Performance Evaluation and Benchmarking of Intelligent Systems

Preface

To design and develop capable, dependable, and affordable intelligent systems, their performance must be measurable. Scientific methodologies for standardization and benchmarking are crucial for quantitatively evaluating the performance of emerging robotic and intelligent systems' technologies. There is currently no accepted standard for quantitatively measuring the performance of these systems against user-defined requirements; and furthermore, there is no consensus on what objective evaluation procedures need to be followed to understand the performance of these systems. The lack of reproducible and repeatable test methods has precluded researchers working towards a common goal from exchanging and communicating results, inter-comparing system performance, and leveraging previous work that could otherwise avoid duplication and expedite technology transfer. Currently, this lack of cohesion in the community hinders progress in many domains, such as manufacturing, service, healthcare, and security. By providing the research community with access to standardized tools, reference data sets, and open source libraries of solutions, researchers and consumers will be able to evaluate the cost and benefits associated with intelligent systems and associated technologies. In this vein, the edited book volume addresses performance evaluation and metrics for intelligent systems, in general, while emphasizing the need and solutions for standardized methods.

To the knowledge of the editors, there is not a single book on the market that is solely dedicated to the subject of performance evaluation and benchmarking of intelligent systems. Even books that address this topic do so only marginally or are out of date. The research work presented in this volume fills this void by drawing from the experiences and insights of experts gained both through theoretical development and practical implementation of intelligent systems in a variety of diverse application domains. The book presents a detailed and coherent picture of state-ofthe- art, recent developments, and further research areas in intelligent systems.

This edited book volume is a collection of expanded and revised papers presented at the 2008 Performance Metrics for Intelligent Systems (PerMIS'08: http://www.isd.mel.nist.gov/PerMIS_2008/) workshop held at the National Institute of Standards and Technology (NIST) from August 19–21, 2008. PerMIS is the only workshop of its kind dedicated to defining measures and methodologies of evaluating performance of intelligent systems. The Intelligent Systems Division of NIST, under the leadership of Dr. John Evans as Division Chief, initiated this series in 2000 to address the lack of measurements and basic understanding of the performance of intelligent systems. Prof. Alexander Meystel of Drexel University who had a long-standing collaboration with NIST's Intelligent Systems Division was a prime mover behind PerMIS from the very beginning. Dr. Meystel was instrumental in shaping the PerMIS workshops and their content for several years, with an emphasis on the theoretical foundations of the field of intelligent systems. Dr. James Albus, Senior NIST Fellow, provided valuable guidance throughout the workshop series, through his deep insights into intelligence. Over the years, the workshops have increased their focus on applications of performance measures to practical problems in commercial, industrial, homeland security, military, and space applications, while still retaining elements of theoretical examination. It has proved to be an excellent forum for discussions and partnerships, dissemination of ideas, and future collaborations between researchers, graduate students, and practitioners from industry, academia, and government agencies. Financial sponsorship has been primarily by NIST and Defense Advanced Research Projects Agency (DARPA). Additional support throughout the years in logistical terms has come, at various times, from the IEEE (originally known as the Institute of Electrical and Electronic Engineers, Inc.), IEEE Control Systems Society, IEEE Neural Net Council (which became the IEEE Computational Intelligence Society), IEEE Systems, Man, and Cybernetics Society, IEEE Robotics and Automation Society, the National Aeronautics

and Space Administration (NASA), and the Association for Computing Machinery (ACM).

In the years since its inception in 2000, the PerMIS workshop series has brought to light several key elements that are necessary for the development of the science and engineering of performance measurement for intelligent systems. The endeavor of measuring the performance of intelligent systems requires development of a framework that spans the theoretical foundations for performance measures to the pragmatic support for applied measurements. The framework would serve to guide development of specific performance metrics and benchmarks for individual domains and projects. The chapters in this book provide a broad overview of several of the different, yet necessary perspectives that are needed to attain a discipline for performance measurement of intelligent systems. If the field of intelligent systems is to be a true engineering or scientific endeavor, it cannot exist without quantitative measurements.

