Proceedings of the Applied Imagery Pattern Recognition (AIPR) Workshop, Washington, DC, October, 11-13, 2006.

Performance Metrics for Intelligent Systems (PerMIS) 2006 Workshop: Summary and Review

R. Madhavan and E. Messina
Intelligent Systems Division
National Institute of Standards and Technology
Gaithersburg, MD 20899-8230.
Email: $raj.madhavan@ieee.org$, $elena.messina@nist.gov$

Abstract

The Performance Metrics for Intelligent Systems (PerMIS 2006) workshop was held during August 21-23, 2006 at the National Institute of Standards and Technology (NIST) in Gaithersburg, Maryland, USA. The PerMIS series (the current workshop is the sixth) is targeted at defining measures and methodologies of evaluating performance of intelligent systems. PerMIS 2006 focused on applications of performance measures to practical problems in commercial, industrial, homeland security, and military applications. An important element of overall performance evaluation is that of assessing the technical maturity of a given technology or system. One approach for accomplishing this is known as Technology Readiness Level (TRL) assesment. TRL evaluations have been the focus of past PerMIS workshops and continue to be a foundational theme. This paper will provide an overview of the workshop and various topics that are closely related to the theme of the AIPR workshop.

1. Introduction

The number of development and deployment programs for advanced systems that are designed to work remotely, independently, or even in some cases, autonomously, is ever-increasing. This puts a greater priority on the assessment of the viability of systems and their components that enable "intelligent" behavior. There is a range of evaluation approaches that can be considered for a given algorithm, component, or system. The National Institute of Standards and Technology (NIST), as part of its core mission of advancing measurement science, standards, and technology in ways that enhance economic security and improve our quality of life, is working on facilitating and advancing the state of the science in evaluation of intelligent systems. In the many programs that it carries out, NIST is working with partners in industry, academia, and other government agencies to develop metrics, test methods, and supporting artifacts, as well as a foundational science of performance evaluation for intelligent systems. An annual workshop has been dedicated to bringing together the community to share progress and ideas on these very topics. Begun in 2000, the Performance Metrics for Intelligent Systems (PerMIS) workshops have become a forum for exchanging ideas and results among researchers, developers, and managers of intelligent systems programs.

Although there are myriads of proposed approaches that can be taken for measuring capabilities with regard to intelligent, independent, or autonomous behavior, when a particular component or system is being considered for fielding, an important assessment is that of its technological maturity. A commonly-used scale for this is the Technology Readiness Level (TRL), developed by the National Aeronautics and Space Administration (NASA) in 1995 [6]. It serves as a maturity assessment of evolving technologies (materials, components, devices, etc.) prior to incorporating that technology into a system or a subsystem. TRL can therefore serve as an important performance measurement technique for the intelligent systems community [3, 8, 2].

Originally designed to validate the readiness of equipment that would eventually be fielded in outer space, this 9 level scale has been adopted by the Department of Defense and is mandated for all major acquisition programs [1]. Adapting the scale to use in measuring the maturity of intelligent systems technologies is an interesting challenge. The Army Research Laboratory (ARL) conducted a TRL evaluation in 2002-2003 to see whether the Demo III Autonomous Navigation System (ANS) had attained TRL 6 (defined as "demonstrated...in a relevant environment") [4].

The ANS was a subsystem resident on an unmanned ground vehicle that was able to navigate between commanded waypoints while avoiding obstacles of various sorts. The missions were carried out in three main types of terrain: gently rolling vegetated, rolling arid, and urban.

Subject matter experts provided course waypoints based on aerial maps of the regions. Once the courses were defined, missions were formed with randomized start points, direction, and lengths. The number of missions that were run was designed to be statistically significant. A total of 646 missions, covering 559.70 km, were run. The key metrics were whether the missions were completed, the distance and time during which the vehicle operated autonomously, and the corresponding percentages. If a mission was not completed, the various possible causes of its premature termination were noted. More details are contained in the final report.

