Benchmarking and Standardization of Intelligent Robotic Systems

Raj Madhavan, Rolf Lakaemper and Tamás Kalmár-Nagy

Abstract— **From mundane and repetitive tasks to assisting first responders in saving lives of victims in disaster scenarios, robots are expected to play an important role in our lives in the coming years. Despite recent advances in mobile robotic systems, lack of widely accepted performance metrics and standards hinder the progress in many application areas such as manufacturing, healthcare, and search and rescue. In this paper, we outline the importance of the development of standardized methods and objective performance evaluation/benchmarking of existing and emerging robotic technologies. We provide a survey of significant past efforts by researchers and developers around the globe and discuss how we can leverage such efforts in advancing the state-of-the-art. Using an example of designing a 'standard' evaluation toolkit for robotic mapping, we illustrate some of the problems faced in developing objective performance metrics whilst accommodating the requirements and restrictions imposed by the intended domain of operation and other practical considerations.**

I. INTRODUCTION

A new frontier of research has been opened up by advances in collaborative operations of man and machine. Mobile robots present almost limitless possibilities by serving as an indispensable aid in dirty, dull, and dangerous environments. Robots will play an increasingly vital role in assisting humans in a variety of domains ranging from innocuous daily chores around the household to potentially harmful situations. Quite aptly, parallels have been drawn between the emergence of the robotics industry and the development of computers 30 years ago [40]. The use of robots in dangerous situations, either in tele-operated or autonomous mode, can not only save lives but also can improve productivity (e.g. factory floors) and in some cases provide solutions which are not possible by humans alone (e.g urban search and rescue). Availability of increased computing power, advances in sensor systems, investments from both the defense and industrial sectors have driven the development of robotic systems with a renewed vigor. In the coming decade, significant progress can be expected in automotive, service, and health care robotics, demonstrating the utility of robotic systems, and, as a result, helping their societal acceptance.

R. Madhavan, R&D Staff, Computational Sciences and Engineering Division, Oak Ridge National Laboratory (ORNL), Oak Ridge, TN 37831 and Guest Researcher, Intelligent Systems Division, National Institute of Standards and Technology (NIST), Gaithersburg, MD 20899, USA raj.madhavan@ieee.org

R. Lakaemper, Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122, USA lakamper@temple.edu

T. Kalmár-Nagy, Department of Aerospace Engineering, Texas A&M University, College Station, TX 77843, USA icar09@kalmarnagy.com

Leaving emerging robotic technologies to proliferate in an unguided direction comes with a high price: synergistic opportunities remain unrealized, lack of cohesion in the community hinders the progress in many domains such as manufacturing, service, search and rescue, and healthcare to name a few. Lack of standards could lead to confusion and frustration for consumers as witnessed in the case of Blu-Ray and HD DVD as the home-video format of choice. A noteworthy example of successful standardization is that of the widely used IEEE 802.11 set of standards for wireless local area network computer communication implemented by the IEEE LAN/MAN Standards Committee [7]. While there are initiatives to provide collections of standard robotics data sets (e.g. The Robotics Data Set Repository [12] and the Robotics Advancement through Web-publishing of Sensorial and Elaborated Extensive Data Sets [13]) and source codes of various robotics algorithms [11], [23], these do not address objective performance evaluation and replication of algorithms is anything but straightforward.

It is our firm belief that in order to facilitate and fast-track the wide acceptance of robotic technologies, scientific methodologies for standardization and benchmarking are crucial. The reasons for these are manyfold:

- Accepted standards and procedures for quantitatively measuring the performance of robotic systems against user-defined requirements will not only improve the utility of mobile robots in already established application areas but will enable the proliferation of such technologies in other emerging markets. Currently, there is no consensus on what objective evaluation procedures need to be followed to deduce 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, comparing robot performance, and leveraging previous work that could otherwise avoid duplication and expedite technology transfer. It is important to develop test artifacts and measurement methodologies to capture performance data in order to focus research efforts, provide direction, and accelerate the advancement of mobile robot capabilities.
- Reuse of information and software interoperability, i.e. common exchange formats for both data and files, is key to providing the research community access to standardized tools, reference data sets, and an open source library of solutions. As a result, researchers and consumers will be able to evaluate the cost and benefits associated with intelligent systems and associated technologies.

