is three-dimensional, hence they should reason in and be able to map in three dimensions. The arenas may change dynamically during a competition (as a building might further collapse while rescuers are searching for victims). Therefore, the ability to create, maintain, and use, maps to find alternate routes is important.

The planning or behavior generation elements of the robots build on the knowledge representation and the sensing elements. The robots must be able to navigate around obstacles, make progress in their mission (that is to explore as much as possible of the arenas and find simulated victims), take into account time as a limited resource, and make time critical decisions and tradeoffs. The planner should make use of an internal map generated by the robot and find alternative routes to exit the arenas that may be quicker or avoid areas that have become no longer traversible.

The overall autonomy of the robot is the next element to be evaluated. The robots are designed to operate with humans. The level of interaction may vary significantly, depending on the robot's design and capabilities, or on the circumstances. The intent is to allow for "mixed initiative" modes to limit human interaction and maximize the effectiveness and efficiency of the collaboration between robot and humans. Robots may communicate back to humans to request decisions, but should provide the human with meaningful communication of the situation. Pure teleoperation is not a desirable mode for the robot's operation. The human should provide the robot with high level commands, such as "go to the room on the left" rather than joystick the robot in that direction.

The final element to be evaluated in the robot's overall capabilities is collaboration among teams of robots. One very rich area of research is in cooperative and collaborative robotics. Multiple robots, either homogeneous or heterogeneous in design and capabilities, should be able to more quickly explore the arenas and find the victims. The issues to be examined are how effectively they maximize coverage given multiple robots, whether redundancy is an advantage, and whether or how they communicate among themselves to assign responsibilities. The human may make the decisions about assignments for each robot a priori, but that would not be as desirable as seeing the robots jointly decide how to attack the problem when confronted in the field.

A continuum of challenges

There are three separate indoor arenas focused toward the Urban Search and Rescue domain, each labeled by a color. A graphic of all three arenas assembled together is shown in Plate 1.

The yellow arena is the easiest in terms of traversability. Researchers, who may not have very agile robot platforms, yet want to test their sensing, mapping, or planning

Plate 1 Model of the three reference test arenas for autonomous mobile robots

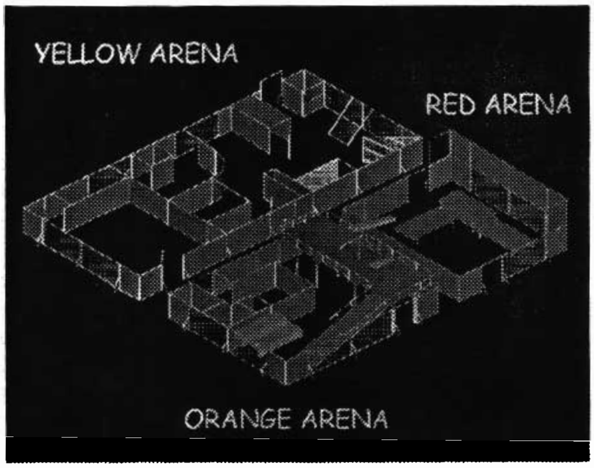
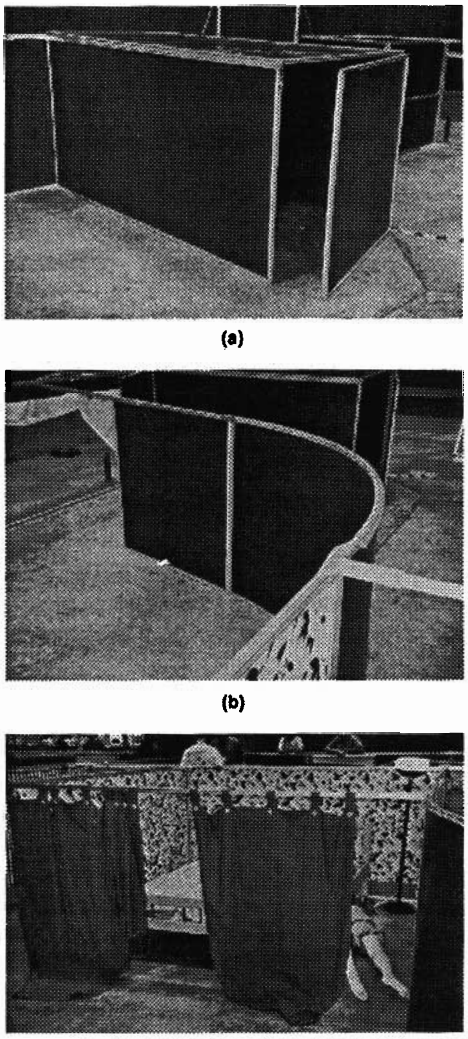