Past PerMIS workshops (for example, 2003 [7]) have in fact had TRL as a theme and included dedicated sessions on TRL evaluation, such as of that carried out for ARL on autonomous navigation systems for unmanned ground vehicles. In 2006, there were no sessions explicitly focusing on this topic, but there was a recurring emphasis on assessing systems that were closer to fielding (or already fielded). PerMIS attendees were given the opportunity to observe a Response Robots Evaluation Exercise, held at a Fire Rescue Training Facility, which would qualify as a "representative environment" that robots for urban search and rescue (US&R) operations could be expected to confront. One important aspect of these Response Robot exercises is to expose the robots (and their developers) to representative environments. Although not providing formal measures of TRL and lacking the statistically significant experiments for data capture, these exercises do provide a bridge to the full-blown TRL exercises which are a necessity along the path to succesfully fielding robots to assist responders in US&R missions.

Figure 1 shows TRL levels from an intelligent systems perspective. This table is adapted from the PerMIS 2003 White Paper, which was entitled "Performance Measures for Intelligent Systems: Measures of Technology Readiness" [3].

2. PerMIS 2006

The PerMIS 2006 workshop was held at NIST from August 21-23, 2006 [5]. The authors served as the program and the general chairs, respectively. Sixth in a series of workshops since 2000, PerMIS is targeted at defining measures and methodologies of evaluating performance of intelligent systems. PerMIS 2006 focused on applications of performance measures to practical problems in commercial, industrial, homeland security, and military applications. In 2006, PerMIS was co-located with the Institute of Electrical and Electronics Engineers (IEEE) Safety, Security, and Rescue Robotics (SSRR) workshop (at the same venue from August 22-24, 2006).

The PerMIS and SSRR workshops were complemented by a series of related events. There was a response robot evaluation exercise held August 19-21 at a nearby responder training facility. A demonstration by local bomb squads using robots on the NIST campus took place on August 23rd (See Section 3 for more details). PerMIS attendees also benefitted from robot and technology exhibits during the workshops.

Figure 2 shows the PerMIS program at a glance. The workshop included several plenary addresses and featured presentations, a panel discussion and robot events and exercises in addition to the regular technical sessions. Plenary addresses were each an hour long and the featured presentations were 45 minutes each. The technical session papers were presented in a 25 minute format (20 minutes talk and 5 minutes discussion).

2.1 Plenary Addresses & Featured Presentations

On the first day of the workshop, Prof. Henrik Christensen (Georgia Institute of Technology, USA/Royal Institute of Technology (KTH), Sweden) delivered a plenary address entitled *Evaluation of Robots for Human-Robot Interaction*. The second day witnessed two plenary addresses by Prof. Shigeo Hirose (Tokyo Institute of Technology, Japan) and Prof. Hugh Durrant-Whyte (The University of Sydney, Australia), respectively. Prof. Hirose's address discussed *Development of Rescue and Demining Robots*. Prof. Durrant-Whyte's address was on *Maximal Information Systems*.

The final day of PerMIS was a veritable intellectual feast book-ended by two plenary addresses. Dr. Martin Buehler (Boston Dynamics, USA) kicked off the day with a presentation on Developing *Dynamic Legged Robots - Towards Greater Mobility Without Falling Over*. Dr. James Albus (NIST) concluded the day - and the workshop - with a banquet address on *Building Brains for Thinking Machines*. The day also included two featured presentations by Mr. Chuck Shoemaker (Robotic Research, LLC and formerly with the Army Research Lab., USA): *Army Autonomous Tactical Unmanned Ground Vehicles* and Dr. Mike Montemerlo (Stanford University, USA): *Winning the DARPA Grand Challenge* in addition to an *Emergency Responder Panel Discussion* moderated by Prof. G. Kemble Bennett (Texas A & M, USA) with the participation of US&R responders from several Federal Emergency Management Agency (FEMA) task forces.