To design and develop capable, dependable, and affordable robotic systems, their performance must be measurable. In

this paper, we discuss the advantages of timely standardization of mobile robotics and the risk in not addressing it. We provide an overview of existing standardization efforts of robotic systems around the globe. In addition to the higher level perspective, the design of a standard evaluation toolkit for mapping is discussed as an example to illustrate and provide insight into problems encountered while developing quality measurement and performance evaluation tools. We also include our suggestions on how to leverage current efforts in advancing the state-of-the-art towards benchmarking and standardization.

The rest of the paper is structured as below: Section II discusses related efforts in benchmarking and standardization and the lessons that can be learnt from their outcomes. Section III provides an illustrative example of a typical lower level problem in evaluation design, using the case of robot map evaluation. Section IV describes our suggestions followed by conclusions and continuing work in Section V.

II. Related Efforts and Lessons

Within the European Union (EU), several programs have been administered under the European Commission (EC) Framework Programme (FP) since 2000. Many of the European research projects funded under the current FP7 (EU's chief instrument for funding research over the period 2007 to 2013), especially under the Challenge 2 – Cognitive Systems, Interaction, Robotics, have focused on robot and robotic systems with a stated goal "to have significant industrial and societal impact" [22]. One significant effort is EURON [37], a network of excellence setup by the EC. A EURON benchmarking initiative considered four core robotics subareas: manipulation and grasping, motion planning, networked robotics, and visual servoing, within which attempts to define benchmarks were carried out [6], [38].

Another notable European strategy in robotics is the EUROP (EUropean RObotics Platform) [20] which "brings together all the main European robotics stakeholders with the aim to formulate and implement a consolidated European robotics strategy". This effort focused mainly on three application domains: industrial robot systems, service robots, and space and security robots. *Normalisation* and *standards* are identified as key challenges but it is not clear how much progress has been made and what has been achieved in these areas [21]. Partly funded by the European Commissions Sixth Framework Programme, the Robot Standards and Reference Architectures (RoSta) effort [17] has focused on benchmarking and standards activities for mobile manipulation and service robots [5]. Recently, an IEEE Robotics and Automation (RAS) Standing Committee on Standards Activities has been established to promote common measures and definitions, measurability and comparability, and integratability, portability and reusability of robotics and automation technology [8].

Various standards for industrial robots and robot systems have been propagated through the Robotic Industries Association (RIA) [19]. Typically the emphasis is on safety-related requirements through the R15.06 Robot Safety committee and on robot system and integration. Work on performance standards for urban search and rescue robots is being carried out under the ASTM E54.08 subcommittee on operational equipment within the E54 Homeland Security Application Committee [4], [45]. The Robotics Domain Task Force of the the Object Management Group (OMG) fosters 'integration of robotics systems from modular components through the adoption of OMG standards' [18]. Established in 2005, it facilitates and promotes standardization of OMG technologies by connecting the OMG community with the robotics community by sharing its expertise. The focus is on modularization of robotic systems and standardization of robotic technology components [25], [36]. The Joint Architecture for Unmanned Systems (JAUS), initially developed by an ad hoc working group, is built on the five principles of vehicle platform independence, mission isolation, computer hardware independence, technology independence and operator use independence to accommodate both current and future unmanned systems and is now within the purview of the Society of Automotive Engineers International (SAE) [10]. It has now entered the consensus standards process under SAE AS-4 Unmanned Systems Technical Committee (USTC) [1]. A closely related effort based on the Autonomy Levels for Unmanned Systems (ALFUS) [2] work is also being carried out under the AS-4 USTC.

Such programs, with emphasis on long-term benefits, are critical to sustenance and realization of robotic systems to be a part of everyday life. With the current global economic crisis, market competitiveness has probably never been more apparent. The United States is uniquely positioned to turn the current downturn into an advantage by wisely investing in research and development of robotic systems. In fact, the importance of such investments cannot be overstated if the US were to maintain a competitive edge in a global, volatile, and inter-dependent market. Similar research funding in Asia, particularly in Japan and South Korea, has allowed these countries to make significant strides with a strong correlation between the funding and demonstrable successes [28], [27], [48]. The Computing Community Consortium [3] recently organized a series of workshops to formulate a targeted R&D roadmap for robotics. While stressing the need for the US to invest its resources into robotics-related research, their emphasis is not on performance evaluation and standardization.

