2010 NDIA GROUND VEHICLE SYSTEMS ENGINEERING AND TECHNOLOGY SYMPOSIUM
ROBOTICS MINI-SYMPOSIUM
AUGUST 17-19 DEARBORN, MICHIGAN

PERFORMANCE EVALUATION AND BENCHMARKING FOR UNMANNED GROUND VEHICLES

Phillip J Durst
US Army Engineer Research and Development Center
Vicksburg, MS

ABSTRACT

Unmanned ground vehicles (UGVs) are being fielded with increasing frequency for military applications. However, there is a lack of agreed upon standards, definitions, performance metrics, and evaluation procedures for UGVs. UGV design, development, and deployability have suffered from the lack of accepted standards and metrics. Developing these standards is exceptionally difficult, because any performance metric must not only be evaluated through controlled experiments, but the metric itself must also be checked for relevance. Several committees and workgroups have taken up the challenge of providing standardized performance metrics, and an overview of the current state of performance evaluation for UGVs is presented. The ability to evaluate a potential metric through simulations would greatly enable these work efforts. To that end, an overview of the Virtual Autonomous Navigation Environment (VANE) computational test bed (CTB) and its potential use in the rapid development of meaningful UGV performance metrics is presented.

INTRODUCTION

What does an unmanned ground vehicle (UGV) need to do well in order to complete its mission, and how well does it need to do these things? In other fields of development, these questions would be answered using a set of standard tests and performance evaluations. These standards would allow comparison across different platforms and hardware configurations. Moreover, the system's performance as evaluated by the standard procedures would lend confidence to the system, insuring its performance was both predictable and repeatable. Unfortunately, no such standards exist for ground robotics.

Currently, UGVs are evaluated on a 'case-by-case' basis. Testing involves making some educated guesses about what the platform needs to do in order to complete its mission. Guesses are made concerning how to measure the UGV's success at these tasks, which is often measured simply in terms of whether or not the UGV completed its mission. Testing conducted in this manner is both expensive and inconsistent. Additionally, UGVs performing well in assumed mission environments do not necessarily perform well in the field.

A lack of standard test methods for performance evaluation has severely hampered UGV development. This lack is seen the most in the area of UGV autonomy. The autonomous capabilities of ground robotics is extremely under-represented in the field, because in the absence of agreed upon performance evaluations, no mechanism exists to transition autonomous capabilities to fielded platforms.

Because of the wide variety of missions for which UGVs are used, a standard set of performance evaluation procedures is difficult to construct. UGV performance must be evaluated within the larger context of the mission, and performance evaluations must take into account the environment in which the UGV operates. Moreover, any proposed performance evaluation method would itself have to be evaluated. Test procedures designed to evaluate UGV performance must be able to predict a UGV's mission performance capability. The difficulty in creating standard performance evaluation procedures for UGVs, and the reason so few exist, is that discovering these procedures requires numerous iterations of a complete UGV mission under identical conditions.

Several workgroups within the robotics community have recognized the growing need for standardization. The following paper gives details on the current efforts across the robotics community to develop standard performance evaluation procedures. In light of the efforts of these groups, a conceptual model for how UGV test procedures could be developed is presented. Due to the demand for repeatable, controllable settings for the development of metrics for test performance measurement, a high-fidelity simulation environment for testing, evaluation, and analysis would be a major enabler. To that end, the VANE CTB is presented, and its capacity for rapid test procedure development is discussed.

CURRENT STANDARDIZATION EFFORTS

Given the interdisciplinary nature of robotics, the workgroups focusing on performance evaluation encompass a variety of research groups. The Institute of Electrical and Electronics Engineers has formed the Technical Committee on Performance Evaluation and Benchmarking of Robotic and Automation Systems (TC-PEBRAS) [1]. The focus of the TC-PEBRAS is the quantitative measurement of robotic system performance. The goal of the TC-PEBRAS is to observe the performance evaluations being conducted by researchers and foster the development of standard tests for evaluation. Their efforts are directed at not only robot platforms but also the algorithms deployed on those platforms. Some of the workgroups's findings can be found in [2] and [3].

A North Atlantic Treaty Organization (NATO) Applied Vehicle Technology (AVT) workgroup, referred to as AVT-175 Unmanned Systems (UMS) Platform Technologies and Performances for Autonomous Operations, has undertaken a similar effort. The AVT-175 workgroup, formed in 2007, has endeavored to develop a unified framework for the design of unmanned systems [4]. While their task is not centered specifically on developing performance metrics, a large part of their efforts are focused on finding a means of evaluating UMS autonomous mission performance. A driving force behind the AVT-175 is the need for a quantitative metric for a UMS's level of autonomy.