2.2 Technical Sessions

MON-AM1 Autonomy and Intelligence

The first technical session kicked off with an invited talk by Dr. Gary Berg-Cross entitled *Improving Knowledge for Intelligent Agents: Exploring Parallels in Ontological Anal-*

Technology Readiness Level	Description
1. Basic principles and broad vision of the system observed and reported	The most general discussion of the system, i.e. the lowest level of resolution in system analysis. It corresponds to the lowest level of technology readiness. The results of this level of analysis are usually presented as paper studies of a system's basic properties. Correspondingly, it is also the lowest level of software readiness. Basic research begins to be translated into applied research and development.
2. Conceptual design of a system and/or technology and its application formulated	Beginning of the system's refinement: resolution grows. Key engineering solutions are proposed, innovations are introduced, key resource limits are chosen. Practical applications are invented and tested. Applications are partially tested, partially hypothesized, and there may be no exhaustive proof or reliable analysis to support the assumptions and visions of the developing team.
3. Thorough theoretical and experimental critical analysis of system's function; detailed characteristic proof of concept	More detail is addressed. Active research and development are initiated. Theoretical studies are conducted in the laboratory targeting physical and/or computational (simulation) validation of analytical predictions for separate sub-systems of the system. Those sub-systems are being scrutinized that are innovative and have not been integrated. Similar active research and development is initiated for the software subsystems. The number of resolution levels must be properly chosen. The programs are written that can validate theoretical predictions for separate software subsystems. Algorithms are tested in laboratory environment or in simulation.
4. Component and/or breadboard validation is conducted in the laboratory environment	All basic subsystems and components are integrated to establish that they will work together. This usually includes ad hoc sub-systems integration. This includes integration of software components are integrated to determine how they will work together. They are relatively primitive with regard to efficiency and reliability compared to the eventual system. System Software architecture development initiated to include interoperability, reliability, maintainability, extensibility, scalability, and security issues. At this point, we are able to check the matching between computational parameters of the algorithms and programs on one hand and the parameters of other components (sensors, actuators) on the other.
5. Component and/or breadboard validation in more realistic relevant environment	Fidelity of breadboard technology increases significantly. The basic technological components are integrated with reasonably realistic supporting elements: it includes "high fidelity" ("high resolution") laboratory integration of software components. Configuration control is initiated. Verification, Validation, and Accreditation (VV&A) initiated. At this point, we have an opportunity to check whether the state-space is tessellated properly, whether the parameters of sampling, or parameters of randomization are proper ones.
6. System/subsystem model or prototype demonstration in a relevant environment	Representative model or prototype system, which is well beyond that of TRL 5, is tested in a relevant environment. Represents a major step up in a technology's demonstrated readiness. Examples include testing a prototype in a high-fidelity laboratory environment or in a simulated operational environment. This stage represents a major step up in software demonstrated readiness. Software support structure is in development. VV&A is in process. At this stage we check the value of parameters such as carrying frequencies, bandwidths, etc.
7. System prototype demonstration in an operational environment	Prototype near, or at, planned operational system. Represents a major growth in resolution comparatively with TRL 6, requires demonstration of an actual system prototype in an operational environment. Examples include testing the prototype in a test bed aircraft. Software support structure is in place. Software releases are in distinct versions. Frequency and severity of software deficiency reports do not significantly degrade functionality or performance. VV&A completed.
8. Actual system completed and qualified through test and demonstration	The system has been proven to work in its final form and under expected conditions. In almost all cases, this TRL represents the end of the system development. Examples include developmental test and evaluation of the system in its intended application to determine if it meets design specifications. Software has been demonstrated to work in its final form and under expected conditions. In most cases, this TRL represents the end of system development. Examples include test and evaluation of the Software in its intended system to determine if it meets design specifications. Software deficiencies are rapidly resolved through support infrastructure.
9. Actual system proven through successful mission operations	Actual application of the technology in its final form and under mission conditions, such as those encountered in operational test and evaluation. Examples include using the system under operational mission conditions. Actual application of the Software in its final form and under mission conditions, such as those encountered in operational test and evaluation. In almost all cases, this is the end of the last debugging aspects of the system development. The system is used under operational mission conditions. Software releases are production versions and configuration controlled.

Figure 1. Technology Readiness Levels (TRL) in the context of intelligent systems performance.