Competitions and field exercises are two different yet effective ways of systematically evaluating the performance of robotic systems. The National Institute of Standards and Technology (NIST) has been active in both of these areas via development of performance metrics and reference test arenas. In 2005, the Department of Homeland Security initiated a project to develop performance metrics and standards for robots applied to urban search and rescue. NIST is coordinating and leading this multi-disciplinary effort and is working closely with Federal Emergency Management Agency (FEMA) USAR Task Force members, robot manufacturers, researchers, and other government agencies. A preliminary set of performance requirements has been devel-

oped, primarily based on FEMA responder input [46]. Test methods are being developed that measure the performance of robots against the different requirements. However, great care is taken not to explicitly test for particular technological solutions; rather, the tests measure how effectively or efficiently a robot can complete certain tasks, without assuming a particular approach. This is intended to encourage and foster creative solutions to improve the robots' capabilities.

Starting in 2000, Rescue Robot competitions have been held to increase awareness of the challenges involved in urban search and rescue applications, provide objective evaluation of robotic implementations in representative environments, and promote collaboration between researchers. Initially hosted by the American Association for Artificial Intelligence, these have been adopted by the RoboCup Federation and have grown and expanded to include not just physical robots but also virtual robots and arenas (environments). In the physical competitions, there is an emphasis on mobility challenges, although there are competitors that exhibit autonomous exploration and mapping behavior. The rules for the competitions have evolved each year to encourage robots to negotiate complex and collapsed structures, find simulated victims, determine their condition and location, and generate human readable maps to enable victim recovery. The associated performance metric has also evolved as it attempts to quantify and encourage these and other behaviors [41]. The competition environments (known as arenas) that represent aspects of a collapsed building and contain simulated victims have also been correspondingly modified to increase the difficulty and stress particular robotic capabilities. Since the competition scoring formula penalizes teams that use many operators and the events are timed, autonomy (partial or complete) can be an advantage. The virtual competition provides many of the same mobility, communications, and even sensing challenges that the physical competitions offer, but is able to do so on a much larger scale and is well-suited to focus on development of autonomous behaviors. Hence, robots competing in the virtual realm must operate fully autonomously. Collaboration and cooperation amongst teams of robots is also encouraged [24].

NIST also administers the response robot evaluation exercises at Disaster City, a 52-acre, state of the art training facility which features full-scale, collapsible structures designed to simulate various levels of disaster and wreckage. These response robot evaluation exercises for urban search and rescue teams introduce emerging robotic capabilities to emergency responders within their own training facilities, while educating robot developers regarding the necessary performance requirements and operational constraints to be effective. These events are conducted in training scenarios to help correlate proposed standard test methods with envisioned deployment tasks and to lay the foundation for usage guides identifying a robot's applicability to particular response scenarios [15].

Another effort to stimulate research and development is the dissemination by NIST of sensor datasets for use in map-building or other perception algorithms. The datasets are systematically collected in a maze environment that includes ramps so that off axis rotations and elevations are induced as well as displacements. The data are collected at fixed locations within the maze in four perpendicular directions. Datasets collected using line scan laser range finders have been disseminated to various teams. Additional sets will be collected using stereo color cameras and more advanced sensors such as array range imagers. NIST plans to make these available as downloads on a web site. The results can be quantitatively evaluated against the ground truth for the maze geometry [14].

The Performance Metrics for Intelligent Systems Workshop (PerMIS) [26] is the only one of its kind dedicated to defining measures and methodologies of evaluating performance of intelligent systems. The workshop focuses on applications of performance measures to practical problems in commercial, industrial, homeland security, and military applications and 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.

There have been similar efforts carried out by various researchers, cf. [29], [39], [32]. It is not possible to include all such individual efforts in this paper though a comprehensive survey of related efforts is being compiled. The authors also acknowledge that there may be other significant attempts that are not listed here but believe that the described projects and programs provide a good sampling of the past and existing efforts in the community based on which a picture of the state-of-the-art can be deduced.

The current landscape for robotics standardization in the United States is characterized by promising yet isolated efforts. Although some existing initiatives are attempting to bring together scientists, engineers, and end-users of robot technologies for evaluation and standardization purposes, research and industry have still not yet acknowledged the synergistic opportunities of common underlying principles to an adequate extent. Reaching out to international partners with a high degree of expert-knowledge and experience will enable the US research and industry to catch up with the international standardization efforts, and to establish itself as both a leader and partner in promoting and sharing standards development efforts in the emerging application areas of robotics and automation [34], [33].