Plate 2 (a) Darkened chamber with door in the yellow arena; (b) cu wall in the yellow arena; (c) soft material, simulated victim under be yellow arena



algorithms, can use the yellow arena only. The arena consists of a planar maze. There are isolated sensor tests, based on obstacles or simulated victims. The arena is reconfigurable in real time, with, for instance, doors that can be closed and blinds that can be raised or lowered. The reconfigurability provides challenges to the mapping and planning algorithms of the robots. Some features of the Yellow arena are shown in Plate 2.

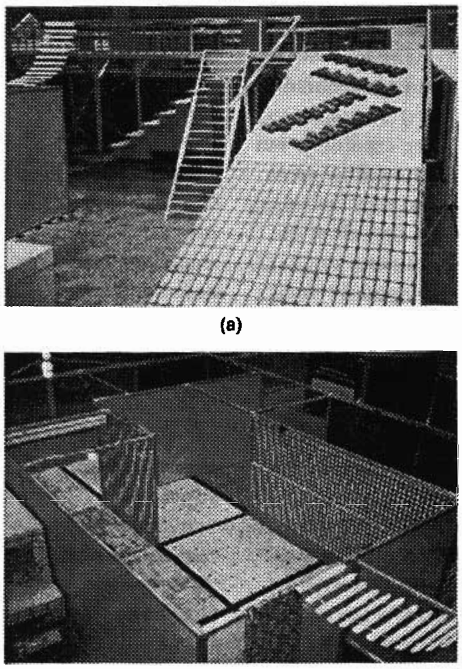
The orange arena provides traversability challenges. Different types of flooring materials are present and there is an elevated floor section, reachable via ramp, stairs, and ladders. There are holes in the second story floors. The mapping and planning capabilities of the robot must therefore be able to consider a three-dimensional world. The orange arena is also reconfigurable in real time. Some features of the orange arena are shown in Plate 3.

The red arena provides the least structure and the most challenges. It is very difficult to traverse, with debris of various sorts throughout the arena. The debris is very problematic for most robot locomotion mechanisms and includes steel rebar, gravel, plastic bags, and thin pipes. Simulated rubble resembling cinder blocks is strewn throughout. There are simulated pancaked floors (floors collapsing onto lower floors). The flooring in certain sections is unstable and will collapse if a heavy robot moves onto it. A robot must determine the stability of these flooring sections prior to applying weight to the structure, which is an advanced sensory challenge. A view of the red arena is shown in Plate 4.

3. The 2001 competitions

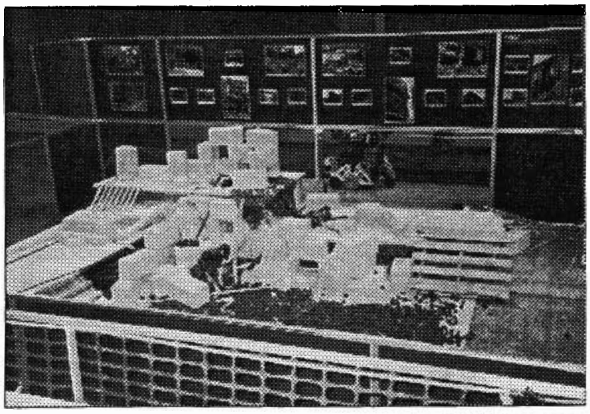
The NIST arenas made their debut at the 2000 AAAI Mobile Robot Competition (Murphy *et al.*, 2000; Schultz, 2001). Their second deployment was at the 2001 IJCAI, where the RoboCup Robot Rescue and AAAI Mobile Robot Rescue competitions were jointly held.

Plate 3 (a) Ramp and other routes to the elevated floor section in the orange arena; (b) different flooring materials within the maze section in the orange arena


(a)

(b)

Plate 4 Simulated rubble pile and collapsed flooring in the red arena



For the second competition, a rules committee developed a coherent scoring formula and other rules. The competition rules were designed to produce a final scoring distribution that defines clear winners. The focus of the competition was on robot intelligence; hence the scoring system had to favor solutions that demonstrated on-board perception, world modeling, planning, autonomy and collaboration.