<table>
<tr><th></th><th>Monday
August 21</th><th>Tuesday
August 22</th><th>Wednesday
August 23</th></tr>
<tr><td>8:00</td><td>Welcome to PerMIS</td><td></td><td></td></tr>
<tr><td>8:30</td><td>Plenary: Henrik Christensen</td><td>Plenary: Shigeo Hirose</td><td>Plenary: Martin Buehler</td></tr>
<tr><td>9:30</td><td>Coffee Break</td><td>Coffee Break</td><td>Coffee Break</td></tr>
<tr><td>10:00</td><td>Mon AM1: Autonomy and Intelligence</td><td rowspan="2">Tue AM1: DARPA ASSIST (Special Session))</td><td rowspan="2">Wed AM1: Autonomous Systems Evaluation: Testbeds and Tools</td></tr>
<tr><td></td><td>Mon AM2: Performance Metrics</td></tr>
<tr><td>13:00</td><td>Lunch</td><td>Lunch</td><td>Lunch</td></tr>
<tr><td>14:00</td><td rowspan="2">Mon PM1: Performance Evaluation</td><td>Plenary: Hugh Durrant-Whyte</td><td rowspan="2">Featured Presentations: Chuck Shoemaker and Mike Montemerlo</td></tr>
<tr><td>15:00</td><td>Coffee Break</td></tr>
<tr><td>15:30</td><td>Coffee Break</td><td rowspan="2">Tue PM1: Performance Analysis</td><td>Coffee Break</td></tr>
<tr><td>16:00</td><td rowspan="3">Travel to Montgomery County Fire Rescue Training Academy and Observe Responders Using Robots in US&R Operations</td><td>Responder Panel Discussion led by Prof. Kemble Bennett</td></tr>
<tr><td>17:00</td><td rowspan="2">Welcome Reception, Posters, Exhibits, Demonstrations</td><td rowspan="2">Robot Demonstrations by Bomb Squads</td></tr>
<tr><td>17:30</td></tr>
<tr><td>19:00</td><td></td><td></td><td>Banquet: James Albus, Speaker</td></tr>
</table>

Figure 2. PerMIS 2006 program at a glance.

ysis and Epigenetic Robotics. This session included talks that discussed robot autonomy and machine intelligence in a general sense.

MON-AM2 Performance Metrics

This session included an invited talk by Dr. Douglas Gage on *Meaningful Metrics and Evaluation of Embodied, Situated, and Taskable Systems*. The rest of the session saw the presentation of robot performance metrics for robots operating in a variety of environments.

MON-PM1 Performance Evaluation

In this session, the talks centered on evaluation of performance of various algorithms against ground truth and their ability to cope with external disturbances. Application domains were varied and included threat evaluation, urban search and rescue robots, integrated vehicle-based safety systems, and learning-based systems.

TUE-AM1 DARPA ASSIST Special Session

On Tuesday morning, a half-day special session was held focusing on the DARPA ASSIST (Advanced Soldier Information System and Technology) program. The goal of the ASSIST program is to exploit soldier-worn sensors to augment the soldier's recall and reporting capability to enhance situational understanding. NIST is serving as the independent evaluation team for this program, and the special session focused on the approaches that NIST took to evaluate the technology being developed. There were three presentations by research teams describing technologies that tracked soldiers movement (both indoors and outdoors), recognized objects in the environment, and characterized what action a soldier was performing. There were also five presentations by members of the evaluation team, describing technology tests, utility tests, ontology-related evaluation, and descriptions of what possible future evaluation efforts may look like.

TUE-PM1 Performance Analysis

Dr. Robert Finkelstein's talk *Memetics and Intelligent Systems* started off this session. The talks in this session analyzed the performance of robot and control algorithms across robot mobility and control system domains.

WED-AM1 Autonomous Systems Evaluation: Testbeds & Tools

Dr. David Sparrow gave an invited talk discussing the *Challenges in Autonomous System Development*. The rest of the session discussed various testbeds and tools (both virtual and real) for evaluating the performance of autonomous systems in general.