III. An Illustrative Example: Evaluating Robot Mapping

In this Section, we give a low level example of problems encountered when evaluating a certain process in autonomous robotics, the process of robot mapping. It will be shown how a more general, task-related view leads to a more versatile evaluation tool.

Robot mapping is the process of creating an internal representation of the robot's physical environment. Maps are created based on sensor data, currently typical sensors are 2D laser range scanners, (stereo) cameras, and a hybrid of them,

range cameras. A 'correct' robot map is a basic requirement for most basic navigational tasks in autonomous mobile robotics, e.g. path planning or (self) localization. Being a pre-process in autonomous robotics, mapping can also be the actual result of a robot's task: in semi autonomous or robot aided settings, human actions can be based on the robotic map. An example for such a setting is the robot aided task of (urban) search and rescue (SAR), where robots explore disaster environments to localize and report victims, general conditions and hazards (e.g. the stability of a building, leaking gas pipes) to first responders. With the goal in mind to autonomously extricate trapped victims, the current state of the art in rescue robotics is to use teleoperated robots to create 2D or 3D maps. Typically, the robot has two different kinds of sensors: first, cameras to transmit the current view (e.g. front and back view) from the robot's current position to the human operator, and second, laser range scanners for the map creation process.

In online robotic mapping, the laser data of the local environment is immediately processed to iteratively create a 'global map' of the environment visited. In this case, the human operator can steer the robot using local camera information, e.g. to avoid obstacles, and information provided by the global map, e.g. for path planning to return the robot to its home location. In offline mapping, all local laser scanner information is first collected and then made available to the mapping process afterwards. Since robust and correct robot mapping is crucial in autonomous and non autonomous rescue robotics, the evaluation of different mapping techniques is of paramount importance.

Mapping, in general, is spatial analysis of environmental features of interest. Inherent to this process is its task dependency, hence there is no 'optimal general mapping'. Mapping of spatial features can be divided into two classes, topographic and topological mapping. While topographic mapping is concerned with detailed, correct geometry of the spatial features, topological mapping aims for correct spatial relation between features only; it favors topological correctness over geometric accuracy, also often referred to as 'global correctness' vs. 'local accuracy'. In robot mapping, two approaches relate to topographic and topological mapping. They are *grid based* and *pose based* approaches (pose based approaches are still geometric, and not real topological mapping. However, looking at low geometric feature level approaches only, pose based evaluation can be utilized for topological mapping). The following will illustrate both approaches and propose a new, hybrid approach.

A. Grid Based Approaches

As mentioned earlier, the yearly RoboCup Rescue competition [16] is a forum to compare the performance of rescue robots. It creates standard environments, tools and algorithms to develop comprehensive systems for rescue scenarios. The robots' performance is evaluated with respect to different aspects with robot map quality being one of them. The tool for map evaluation used in the 2008 RoboCup Rescue competition was the *Jacobs Map Analysis Toolkit* [50], [9], an open source visual toolkit to assist a human referee in the scoring process. It consists mainly of three steps. First, the arena building module is utilized to draw a ground truth map, specifically supporting standard RoboCup Rescue arena type environments. The second is a visualization tool, which allows for transparent superimposition of robot generated maps with the ground truth for subjective, visual inspection and scoring. Third, the Jacobs Toolkit's offers a map similarity measure, which is an attempt to create a mapping evaluation standard. The Jacobs Toolkit's similarity measure is a *grid based* approach: the map to be evaluated (target map) and the ground truth map are both embedded into a grid. The grid cells are labeled using properties like 'object', 'empty space' or 'hidden'.