There exists a need to balance the desire for overarching theories and generality of measures and the pragmatic specificity required for applied evaluations. Similarly, there is a need for the measurement of performance of components and subsystems and for the overall integrated system. The papers that were selected from PerMIS'08 illustrate many of the dimensions of performance evaluation. Biologically-inspired measures are an example of foundational, overarching principles for the discipline. Areas covered in these selected papers include a broad range of applications, such as assistive robotics, planetary surveying, urban search and rescue, and line tracking for automotive assembly. Subsystems or components described in this book include human-robot interaction, multi-robot coordination, communications, perception, and mapping. In this emerging and challenging field of performance evaluation, supporting tools are essential to making progress. Chapters devoted to simulation support and open source software for cognitive platforms provide examples of the type of enabling underlying technologies that can help intelligent systems to propagate and increase in capabilities.

The edited book volume is primarily intended to be a collection of chapters written by experts in the field of intelligent systems. We envisage the book to serve as a professional reference for researchers and practitioners in the field and also for advanced courses for graduate level students and robotics professionals in a wide range of engineering and related disciplines including computer science, automotive, healthcare, manufacturing, and service robotics. The book is organized into 13 chapters. As noted earlier, these chapters are significantly expanded and revised versions of papers presented at PerMIS'08. Out of 58 papers presented at the 2008 workshop, these papers were selected based on feedback and input from the reviewers during the workshop paper acceptance process. A summary of the chapters follow:

- Multiagent systems (MAS) are becoming increasingly popular as a paradigm for constructing large-scale intelligent distributed systems. In "Metrics for Multiagent Systems", Robert Lass, Evan Sultanik, and William Regli provide an overview of MAS and of currently-used metrics for their evaluation. Two main classes of metrics are defined: Measures of Effectiveness (MoE), which quantify the system's ability to complete a task within a specific environment, and Measures of Performance (MoP), which measure quantitatively performance characteristics, such as resource usage, time to complete a task, or other relevant quantities. Metrics can be further classified by type of mathematical scale, namely, whether these quantities are defined as nominal, ordinal, intervals, or ratios. The authors classify a set of existing agent metrics according to whether they are MoE, MoP, type of mathematical formulation, their community of origin, and which layer they apply to within a reference model for agent systems. Using the defined metrics, the authors present a framework for determining how to select metrics for a new project, how to collect them, and ultimately analyze a system based on the chosen metrics and demonstrate this process through a case study of distributed constraint optimization algorithms.

- A list of evaluation criteria is developed in the chapter by Birsen Donmez, Patricia Pena and Mary Cummings entitled "Evaluation Criteria for Human-Automation Performance Metrics" to assess the quality of humanautomation performance to facilitate the selection of metrics for designing evaluation experiments in many domains such as human-robot interaction and medicine. The evaluation criteria for assessing the quality of a metric are based on five categories: experimental constraints, comprehensive understanding, construct validity, statistical efficiency, and measurement technique efficiency. Combining the criteria with a list of resulting metric costs and

benefits, the authors provide substantiating examples for evaluating different measures.

- Several areas of assistive robotic technologies are surveyed in the chapter "Performance Evaluation Methods for Assistive Robotic Technology" by Katherine Tsui, David Feil-Seifer, Maja Matari´c, and Holly Yanco to derive and demonstrate domain-specific means for evaluating the performance of such systems. Two cases studies are included to detail the development of performance metrics for an assistive robotic arm and for socially assistive robots. The authors provide guidelines on how to select performance measures for end-user evaluations of such assistive systems. It is their conclusion that end-user evaluations should focus equally on human performance measures as on system performance measures.