3. Robot Events and Exercises

NIST has been tasked by the Department of Homeland Security (DHS) to develop US&R robot performance standards. On the days preceding the workshop (August 19-20, 2006), a response robot informal evaluation was held at the Montgomery County Fire Rescue Training Academy in nearby Rockville, MD. These exercises for US&R teams introduce emerging robotic capabilities to emergency responders while educating robot developers regarding the necessary performance requirements to be effective, along with the associated environmental conditions and operational constraints to be useful. Standard test methods and usage guides for US&R robot performance are under development within the ASTM International E54.08 Subcommittee on Operational Equipment, which is under the Homeland Security Committee. These events help refine the proposed standard test methods and artifacts that developers can use to practice critical capabilities and measure performance in ways that are relevant to the end user, i.e. responders. These events are conducted in actual US&R training scenarios to help correlate the proposed standard test methods with envisioned deployment tasks and to lay the foundation for the usage guides, which will provide guidance on which robots are best suited for which response situations.

In the afternoon of August 21st, workshop attendees traveled to the nearby Maryland Fire and Rescue Training Academy to observe FEMA US&R task force members putting a wide variety of robots through their paces in operational scenarios, test methods, and radiation sensor integrations. DHS/FEMA US&R teams participated in this event. More information is available from `www.isd.mel.nist.gov/US&R_Robot_Standards`.

On the evening of the 22nd of August, the workshop attendees also saw demonstration of bomb disposal robots being operated by bomb squads from Maryland, Virginia, and Michigan, with emphasis on training procedures, performance test methods, operator interfaces, and deployment strategies.

4. Summary

PerMIS 2006 included thirty three regular presentations, four invited talks, two featured presentations, five plenary addresses and one panel discussion in addition to robot events, demos, and exercises Attendees of the workshop consisted of researchers, students, practitioners from industry, academia, and government. The workshop proved to be an excellent forum for discussions and partnerships, dissemination of ideas, and future collaborations

A special issue entitled *Quantitative Performance Evaluation of Robotic and Intelligent Systems* (edited by R. Madhavan, A. Jacoff, and E. Messina) consisting of selected papers from PerMIS 2006 and SSRR 2006 workshops will be published in a forthcoming volume in the Journal of Field Robotics. This volume is expected to be published in the first quarter of 2007.

References

[1] DoD Directive 5000.2-R, Mandatory Procedures for Major Defense Acquisition Programs (MDAPs) and Major Automated Information System (MAIS) Acquisition Programs, Mar 1996.

[2] PerMIS 2000 White Paper: Measuring Performance and Intelligence of Systems with Autonomy. In *Proceedings of the Performance Metrics for Intelligent Systems (PerMIS) Workshop, NIST Special Publication 970*, Aug 2000.

[3] PerMIS 2003 White Paper: Performance Measures for Intelligent Systems - Measures of Technology Readiness. In *Proceedings of the Performance Metrics for Intelligent Systems (PerMIS) Workshop, NIST Special Publication 1014*, 2003.

[4] R. Camden, J. Borenstein, F. French, C. Shoemaker, B. Bodt, S. Schipani, T. Runyon, A. Jacoff, and A. Lytle. Autonomous Mobility Technology Assessment Final Report. Technical Report ARL-TR-3471, Army Research Laboratory, 2005.

[5] R. Madhavan and E. Messina, editors. *Proceedings of the 2006 Performance Metrics for Intelligent Systems (PerMIS'06), NIST Special Publication 1062*, Aug 2006.

[6] J. Mankins. Technology Readiness Levels A White Paper, April 1995. http://www.hq.nasa.gov/office/codeq/trl/trl.pdf.

[7] E. Messina and A. Meystel, editors. *Measuring the Performance and Intelligence of Systems: Proceedings of the 2003 Performance Metrics for Intelligent Systems (PerMIS'03)*, NIST Special Publication 1014, Sep 2003.

[8] E. Messina, A. Meystel, and L. Reeker. PerMIS 2001 White Paper: Measuring Performance and Intelligence of Intelligent Systems. In *Proceedings of the Performance Metrics for Intelligent Systems (PerMIS) Workshop , NIST Special Publication 982*, 2001.