Fostering Progress in Performance Evaluation and Benchmarking of Robotic and Automation Systems

by Fabio Bonsignorio, Elena Messina, and Angel P. del Pobil

We have shared benchmarks for many engineering systems and products in the market that can be used to compare solutions and systems. We can compare cars in terms of maximum speed, acceleration, and maximum torque; computers in terms of flops, random access memory, and hard disk capacity; and smartphones in terms of battery life and screen dimensions. We also have shared usability metrics based on human factors, which are used to compare the ease of use of different software interfaces. When we come to the evaluation and the comparison of how intelligent, robust, adaptive, and antifragile the behaviors of robots are in performing a given set of tasks, such as daily life activities with daily life objects such as in a kitchen or a hospital room, we are in trouble.

The scope of the workshops is gradually shifting from general purpose meetings to more focused ones targeting the definition of practical protocols.

So far, there are no shared methods to compare intelligent robot system capabilities. This is actually a bottleneck, at the same time, for research progress and technology transfer. To a significant extent, the evaluation of the state of the art in a subfield, such as the autonomous navigation of drones, is based more on the subjective judgments of experts than on objective benchmarking methodologies. A collateral effect of this situation is that investors in robotics are forced to assume more risks than necessary, as they do not have objective ways to evaluate the novelty of a prototype with respect to the state of the art, thus slowing the technology transfer in our area of research. This lack of shared performance evaluation procedures is due to a lack of complete scientific understanding of intelligence and cognition as well as weak experimental research practices. There are several involved issues and alternative options. They span, for example, from the selection of the proper level of modeling and abstraction (shown in Figure 1), to the choice of an appropriate statistical description (shown in Figure 2), and to the development of adequate mechanical devices (shown in Figure 3). A prerequisite to the performance comparison of intelligent systems is the ability to replicate research results; however, this is still difficult in many if not most cases.

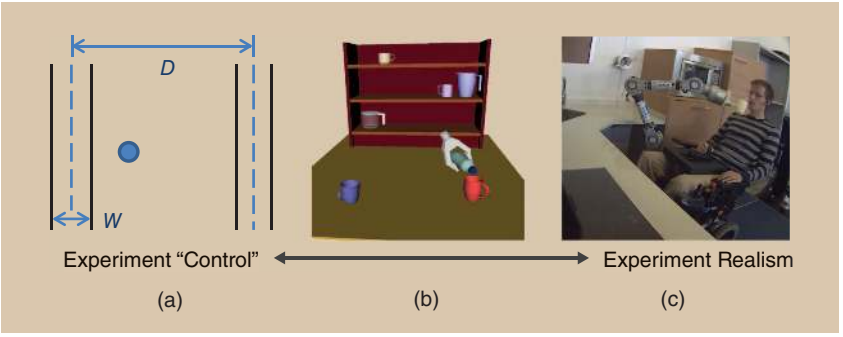

Figure 1. The different levels of modeling in the experimental evaluation of human robot interaction: (a) experiment control, (b) experiment design, and (c) real-world experiment. (Image courtesy of RoboticsLab, UC3M.)

The Technical Committee on Performance Evaluation and Benchmarking of Robotic and Automation Systems (TC-PEBRAS) is intended to serve as a forum to address performance evaluation and benchmarking issues pertaining to robotic and automation systems. This means it is active on both the scientific and methodological challenges, hindering the design and the sharing of mature and usable performance evaluation and benchmarking methods for intelligent and autonomous systems. The Technical Committee (TC) is now in its fourth year of activity since it was approved at the Technical Activities Board (TAB) meeting held at the 2009 International Conference on Robotics and Automation (ICRA) in Kobe, Japan.

The TC-PEBRAS contributes to the progress of performance evaluation and benchmarking, focusing on intelligent robots and systems including those with some high-level cognitive capabilities and

Digital Object Identifier 10.1109/MRA.2014.2298363
Date of publication: 10 March 2014

some degree of autonomy, by providing a forum for researchers and engineers in the industry to exchange their ongoing work and ideas in this article. One purpose of the TC is to achieve better and agreed-upon ideas on how to define and measure system-level characteristics like autonomy, cognition, and intelligence. The development of proper evaluation methods facilitates the technological transfer of research results. The TC-PEBRAS also supports the definition and sharing of research and reporting methods allowing the replication of research results, seen as a necessary precondition to performance comparison.

Figure 2. A subset of the possible end-effectors path for a shared control assistive arm in daily life activities. (Image courtesy of RoboticsLab, UC3M.)