- Given that biological systems set the standard for intelligent systems, it is useful to examine how bio-inspired approaches can actually assist researchers in understanding which aspects derived from biology can enable desirable functionality, such as cognition. Gary Berg-Cross and Alexei Samsonovich discuss several schools of thought as applied to intelligent systems and how these approaches can aid in understanding what elemental properties enable adaptability, agility, and cognitive growth. High-level, "Theory-of-Mind" approaches are described, wherein a pre-existing cognitive framework exists. But what are the essential elements of cognition (what the authors term "critical mass")? "Cognitive decathlons" are appealing methods of trying to elucidate exactly what this critical mass is. An alternative view to the high level one is that of developmental robotics, which posits that cognitive processes emerge through learning and self-organization of many interacting sub-systems within an embodied entity. Developmental robotics offers the potential advantage of side-stepping the rigid engineering that may be present in top-down approaches. The authors conclude that there are yet pitfalls in the developmental approaches, including biases in the systems' design. They point the way forward by suggesting that developmental approaches should be guided by principles of how intelligence develops, but should not slavishly follow bio-inspiration on all fronts.

- In the chapter "Evaluating Situation Awareness of Autonomous Systems", Jan Gehrke argues that higher-level situation analysis and assessment are crucial for autonomous systems. The chapter provides a survey of situation awareness for autonomous systems by analyzing features and limitations of existing approaches and proposes a set of criteria to be satisfied by situation-aware agents. An included example ties together these ideas by providing initial results for evaluating such situation-aware systems.

- The role of simulation in predicting robot performance is addressed by Stephen Balakirsky, Stefano Carpin, George Dimitoglou, and Benjamin Balaguer in the chapter "From Simulation to Real Robots with Predictable Results: Methods and Examples". The authors reinforce the theoretical argument that algorithm development in simulation accelerates the motion planning development cycle for real robots. With a focus on model deficiencies as sources of simulation system brittleness, the chapter presents a methodology for simulation-based development of algorithms to be implemented on real robots including comparison of the simulated and physical robot performance. Examples using simulated robot sensor and mobility models are used to demonstrate the approach and associated techniques for validating simulation models and algorithms.

- The chapter "Cognitive Systems Platforms using Open Source" by Patrick Courtney, Olivier Michel, Angelo Cangelosi, Vadim Tikhanoff, Giorgio Metta, Lorenzo Natale, Francesco Nori and Serge Kernbach reports on various open source platforms that are being developed under the European Union Cognitive Systems program. In particular, significant research efforts in cognitive robotics with respect to benchmarking are described in detail.

- The chapter "Assessing Coordination Demand in Coordinating Robots" by Michael Lewis and JijunWang presents a means to characterize performance of the dual task of coordinating and operating multiple robots using a difficulty measure referred to as coordination demand. Several

approaches are presented for measuring and evaluating coordination demand in applications where an operator coordinates multiple robots to perform dependent tasks. Simulation experiments involving tasks related to construction and search and rescue reveal the utility of the measure for identifying abnormal control behaviors and facilitating operator performance diagnosis as well as aspects of multi-robot tasks that might benefit from automation.

- Intelligent systems such as mobile robots have to operate in demanding and complex environments. Receiving transmissions from human operators, other devices or robots and sending images or other data back are essential capabilities required by many robots that operate remotely. In their chapter "Measurements to Support Performance Evaluation of Wireless Communications in Tunnels for Urban Search and Rescue Robots", Kate Remley, George Hough, Galen Koepke, and Dennis Camell describe the complexities that confront the wireless communications systems for a robot that is used to explore unknown and difficult environments. They define the types of measures that can be taken to characterize the environment in which a robot must operate. The data can be used to help define reproducible test methods for the communications sub-systems and can enable modeling performance so as to allow a more advanced robot to compensate for degradation in its network by changing parameters or deploying repeaters. A detailed description of an experiment conducted in a subterranean tunnel illustrates the environment characterization and modeling process.