Scoring was biased toward high quality interactions with humans, meaning that there would be low-bandwidth, high content, infrequent communications to and from humans. The robots were expected to produce human-understandable maps of their findings. Teleoperation, where the human was doing the perception and making the navigation decisions, was penalized. This control scheme is contrary to the goals of the AAAI, which are to push the state of the art toward full autonomy. Since this is an ambitious goal, we are accepting mixed initiative approaches as near term solutions. But since it is very difficult for humans to gain enough situational awareness through remote sensory feedback to effectively steer a robot in unknown and difficult surroundings, such as those of a collapsed building, we hope to encourage bounded autonomous, and eventually fully autonomous, solutions in the near future.

Disincentives were built into the scoring to discourage other undesirable traits in the robots. For example, using simple redundancy in robots, without demonstrating clear collaboration among the robots or benefits due to heterogeneous robots, was clearly discouraged in the scoring formula. The goal was to reward innovation and robustness in particular designs while allowing use of multiple robot teams to explore larger areas. Teleoperation, as discussed above, was also discouraged in the scoring formula. In particular, teams were penalized when they could not demonstrate a sufficient level of robot autonomy to allow one operator to control multiple robots simultaneously using higher level commands.

Other considerations in the design of the scoring were reflective of the course's design. "Gaming" of the arena, that is, learning the course layout at competition time and programming that layout into the robot in order to perform well, is obviously undesirable from a competitive standpoint. Generally, rearranging the location of simulated victims or closing doors between rounds is sufficient to thwart these efforts. Since there were some fairly easy simulated victims to find, a minimum score was required to qualify for one of the place awards. The scoring formula also was weighted to reflect the increasing difficulty of navigating and searching each progressively more challenging arena.

Six teams registered for the competition, but only four actually competed due to hardware or other problems. No team scored enough points to qualify for either first, second, or third place awards. The two most successful teams earned "qualitative" awards for demonstrating very different capabilities.

Swarthmore College (USA) demonstrated the most artificial intelligence capability, but they only navigated within the yellow arena. The scoring formula required that the robots confined to the yellow arena find all of the victims to earn the minimum score to qualify for a place award and be competitive with robots entering the other two more difficult arenas. They came close, finding all but one of the victims during one of their runs.

Sharif University (Iran) demonstrated a more rugged, skid-steered, tracked robot, and attempted to negotiate the red arena. However, they had problems with their control strategy, bumping walls and obstacles frequently. They even collapsed the pancaked flooring in the red arena. They resorted to identifying victims from outside the arena, but suffered from inherent victim location inaccuracies with this approach. They also required too many human operators to score well, keeping them from earning a place award.

4. Toward more capable industrial mobile robots

Our approach toward developing useful performance metrics through task-based performance testing is still in the early stages. No definitive metrics have been identified and rigorously tested thus far. We are actively promoting the arenas to increase the number of observable entrants. And we are developing a robot tracking system to collect objective position data while the robots are in the arenas. To this point, our observations and video capture of robot entrants interacting with the reference obstacles is still somewhat anecdotal. However, it appears evident that certain robot capabilities currently under development, when fully realized, will be particularly useful for performing industrial tasks such as automated inspection and material handling. As discussed earlier, identifying a "best in class" solution for a given set of arena obstacles will be the intended way to transfer successful implementations (e.g. algorithms, sensors, or mechanisms) from evaluation demonstrations to industrial applications.

Recognizing and negotiating obstacles

As the automatically guided vehicle (AGV) industry moves away from wire guided systems in favor of more intelligent, reconfigurable systems, they rely much more heavily on sensor fusion to perform obstacle detection and avoidance (Bostelman *et al.*, 2001). All AGV systems found in industry are required to sense obstacles in their path and stop if contacted. The question is how do they recover from their stopped state. Most AGVs must be reset manually, which is tedious. The more advanced AGV systems allow vehicles to pass stopped vehicles when necessary. But they often receive their commands from a centralized controller sensing their respective positions via laser guidance systems. Very little sensory processing is done onboard the vehicles, and no vehicle to vehicle communication typically takes place. If the obstacle type is unknown, even these AGVs are often incapacitated. The advent of light/laser direction and range (LIDAR/LADAR) sensors as proximity bumpers on some advanced AGVs (which are ubiquitous in research oriented mobile robot implementations) shows the willingness of these manufacturers to adopt advanced capabilities when practical.