In RoboCup Rescue 2008, the target maps were submitted to the system either as GeoTIFF formatted image-files, or as a printout; one advantage of a grid based approach is that it can deal with such basic map representations. The similarity is computed as sum of nearest distances between target and ground truth cells of same label. The underlying similarity function is an efficiently implemented version of the distance transform. The Jacobs Toolkit measures the local geometric accuracy of the map. Since only low level features (object/empty space) are incorporated, the target map must be close to the ground truth map: it is assumed that low level correspondences imply higher level correspondences (e.g. object-object $\Rightarrow$ 'door'-'door'). Larger errors in the global appearance of maps can not be quantified, globally erroneous maps are classified as 'wrong' - even if they are locally correct, see the example in Figure 1. The example illustrates a case which is typical for non-autonomous mapping, where first responders need a map to get an overview of the environment to rescue a victim. Global geometric correctness might be of minor interest compared to locally geometric, yet global topological accuracy. Figure 1, left, shows the ground truth map. Figure 1, center, illustrates a mapping result with high global geometric correctness, although the bottom part is wrong in details. The right example is for a map with a high global geometric error. However, all details (obstacles, victim's position in bottom room) are mapped correctly, the map is also topologically correct (two rooms are connected by hallways). A grid based approach will prefer the center map to the right one. However, if the map is intended to be a navigational aid for first responders, the right map is of higher quality: it shows correctly that the victim (red dot) is reachable from the current position (black dot) using the right hallway. The center map misleads the first responders to take the left hallway, a probably fatal mistake.

Grid based approaches like the Jacobs Toolkit aim to measure the global topographic quality of a robot map, they can not quantify the topological qualities of a map.

B. Pose Based Approaches

A different approach to mapping evaluation is *pose based map quality estimation*. Pose Based fitness exploits the fact that precise robot localization is dual to robot mapping: if the robot pose is precisely known in the ground truth map, the

Fig. 1. Grid based evaluation. Left: Ground truth map. Two rooms (green, top and bottom) with obstacles (gray) are connected by two hallways (green, center). The victim (red dot) can only be reached from the current position (black dot) using the right hallway. Center: mapping example with high global and low local correctness. Right: mapping example with low global and high local correctness. A grid based map evaluation will prefer the center map, although for first responders the right map is of better use.

scans can be registered based on the pose estimates. Since robot pose measurements are imprecise, the scan data itself has to be taken into account to register the scans in a common coordinate system. Successful registration of scans adjusts the robot poses defined by the target map into the ground truth coordinate system.

Evaluation based on pose information compares the adjustment of robot poses, i.e. an error e is computed for every pose as

$$e(x, z, \theta) = |(x, z) - (x_G, x_G)| + \gamma|\theta - \theta_G| \qquad (1)$$

with x, z, θ defining the robot's estimated pose and x_G, z_G, θ_G defining the ground truth pose. γ is a weight factor, modeling the perceptually different influences of rotational and translational errors. The sum of all pose errors yields the overall error. The main advantage of pose based evaluation is its applicability in higher dimensions (e.g. 6D-SLAM). The number of poses to be evaluated is dimensionality independent, whereas the memory consumption of a grid based evaluation approach increases for 3D applications to a prohibitive cubic behavior. Hence there is a high interest in gaining knowledge about pose based evaluation.

Fig. 2. Pose based evaluation. The target map (right) is transformed to match the ground truth (left). The transformation parameters (here: rotation, arrows) are used to quantify the map quality. The topological correctness of the target map is reflected by the fact that only two rotations are needed to achieve the optimal map.

Although still a geometric measure, pose based map estimation is closely related to evaluation of topographical maps, see the example in Figure 2. The global topological correctness is captured by the fact that only a few rotations are needed to achieve the optimal result.

Due to different local and global influence of rotational errors, it is hard in pose based estimation to precisely quantify local geometric correctness, a major drawback compared to grid based approaches. Also, pose based evaluation requires the less intuitive single scans along with their aligned poses as input. In practice this is not a drawback, since mapping approaches are naturally based on single scans. Once a pose based evaluation standard is established, mapping algorithms can be required to save single poses before merging the scans to a global map. Another critical factor is the parameter γ in Eq. 1. Angular errors translate radius dependent to absolute errors, which is hard to model in a single parameter. Additionally, simple summing of errors, in general, does not reflect the influence of pose errors in real world settings.

C. A Hybrid Evaluation Approach

For combined evaluation with respect to topographical and topological map properties, a new hybrid pose/grid-based evaluation has been proposed. Emerging from a mapping approach, Virtual Scan assisted Force Field Simulation (VFFS) [42], is designed to eliminate the drawbacks of pure pose or grid based evaluation and to combine their advantages. VFFS is an offline scan alignment technique, which rigidly transforms (rotation/translation) single scans to achieve an optimal map. Additionally it takes into account expected structures (e.g. straight walls), which augment the original sensor data with hypotheses about objects in the environment. These data, called virtual scans, are added to the physical sensor data with a certain weight of confidence; a high weight forces the algorithm to align the real scans to the virtual scans.