A workgroup similar to the TC-PEBRAS is the Society of Automotive Engineers (SAE) AS-4D Unmanned Systems Performance Measures subcommittee [5]. This subcommittee has produced the Performance Measures for Unmanned Systems (PerMFUS) framework document [6]. The goal of the PerMFUS is to answer the questions of what to measure while testing UMS and how to measure it. The final document will provide a framework for how to evaluate performance as opposed to the benchmark tests themselves. The results from the AS-4D subcommittee are designed to feed directly into the Autonomy Levels for Unmanned Systems (ALFUS) framework.

The ALFUS framework was created by the AFLUS workgroup and forms the only current working model for evaluating UMS's autonomous capabilities [7]. The ALFUS framework lays the groundwork for how an unmanned system's performance evaluations could be combined into a single quantitative measure of a system's autonomy (Figure 1). Within the ALFUS framework, a system's level of autonomy is determined by using a tool called the Contextual Autonomous Capability (CAC). The CAC (Figure 2) is a 3-axis measurement that takes into account the UMS mission, operational environment, and operator workload. Each axis would be comprised of scores from bench tests related to each axis, and the axes would be combined to form one overall level of autonomy score. The obvious shortcoming of the CAC is that no standard bench tests exit to fill in the axes.

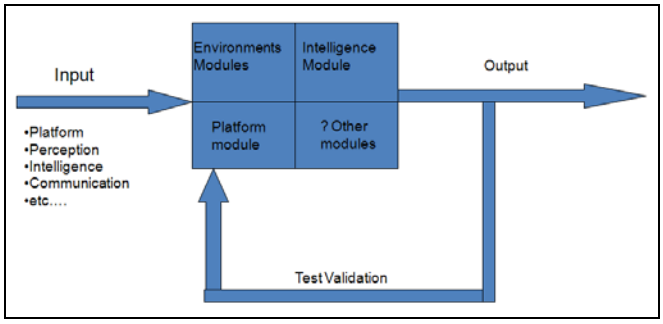

Figure 1. The ALFUS concept for an autonomy performance assessment tool. Information about the robotic hardware, software, and mission are tested using standard procedures, which are combined to output the system's autonomous capabilities.

QUANTITATIVELY ASSESSING UGV PERFORMANCE

While work is just beginning to establish standard performance measurements for UGVs, the idea of quantifying mobile robot performance is by no means a new one [8]. Many researchers have employed statistical assessment methods for their robots [9], and some work has even been done in finding performance measurement methods for UGV algorithms [10]. However, established metrics found in the literature suffer from the problem of being specific to a certain platform and mission. Quantitative performance measures are developed based on the platform or the algorithm to be tested, but no effort is made to standardize these metric or extrapolate them into a larger context.

The greatest problem facing any effort for developing performance metrics for UGVs is the complexity of the interactions between the robot and its environment. Because autonomous robots closely resemble biological agents, any statement made about a UGV's capabilities must be made within the context of its operational environment [9]. Therefore, to properly evaluate UGV performance, metrics related to the environment in which the UGV operates must also be developed. These metrics are accounted for in both the ALFUS standards and the model presented below. But as of writing, no formal work effort has been undertaken to develop standard metrics for UGV operational environment concerns.

MODELS FOR CREATING PERFORMANCE EVALUATION PROCEDURES

One conclusion that is found across all the workgroups and literature is that, unlike with other systems, metrics for evaluating UMS performance cannot be separated from the mission the UMS is to perform. The sensors, software, and hardware needed to perform one mission are often completely worthless for another mission. Therefore, any standard test for UMS would have to be unique to a given mission and environment. But, if standards must be developed for every robot and every mission, then the entire purpose of developing standards is defeated.

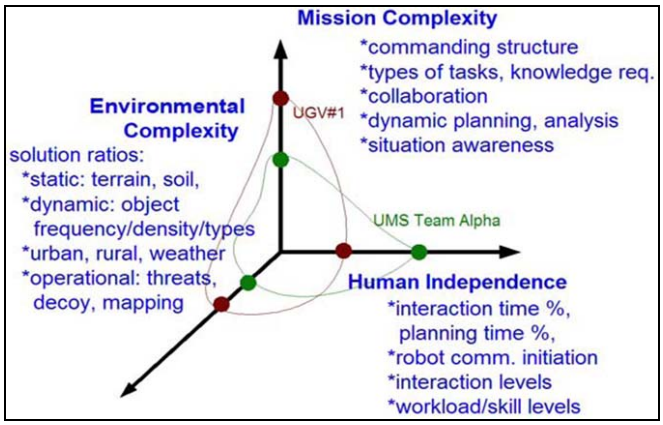

Figure 2. The Contextual Autonomous Capability (CAC). Scores for testing related to the three axes are combined into a single quantitative measurement of the system's level of autonomy.