As the AGVs become more sensor laden, and their onboard processing capabilities necessarily increase, they will evolve from guided vehicles into functional robots capable of making decisions on their own. At that point, they will be ready to incorporate any number sensory processing capabilities found to be successful in our reference test arenas. For example, sensing the thermal signatures of humans, a capability that is virtually required to successfully identify all the simulated victims in our arenas, may initiate a variety of behaviors in industrial mobile robots that would greatly improve their effectiveness in negotiating busy human workplaces. Also, a robot's ability to model and plan its way through ever-changing environments, specifically tested in our arenas, is crucial to improving robot performance in complicated industrial settings. Our arenas seek to isolate and test such capabilities, allowing robots to demonstrate their ability to recognize, negotiate, and map obstacles in their environment. Doing so repeatedly will demonstrate robustness and reliability of particular implementations.

Communication with humans and robots

Some existing service robots, which perform material transport in hospitals, use sophisticated sensor fusion to successfully negotiate ever-changing hallways full of obstacles and people (Evans and Krishnamurthy, 1998). They even communicate with people not familiar with such interactions. These robots provide relatively high level communications to their human handlers (or human obstacles) in the form of voice recorded status messages or prompts. The advent of voice recognition to allow verbal re-tasking of robots, and speech generation to answer human queries in their natural language, will provide a marked improvement in acceptance of these robots. These are the kind of high level communications encouraged (through scoring incentives) during competitions in our arenas. The largest scoring incentive by far rewards a single "manager" (rather than "operator") who can simultaneously control several robots, clearly improving on teleoperative techniques (one operator to one robot ratio, or worse). No particular preference is given for using speech, hand motions, or other communication methods during competition. Although at the end of a run, the robot must produce human readable maps of the environment to evaluate certain aspects of its performance. In general, high-level, intermittent communications between humans and robots are clearly encouraged.

Additionally, active collaboration between robots, without direct human intervention, can greatly improve robot performance and is encouraged in our arenas. Sharing maps, collaborating on plans, and communicating hazards allow teams of robots working under a time limit to search larger areas, reduce redundant searches, and avoid dangerous areas. Communication among autonomous robots, just as communication among humans, is essential for improving efficiency across large, multi-robot, industrial applications where robot to robot encounters happen often and can lead to bottlenecks in flow.

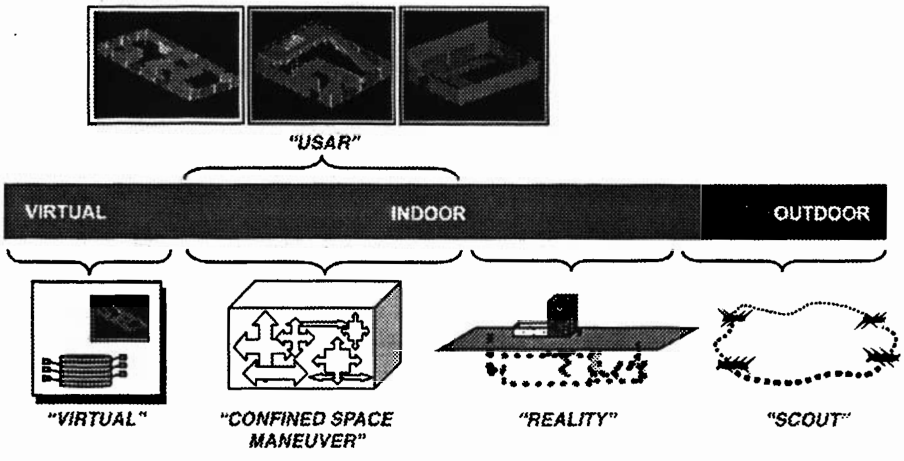
5. Future activities

The NIST reference test arenas will continue to host the AAAI Mobile Robot Competitions and the RoboCupRescue league competitions in 2002. Since no teams have demonstrated the minimum capabilities necessary to earn a place award, the arena configuration will stay relatively static for the time being. But as soon as the robots begin demonstrating repeated successes against the challenges posed, we will elevate the difficulty factor to stay modestly out in front. Our goal is to balance two factors: allow the researchers to see the "next step" goals to design toward while not demoralizing the researchers and students with terrain that is too difficult. Both the AAAI and RoboCupRescue competitions will use the same scoring system. Replicas of the arenas will be built for each RoboCupRescue event and left in the host country. This will result in dissemination of the arenas worldwide and enable researchers to have access to them for year round robot development and practice.