VFFS mapping evaluation is based on 4 design principles:

1) Substitute the alignment of real scans to virtual scans with an alignment between target map and ground truth map.
2) Split the ground truth map into parts with required high geometric accuracy. These parts are the 'single scans', they will be aligned by VFFS to the target map (=the virtual scan).
3) Instead of aligning the target map to the ground truth map, use reverse alignment: align the ground truth map to the target map
4) After alignment, use the part-alignment parameters for a pose based evaluation. Additionally, evaluate the geometric preciseness of each part (after VFFS alignment) using a grid based approach. The weight of each part-transformation as well as the relative weight of pose-based and grid based evaluation scores are predetermined in relation to the task the map was created for.

(1), (2) and (3) define the mapping evaluation in the framework of the VFFS mapping approach. Using the target map as a fixed virtual scan with high confidence weight, align

the single scans of the decomposed ground truth map to the target map, see Figure 3. Observe that in this approach we transform the ground truth map, not the target map. There are two reasons for such an approach: first, it makes the evaluation independent of the target map's data format. Since the target map is not transformed, it can be given in any format, e.g. GeoTiff. Second, and more important, the part-decomposition of the ground truth map can reflect the task specific requirements of the mapping approach. For example, the ground truth map of Figure 3(a) is decomposed into top room, hallways and bottom room. These three parts are required to be mapped with high geometric accuracy. Topic (4) quantifies the map quality, using pose based parameters from VFFS, and grid based parameters from additional evaluations on the single parts. The relative pose defines the global appearance of the map. It will be captured by the transformation parameters. The importance weight of the transformation parameters can be individually determined. In the example of mapping-assistance for first responders, a rotational error among a certain threshold could be omitted, leading to a preference of the right map in Figure 1.

Fig. 3. VFFS based hybrid evaluation. (a) decomposed ground truth map, 3 parts. (b) target map. (c) VFFS transforms (a) to (b). (d) Grid based evaluation on transformed parts. The final score is computed using task adjusted weights for transformation parameters and grid evaluation results.

A tool utilizing the pose and grid based hybrid evaluation approach can adjust to both global appearance and local precision. It is therefore a versatile instrument to evaluate robot maps accounting for different requirements of different tasks.

IV. WHERE DO WE GO FROM HERE?

It is not an exaggeration to claim that interest in benchmarking and standardization of robotic research has reached 'critical mass' as witnessed by the increasing number of workshops and journal special issues [35], [30], [31], [43], [44] dedicated to this topic. Researchers, developers, and end-users alike are aware of the problems [49], the importance of such efforts, and how it can be beneficial to them. One common complaint about standardization is that too early standardization hinders the use of more recent technologies/techniques that would admit desirable performance (and price etc.). For specific technologies for devices and communication it is easier to know when the technology is mature and should be standardized, even if later upgrades are likely (as with Ethernet and related network technologies), but for systems (for navigation, motion control, etc.) that is much harder. The other problem with standards, even if they are done at the right time, is that the complex (sub)system interactions lead to standards that are too 'thick', and therefore are hard to apply; consequently they tend to be obsolete before being passed as a formal standard. While standards are needed for interoperability, experience from our collaborators and others has shown that loosely coupled systems permit reuse without normative standards and thus should precede standardization [47].

We believe that a concerted international initiative to research and develop objective standards and procedures leading towards tangible and meaningful performance metrics is imperative if we are to increase the robots' level of participation in our daily lives (e.g. robotic assistants for the elderly), and the acceptance and proliferation of new technologies to improve the quality of human lives. Once initial efforts gain traction and are eventually accepted within the research and industrial community, it can lead towards the establishment of de facto standards, which can then be propagated through existing standards organizations, resulting in a widely accepted 'standard'. In addition, we believe that substantiating evidence resulting from concomitant research can facilitate the acceptance of emerging standards in different countries which in turn will expedite the worldwide acceptance of normative standards.

V. CONCLUSIONS AND CONTINUING WORK

Though we are beginning to see many instances of man-machine collaborative applications, there is an inherent *mental block* in terms of the social acceptance of robots as a trustworthy aid to supplant a human even in harmful situations. Only through extensive experimentation and continual refinement can this attitude be overcome for which we believe it is crucial to demonstrate the mettle of robotic systems that are capable and dependable. This paper outlined some of our ideas on benchmarking and standardization, and why we think these areas are critical in achieving these goals by drawing upon our experiences in working with fielded robotic systems and in conjunction with end-users, developers, researchers, and vendors.