Activities

TC-PEBRAS fosters the discussions, research, and reporting practices that are useful to make significant progress in performance evaluation and benchmarking for robotic and automation systems, in general.

Those objectives are pursued by organizing a number of activities, such as

- workshops to create a sustainable culture of performance evaluation and benchmarking and to provide forums for exchanging ideas and approaches
- major publications (not counting proceedings from the previously mentioned workshops) to provide references for researchers seeking information on performance evaluation and benchmarking
- competitions to put the benchmarks and performance evaluation methods into action.

Workshops

The complete list of activities can be found at http://www.ieee-ras.org/performance-evaluation and http://www.heronrobots.com/EuronGEMSig.

So far, the TC has supported more than 20 workshops on the related activities at various conferences, such as the International Conference on Intelligent Robots and Systems (IROS), ICRA, and Robotics: Science and Systems (RSS). So far more than 200 people have participated in these workshops.

The latest workshops were the following:

- Workshop on Metrics of Embodied Learning Processes in Robots and Animals at IROS 2013 (Tokyo, Japan, 7 November 2013)

Figure 3. The nodal apparatus for a manipulator dexterity evaluation. This model has six directed inspection targets and can be used for retrieval and insertion evaluation as well. (Image courtesy of NIST.)

- Workshop on Proposals for Experimental Protocols for Robotics Research during the RSS 2013 conference (Berlin, Germany, 27 June 2013)
- Workshop on Metrics of Sensory Motor Integration in Robots and Animals at IROS 2012 (Vilamoura, Algarve, Portugal, 12 October 2012)
- Replicable Robotics Research, Benchmarking, and Result Exploitation: Where We Are During the European Robotics Forum (Lyon, France, 21 March 2013)
- From Theory to Practice of Performance Comparison and Result Replications in Robotics Research—Workshop at RSS 2012 (Sydney, Australia, 12 July 2012).

In parallel, the Performance Metrics for Intelligent Systems Workshops focus on performance measure challenges coming from the application of robotics and automation technologies to practical problems in the commercial, industrial, homeland security, and military domains. More information can be found at http://www.nist.gov/el/isd/permis2012.cfm. The scope of the workshops is gradually shifting from general purpose meetings to more focused ones, targeting the definition of practical protocols, like the one at RSS 2013 in Berlin, or deep theoretical issues, like the one at IROS 2013 in Tokyo.

Publications

Several publications, besides the proceedings of the workshops, have been produced on the topics related to the TC, and others are coming. The following are some of the most relevant:

- R. Madhavan, E. Tunstel, and E. Messina, Eds., "Quantifying the performance of intelligent systems," *Int. J. Intell. Control Syst.*, Special Issue, vol. 16, no. 2, pp. 37–159, June 2011.
- "Quantifying the performance of intelligent systems," *Int. J. Intell. Control Syst.*, Special Issue, vol. 16, no. 2, pp. 37–39, June 2011.
- *Performance Evaluation and Benchmarking of Intelligent Systems*, R. Madhavan, E. Tunstel, and E. Messina (Eds.). Springer, ISBN: 978- 1-4419-0491-1, Sept. 2009.
- R. Madhavan, C. Scrapper, and A. Kleiner (Eds.), *Autonomous Robots*, Special Issue, vol. 27, no. 4, 2009.
- R. Madhavan, A. P. del Pobil, and E. Messina, "Performance evaluation and benchmarking of robotic and automation systems [TC Spotlight]," *IEEE Robot. Autom. Mag.*, vol. 17, no. 1, pp. 120–122, 2010.

A book on the measurement of sensory-motor coordination in robots and animals, to be published in the Cognitive Systems Monographs Springer series, is currently in preparation.

Competitions

The TC activities include robotics competitions as competitions are a valid complement to benchmarks for the comparison of system level intelligent behaviors, which can be difficult to quantify in specific quantitative metrics.

Two of the cosponsored competitions of TC-PEBRAS are the Virtual Manufacturing Automation Competition (VMAC) (http://www.vma-competition.com), which was initiated as a demonstration event at ICRA 2008 and has continued as a robot challenge event, and the Mobile Microrobotics Challenge (http://www.nist.gov/eeel/

semiconductor/mmc/) held for the first time at ICRA 2010 in Anchorage, Alaska. The goal of the VMAC is to provide an industrially relevant scenario and performance benchmarks to assess technologies in the areas of robot navigation in dynamic unstructured environments, including mixed palletizing operations and mobile manipulation. The competition's open source policy is designed to encourage collaboration and the dissemination of ideas and algorithms. The objective of the microrobotics challenge is to inspire innovation in microrobot design and to evaluate the performance of the latest microrobotic technologies. Viewed under a microscope, the microrobots are operated by remote control and move in response to changing magnetic fields or electrical signals transmitted across a playing field located on a microchip.