- In order for intelligent mobile robots to operate effectively and safely in the world, they need to be able to determine where they are with respect to their surroundings and to form a map of their environment, either for their own navigation purposes or, if exploring environments for humans, to transmit to their operators. Chris Scrapper, Raj Madhavan, Rolf Lakaemper, Andrea Censi, Afzal Godil, Asim Wagan, and Adam Jacoff discuss various approaches aimed at defining quantitative measures for localization and mapping in their chapter "Quantitative Assessment of Robot-Generated Maps". First off, physical test environments that abstract challenges that robots may encounter in real-world applications are described. Some theoretical approaches that leverage statistical analysis techniques to compare experimental results are discussed. Force Field Simulation is put forward as a means of evaluating the consistency of maps. Finally, quantitative measures of quality, based on features extracted by three algorithms—Harris corner detector, the Hough transform, and the Scale Invariant Feature Transform— are presented.

- Performance metrics for evaluating and prescribing approaches to surveying land areas on other planets using mobile robots are presented in the chapter by Edward Tunstel, John Dolan, Terrence Fong, and Debra Schreckenghost entitled "Mobile Robotic Surveying Performance for Planetary Surface Site Characterization". The authors apply a geometry-based area coverage metric to several different surveying approaches and assess trends in relative performance when applied to surveys of a common area. Limitations of the metric are highlighted to motivate the need for richer metrics that account for more than geometric attributes of the task. Examples of metrics that further account for system, environment, or mission attributes that impact performance are presented in the context of NASA research on human-supervised robotic surveying.

- The chapter "Performance Evaluation and Metrics for Perception in Intelligent Manufacturing" by Roger Eastman, Tsai Hong, Jane Shi, Tobias Hanning, Bala Muralikrishnan, Susan Young, and Tommy Chang summarizes contributions and results of a workshop special session. Various aspects of the general problem of benchmarking complex perception tasks in the context of intelligent manufacturing applications are covered ranging from camera calibration to pose estimation of objects moving in an environment with uncontrolled lighting and background. The authors discuss the interrelationships of the underlying approaches and bring them together to form a three-step framework for benchmarking perception algorithms and sensor systems.

- The final chapter provides a detailed description of how the results of performance evaluations can be used to understand where advanced technologies can be applied in industry. In "Quantification of Line Tracking Solutions for Automotive Applications", Jane Shi, Rick Rourke, Dave Groll, and Peter Tavora describe in depth experiments that their team conducted to quantify performance of

line tracking for automotive robotic assembly applications. They selected as performance metrics the range of relative positional tracking error between the robot and the vehicle body and the repeatability of the relative positional tracking error as measured in three standard deviations. These metrics provided insight in inter-comparing results from three line tracking solutions. Their analysis of system performance can lead to conclusions about where current technology is applicable and where there are technology gaps that must be addressed before robotic line assembly tracking is possible within the necessary tolerances.

It would be remiss on our part if we did not acknowledge the countless number of people who variously contributed to this effort. Firstly, we would like to express our sincere thanks to the authors of the chapters for reporting their thoughts and experiences related to their research and also for patiently addressing reviewers' comments and diligently adhering to the hectic deadlines to have the book sent to the publisher in a timely manner. We are indebted to the reviewers for providing insightful and thoughtful comments on the chapters which tremendously improved the quality of the expanded workshop papers to be included in this book. Even though not directly involved with the production of this book, we acknowledge the entire PerMIS'08 crew including program and organizing committee members, reviewers, and our local support staff who tirelessly ensure the success of PerMIS making it a pleasant and memorable experience for everyone involved. Our thanks are due to Springer for publishing this book and for accommodating us at various stages of the publication process. Lastly, but certainly not in the order of importance, we are grateful to our immediate family members who paid the "full price of the book" many times over!

We believe that this book is an important contribution to the community in assembling research work on performance evaluation and benchmarking of intelligent systems from various domains. It is our sincere hope that many more will join us in this time-critical endeavor. Happy reading!

Oak Ridge, TN — Raj Madhavan
Laurel, MD — Edward Tunstel
Gaithersburg, MD — Elena Messina

May 2009