The following are current areas of research that we are undertaking to facilitate performance evaluation and standardization:

• It is a daunting task to come up with a mathematical formulation and a framework that is generic enough to quantitatively evaluate performance of intelligent systems. We take an approach that such a framework is better developed by learning from requirements imposed by end-users from multiple domains rather than trying to develop a theoretical

framework that is not grounded in practicality.

• Open source software and technologies enable quick implementations and testing of new algorithms with minimal effort. Our experience has shown us that it is an excellent means to maintain transparency and increase reliability.

• We are bringing together the research community to work collaboratively in developing shared-solutions across different application areas through field exercises, competitions, and scholarly exchange of ideas via workshops, publications, and discussions.

REFERENCES

[1] AS-4 Unmanned Systems Committee. http://www.sae.org/servlets/works/committeeHome.do?comtID=TEAAS4.

[2] Autonomy Levels for Unmanned Systems. http://www.isd.mel.nist.gov/projects/autonomy_levels/.

[3] CCC/CRA Roadmapping for Robotics. http://www.us-robotics.us/.

[4] Committee E54 on Homeland Security Applications. http://www.astm.org/COMMIT/COMMITTEE/E54.htm/.

[5] Defining a Standard Benchmark for Mobile Service Robots. http://wiki.robot-standards.org/index.php/Benchmarks.

[6] EURON Benchmarking Initiative. http://www.euron.org/activities/benchmarks/;http://www.robot.uji.es/EURON/en/index.htm/.

[7] IEEE 802 LAN/MAN Standards Committee. http://www.ieee802.org/.

[8] IEEE RAS Standing Committee on Standards Activities. http://www.ieee-ras.org/industrial/standards/.

[9] Jacobs Map Analysis Toolkit. http://robotics.iu-bremen.de/datasets/MapEvaluation/.

[10] Joint Architecture for Unmanned Systems. http://www.jauswg.org/.

[11] OpenSLAM. http://www.openslam.org/.

[12] Radish: The Robotics Data Set Repository. http://radish.sourceforge.net/.

[13] RAWSEEDS: Robotics Advancement through Web-publishing of Sensorial and Elaborated Extensive Data Sets. http://www.rawseeds.org/.

[14] Reference Datasets. http://www.isd.mel.nist.gov/projects/USAR/2007/refdatasets.htm/.

[15] Response Robot Evaluation Exercise (#5). http://www.isd.mel.nist.gov/US&R_Robot_Standards/disaster_city/eventint%ro5.htm/.

[16] Robocuprescue. http://www.robocuprescue.org.

[17] Robot Standards and Reference Architectures. http://www.robot-standards.eu/.

[18] Robotics Domain Task Force. http://robotics.omg.org/.

[19] Robotics Industries Association. http://www.robotics.org/.

[20] The European Robotics Platoform. http://cordis.europa.eu/ist/europ/.

[21] The European Robotics Platoform: Strategic Research Agenda. ftp://ftp.cordis.europa.eu/pub/ist/docs/europ/rob-plat-2.pdf/.

[22] The FP7 ICT Work Programme. http://cordis.europa.eu/fp7/ict/programme/.

[23] The Mobile Robot Programming Toolkit. http://babel.isa.uma.es/mrpt/.

[24] Urban Search and RescueRobot Competitions. http://www.isd.mel.nist.gov/projects/USAR/competitions.htm/.

[25] Workshop on 'Robotic Standardization', IEEE/RSJ International Conference on Intelligent Robots and Systems. http://staff.aist.go.jp/t.kotoku/conf/iros2006ws/.

[26] Performance Metrics for Intelligent Systems (PerMIS) Workshop, 2000–2008. http://www.isd.mel.nist.gov/PerMIS/.

[27] Robotics in Manufacturing Technology Roadmap. Technical report, Energetics Incorporated, November 2006.

[28] *Rising Above The Gathering Storm: Energizing and Employing America for a Brighter Economic Future*. National Academies Press, 2007.

[29] S. Abdallah, D. Asmar, and J. Zelek. Towards Benchmarks for Vision SLAM Algorithms. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 3834–3839, May 2006.