Using the findings of the presented workgroup efforts and literature, a general model for developing standard test methods for UGVs can be devised (Figure 3). The model is a three step method. First, for a given mission setting, a set of simple bench tests are proposed and measured for the UGV. Second, the UGV is tested during the mission for success/failure. This process must be repeated for several different UGVs. Last, the UGV performances are compared. If the UGVs with a higher score for a given test performed better during the mission, then that test is transitioned into a standard benchmarking procedure.

An excellent example of this type of methodology at work can be found in [11]. The specific mission and environment of urban disaster first response are chosen, and a general class of SUGV is studied. The robots are evaluated through a set of standard tests, with robots achieving higher scores being found to perform better in the field. Standard tests are used to evaluate the SUGVs, with higher scores indicating better mission performance.

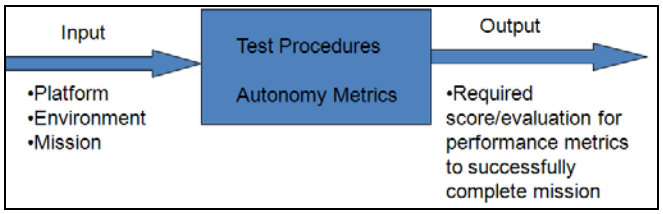

Figure 3. A model for creating UGV assessment procedures. Tests should provide a meaningful measure of the UGV's mission capabilities.

EVALUATING BENCHMARKS THROUGH SIMULATIONS – THE VANE CTB

The model presented for bench test development has many drawbacks. Namely, it is time consuming and expensive, requiring multiple interactions of entire UGV missions, iterated multiple times for a wide range of UGV platforms. Furthermore, these missions would need to be run identically to control for weather effects, etc. Therefore, this method of development for standard tests and metrics works best in a simulation environment. Additionally, testing in a simulation environment would enable evaluations of not only UGV platforms but also mission specific algorithms.

In fact, this paper is not the first to propose using simulations for quantitatively measuring UGV mission performance. To date, the best UGV simulation test bed available is NIST's USARSim, a high fidelity robot simulator built off a commercial gaming engine [12]. USARSim has been validated against real world results, both for robot physics and mobile robot algorithms [13] [14], and it has been used to quantitatively evaluate mobile robot simultaneous mapping and localization (SLAM) algorithms [10]. The drawbacks of USARSim are that it is limited to indoor environments, is focused on disaster first response missions, and does not employ physics based sensor models. For use in developing UGV bench tests, a simulation environment would have to include large, outdoor missions and physics-based sensor-environment interactions.

For a simulation environment to effectively evaluate autonomous UGV mission capabilities, it must reproduce sensor response outputs and geoenvironmental effects at a near truth level. As part of the Safe Operations of Unmanned systems for Reconnaissance in Complex Environments (SOURCE) Army Technology Objective (ATO) research and development program, the US Army Engineer Research and Development Center (ERDC) and its partners have undertaken the development of this type of simulation environment [15]. Dubbed the Virtual Autonomous Navigation Environment (VANE) Computational Test bed (CTB), it encompasses a set of high-resolution environment, terrain, sensor, and vehicle models. The UGV, sensor response outputs, and vehicle-

terrain interactions are simulated using first principles with the ultimate goal of evaluating autonomous navigation capabilities for UGVs.

Sensor response outputs in the VANE CTB are more able to recreate the affects of the environment on UGV autonomy than their empirically derived counterparts [16]. Additionally, UGV robot platform performance is better predicted using ERDC's legacy high fidelity, physics driven vehicle modeling capabilities. Furthermore, a ground contact element (GCE) along with rigorous soil modeling allows the UGV to interact with and influence the terrain it traverses [17]. Scenes for VANE CTB simulations are on the kilometer scale and encompass the entire range of environments in which UGVs operate (Figure 4).

Figure 4. An example scene used for VANE CTB simulations. The scene is roughly one square kilometer and contains several fully realized urban blocks.

The VANE CTB effectively solves the issues with the presented model for creating performance evaluation procedures. Multiple UGVs can be tested repeatedly under identical circumstances. Entire UGV missions can be simulated, often under conditions that cannot be effectively reproduced in a laboratory (enemy forces, pedestrian traffic, etc.). Unlike most field tests, the VANE CTB could evaluate UGV performance within the context of the larger UGV mission under true-to-life mission conditions. Additionally, the same mission could be simulated multiple times under differing conditions to determine the impact on system reliability.