Some of the metrics and test methods are being advanced into draft standards (e.g., the six-degrees of freedom localization from the perception challenge formed the technical foundation for an American Society for Standards and Materials standard being balloted under the E57 Committee on Three-Dimensional Imaging Systems).

The latest supported competitions are as follows:

- ICRA 2013 VMAC, Karlsruhe, Germany
- ICRA 2013 Mobile Microrobotics Challenge (MMC), Karlsruhe, Germany
- ICRA 2012 MMC, Minneapolis, Minnesota, United States
- ICRA 2012 Solutions in Perception Challenge (SPC), Minneapolis, Minnesota, United States
- ICRA 2012 VMAC, Minneapolis, Minnesota, United States
- ICRA 2011 MMC, Shanghai, China
- ICRA 2011 SPC, Shanghai, China
- IICRA 2011 Modular and Reconfigurable Robot Challenge, Shanghai, China
- ICRA 2011 VMAC, Shanghai, China.

In the future, we will look for a tighter cooperation with RoboCup Federation and with other similar initiatives worldwide, such as the European Union-funded RockIN and Eurathlon coordination actions.

Outreach

The TC has coorganized events with Euron and workshops at RSS. Other initiatives have been carried out in collaboration with the RoboCup organization and in conjunction with the U.S. Department of Homeland Security (DHS). The RoboCup Rescue International competitions rely on quantifiable measures of robot performance within draft and existing test methods for evaluating their ability to assist in disaster response. The best-performing robot teams are invited to participate in the NIST-DHS Response Robot Evaluation Exercises, typically held every 18 months at Disaster City, a major responder training facility in the United States.

One important aspect of this TC is the need to cooperate with other TCs. The TC-PEBRAS challenges affect the way research is performed in every area and, for that reason, a mutual exchange of information is needed, as the general methods need to be tailored to the specific necessities of research in specific subfields.

Challenges and Future Work

After several years of activities and some significant successes, there is still a lot of work to be done. As already noticed, a critical precondition for performance comparison of research results is the possibility to replicate them. There is a growing tendency to share data sets and code, in particular in the simultaneous localization and mapping community, and sharing datasets and code is also encouraged and promoted in the workshops we organize. Despite the progress made in defining common protocols and procedures to allow replication, we still lack a venue where we can find experiments that can be replicated following commonly agreed-upon procedures. Achieving this goal will require some time and the continued participation of the community. The definition of shared procedures for benchmarking needs the cooperation of the other TCs. TC-PEBRAS provides a forum to develop a common approach across the diverse areas of research, yet the general methods need to be instantiated with respect to different particular problems addressed by different TCs. In the future, we will have to increase cooperation and joint activities with the other TCs. We will also continue to dig into the nontrivial theoretical issues raised by the measurement of intelligent and cognitive behaviors.

How to Contribute

Much work has been done, but much-work still remains. We are looking for volunteers, in particular volunteers from Asia, Oceania, Latin America, and Africa, which are currently underrepresented. If you wish to join the TC, contribute, or sponsor related activities, you may contact the cochairs by e-mail at fabio.bonsignorio@uc3m.es or fabio.bonsignorio@heronrobots.com, pobil@uji.es, and elena.messina@nist.gov. We look forward to working with you on this challenging and very timely enterprise!

Acknowledgment

We would like to thank Satoshi Tadokoro and the TAB members for their enthusiasm and continued support of this TC. We also thank Raj Madhavan for his efforts as founding chair to constitute this TC and Ken Goldberg who backed its approval.

Fabio Bonsignorio, RoboticsLab, Department of System Engineering and Automation, University Carlos III of Madrid, Spain, and Heron Robots, Genova, Italy. E-mail: fabio.bonsignorio@uc3m.es; fabio.bonsignorio@heronrobots.com.

Elena Messina, Intelligent Systems Division, National Institute of Standards and Technology, Gaithersburg, Maryland. E-mail: elena.messina@nist.gov.

Angel P. del Pobil, Engineering and Computer Science Department, Universitat Jaume I, Spain, and Department of Interaction Science, Sungkyunkwan University, Seoul, South Korea. E-mail: pobil@uji.es.