In an effort to augment our approach toward standardized tests and to further disseminate the arena's existing challenges, NIST is developing virtual versions of the indoor arenas (see Figure 2). The effort is two-fold. Initially, sensor data sets captured systematically within the actual arenas will be organized and made available on the Web. This will allow researchers the ability to process real sensor data during development of their algorithms, while freeing them from the costs and maintenance associated with real robot hardware. Data from a range-imaging sensor (LADAR) and from a digital color camera will be the first datasets available. This virtual arena should help quicken the pace of software development while maintaining a direct correlation to the real world, representative environment. Once the software is tested and deemed likely to succeed, it can be downloaded to an actual robot equipped with the exact same sensors used to capture the original data set, and allowed to run the actual arenas. We expect to begin hardware independent competitions such as this in association with the Performance Metrics for Intelligent Systems (PerMIS) conferences to be held at NIST in August 2002.

A second, more ambitious effort involves creating a simulated version of the course into

Figure 2 The spectrum of testing environments envisioned



which teams can plug their algorithms, receive simulated sensor data, and send actuation commands to navigate through the arenas. Interaction with the research community is needed to design and develop this environment.

In an effort to develop generalized performance metrics for a variety of autonomous mobile robots and applications, we are considering development of additional arenas. Beyond the existing indoor arenas and the virtual arenas under development, we are considering development of an outdoor arena, or "scout" arena. This equestrian style arena would consist of a collection of sensory obstacles, connected by paths or roads, spanning a relatively large outdoor space (>1km). These sensory obstacles would consist of challenges found in natural settings such as tall grass (with and without unseen rocks and logs), transitions from clearing to forest, negative obstacles (holes), standing water, and the like. These obstacles would also begin to isolate certain urban situations such as fences, signs, curbs, buildings, doors and stairs. This scout arena would allow repeatable testing of autonomous behaviors in outdoor environments.

6. Conlusions

Tangible, realistic challenge problems can provide researchers with direction and focus their efforts. Reproducible and widely known challenges can help evolving fields by providing reference problems with measures of performance. That is the intent of our reference test arenas for autonomous mobile robots. Robot competitions and public demonstrations at international events such as AAAI, IJCAI, RoboCupRescue, and others, can be valuable in spurring advancements in robotic abilities by raising researchers' awareness of the types of challenges that they must confront. Successful implementations against widely known obstacles can achieve a "best in class" status for sensors, algorithms, and mechanisms. This, along with a greater understanding of the capabilities demonstrated through wider familiarity with the standardized test, will hopefully lead to more re-use of algorithms and approaches. The end result: more effort allocated toward enhancing existing, successful implementations rather than "recreating the wheel." In fact, the least successful implementations each year should be as capable as the "best in class" from the previous year. So each successive year should produce new robot capabilities far surpassing the previous lot. Thus the pace of improvement should increase quickly.

As these enhanced robotic capabilities emerge, are proven in competition, demonstrated in representative environments, and challenged beyond their intended domains, the fruits of this research will almost certainly transfer into productive applications. Industrial autonomous vehicles performing inspection or material handling tasks should begin to exhibit the prior year's advances;

robust sensory perception, reactive planning, sophisticated mapping of complex environments, and interactive communication capabilities among robots and with humans. Teams of exploration robots will be able to effectively collaborate in hazardous or remote settings with only sporadic guidance from their human handlers. A productive surge in robotic applications will finally find their way into our everyday lives, as was always promised. We believe that for these eventualities to become realities in the near term, we must begin with standardized performance tests, quantitative performance measures, and objective performance evaluations, to help guide our research efforts, focus our innovations, and facilitate collaboration among the research community.

References

Blitch, J.G. (1996), "Artificial intelligence technologies for robot assisted urban search and rescue", *Expert Systems with Applications*, Vol. 11 No. 2, pp. 109-24.

Bostelman, R., Juberts, M., Szabo, S., Bunch, R. and Evans, J. (2001), "Industrial autonomous vehicle project report", *NIST Internal Report No. 6751*, June.

Evans, J. and Krishnamurthy, B. (1998), "Helpmate, the trackless robotic courier: a perspective on the development of a commercial autonomous mobile robot", *Autonomous Robotic Systems.*

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

Kitano, H., Asada, M., Kuniyoshi, Y., Noda, I., Osawa, E. and Matsubara, H. (1997), "RoboCup: a challenge problem for AI", *AI Magazine*, Vol. 18 No. 1, Spring, pp. 87-101.

Murphy, R., Casper, J., Micire, M. and Hyams, J. (2000), "Assessment of the NIST standard test bed for urban search and rescue competitions", *Proceedings of the 2000 Performance Metrics for Intelligent Systems Workshop*, August, Gaithersburg, MD.

Schultz, A. (2001), "The 2000 AAAI mobile robot competition and exhibition", *AI Magazine*, Vol. 22 No. 1, AAAI, Spring, pp. 67-72.