Using the VANE CTB, UGV performance evaluation procedures could be rapidly developed. Individual, even mission specific, necessary UGV capabilities could be studied in depth. Those capabilities could be tested for relevance across varied missions and settings. If the capability proved useful for mission success, they could be used to derive simple, quantitative bench tests.

Different UGV platforms could be simulated performing the same mission. If the UGVs with higher bench test scores performed better, then the proposed bench test would be considered meaningful. Variations on the mission could then be iterated until the set of bench tests was optimized. Finally, a battery of standard evaluation procedures could be selected for a given class of UGV platform and mission.

DISCUSSION

The field of robotics is growing exponentially, but this growth has not yet been reflected in fielded applications. Due to a lack of agreed upon, standardized methods for evaluating and quantifying UGV performance, there exists a lack of knowledge and confidence in the capabilities of UGVs. This holds especially true for autonomous mobile robots, for no accepted measure of UGV autonomy exists as of yet. For UGVs to reach their full potential for military applications, a standard set of performance evaluation tests must be developed.

A model for quickly determining a set of meaningful performance metrics was presented. The model involved not only evaluating UGV metrics but also evaluating the metrics themselves within the broader mission context. Efficient use of this model necessitates a high fidelity simulation environment, which is provided by the VANE CTB. Using the VANE CTB, proposed performance metrics could be rapidly tested and transitioned into standardized test procedures.

ACKNOWLEDGEMENTS

Permission to publish was granted by the Director, Geotechnical and Structures Laboratory.

REFERENCES

[1] IEEE TC: Performance Evaluation and Benchmarking of Robotic and Automation Systems, http://tab.ieee-ras.org/committeeinfo.php?tcid=35.

[2] Autonomous Robots Journal, special issue-Performance Evaluation and Benchmarking, November 2009.

[3] Raj Madhaven, Edward Tunsel, and Elena Messina, "Performance Evaluation and Benchmarking of Intelligent Systems", Springer 2009.

[4] AVT-175 Unmanned Systems (UMS) Platform Technologies and Performances for Autonomous Operations, http://www.rta.nato.int/Activity_Meta.asp?Act=AVT-175/

[5] AS-4D Unmanned Performance Measures Committee, http://www.sae.org/servlets/works/postDiscussion.do?commtID=TEAAS4D&docID=&forumID=17351&resourceID=134092&inputPage=showAll.

[6] H. Huang, Elena Messina, and Adam Jacoff, "Performance Measures for Unmanned Systems", Proceedings of the Performance Metrics for Intelligent Systems (PerMIS) 2009 Conference.

[7] H. Huang, Kerry Pavek, James Albus, and Elena Messina, "Autonomy Levels for Unmanned Systems (ALFUS) Framework: An Update", 2005 SPIE Defense and Security Symposium.

[8] Eran Gat, "Towards Principled Experimental Study of Autonomous Mobile Robots", Autonomous Robots V 2 pp. 179-189, 1995.

[9] Ulrich Nehmzow, "Quantitative analysis of robot-environment interaction-towards 'scientific mobile robots'", Robotics and Autonomous Systems V 44 pp. 55-68, 2003.

[10] B. Balaguer, S. Carpin, S. Balakirsky, "Towards Quantitative Comparisons of Robot Algorithms: Experiences with SLAM in Simulation and Real World Systems", IROS 2007 Workshop, 2007.

[11] Adam Jacoff, "Urban Search and Rescue Robot Performance Standards: Progress Update", June 2007.

[12] S. Carpin, M. Lewis, J. Wang, S. Balakirsky, C. Scrapper, "USARSim: a robot simulator for research and education", 2007 IEEE International Conference on Robotics and Automation, pp. 1400-1405, 2007.

[13] S. Carpin, T. Stoyanov, Y. Nevatia, M. Lewis, and J. Wang, "Quantitative assessments of USARSim accuracy", Proceedings of PerMIS, 2006.

[14] S. Okamoto, K. Kurose, S. saga, K. Ohno, and S. Tadokoro, "Validation of Simulated Robots with Realistically Modeled Dimensions and Mass in USARSim", 2008 IEEE International Workshop of Safety, Security, and Rescue Robotics, pp. 77-82, 2008.

[15] R. Jones, J. Priddy, D. Horner, J. Peters, S. Howington, J. Ballard Jr, B. Gates, and C. Cummins, "Virtual Autonomous Navigation Environment (VANE)".

[16] C. Goodin, R. Kala, A. Caririllo, and L. Liu, "Sensor Modeling for the Virtual Autonomous Navigation Environment", 2009 IEEE Sensors, pp. 1588-1592, 2009.

[17] C. Cummins and G. McKinely, "A Three Dimensional Ground Contact Element for Wheels and Tracks", 2009 USACE Research and Development Conference, 2009.