[30] R. Madhavan A.P. del Pobil and E. Messina (Organizers). In *Performance Evaluation and Benchmarking for Intelligent Robots and Systems*, November 2007.

[31] R. Madhavan A.P. del Pobil and F. Bonsignorio (Organizers). In *Performance Evaluation and Benchmarking for Intelligent Robots and Systems*, September 2008.

[32] J. Baltes. A Benchmark Suite for Mobile Robots. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1101–1106, 2000.

[33] G. Bekey, R. Ambrose, V. Kumar, D. Lavery, A. Sanderson, B. Wilcox, J. Yuh, and Y. Zheng. *Robotics: State of the Art and Future Challenges*. Imperial College Press, 2008.

[34] G. Bekey, R. Ambrose, V. Kumar, A. Sanderson, and B. Wilcox. WTEC Panel Report on International Assessment of Research and Development in Robotics. Technical report, National Technical Information Service (NTIS), January 2006. Report No. PB2007-102141.

[35] F. Bonsignorio, J. Hallam, and A.P. del Pobil (Organizers). Experimental Methodology and Benchmarking in Robotics Research. In *Workshop Proceedings, Robotics: Science and Systems (RSS) Conference, Zurich, Switzerland*, June 2008. http://www.heronrobots.com/EuronGEMSig/GEMSIGRSS08Program.html/.

[36] A. Bose. The OMG Robotics DTF: Motives, Challenges & Efforts. In *Workshop on 'Measures and Procedures for the Evaluation of Robot Architectures and Middleware', IEEE/RSJ International Conference on Intelligent Robots and Systems*. http://wiki.robot-standards.org/images/1/1a/Omg_robotics.pdf/.

[37] H. Christensen. EURON - The European Robotics Network. *IEEE Robotics & Automation Magazine*, 12(2):10–13, June 2005.

[38] R. Dillmann. KA 1.10 Benchmarks for Robotics Research, April 2004. http://www.cas.kth.se/euron/eurondeliverables/ka1-10-benchmarking.pdf/.

[39] D. Feil-Seifer, K. Skinner, and M. Matarić. Benchmarks for Evaluating Socially Assistive Robotics. *Interaction Studies: Psychological Benchmarks of Human-Robot Inteaction*, 8(3), October 2007.

[40] W. Gates. A Robot in Every Home. In *Scientific American*, December 2006.

[41] A. Jacoff, B. Weiss, and E. Messina. Evolution of Metrics and Performance for USAR Competitions. In *Proceedings of the Performance Metrics for Intelligent Systems (PerMIS) Workshop, NIST Special Publication 1017*, September 2003.

[42] R. Lakaemper and N. Adluru. Improving Sparse Laser Scan Alignment with Virtual Scans. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2915–2921, 2008.

[43] R. Madhavan, C. Scrapper, and A. Kleiner (eds.). Characterizing Mobile Robot Localization and Mapping. *Autonomous Robots*, 2009 (to appear).

[44] R. Madhavan, E. Tunstel, and E. Messina, editors. *Performance Evaluation and Benchmarking of Intelligent Systems*. Springer, September 2009 (to appear).

[45] E. Messina. Performance Standards for Urban Search and Rescue Robots. *ASTM Standardization News*, August 2006.

[46] E. Messina, A. Jacoff, J. Scholtz, C. Schlenoff, H.-M. Huang, A. Lytle, and J. Blitch. Statement Of Requirements For Urban Search And Rescue Robot Performance Standards, 2005. http://www.isd.mel.nist.gov/US&R_Robot_Standards/Requirements%20Report%%20(prelim).pdf.

[47] K. Nilsson. Personal Communication.

[48] The President's Council of Advisors on Science and Technology (PCAST). Leadership Under Challenge: Information Technology R&D in a Competitive World. Technical report, August 2007. http://www.ostp.gov/pdf/nitrd_review.pdf.

[49] E. Prassler and K. Nilsson. 1,001 Robot Architectures for 1,001 Robots. *IEEE Robotics & Automation Magazine*, 16(1):113, March 2009.

[50] I. Varsadan, A. Birk, and M. Pfingsthorn. Determining Map Quality through an Image Similarity Metric. In *Lecture Notes in Artificial Intelligence (LNAI): RoboCup 2008: Robot WorldCup XII*, 2008.