Performance Evaluation of Intelligent Systems at the National Institute of Standards and Technology (NIST)

Craig Schlenoff, NIST, 100 Bureau Drive, Stop 8230, Gaithersburg, MD 20899
craig.schlenoff@nist.gov, 301-975-3456

Harry Scott, NIST, 100 Bureau Drive, Stop 8230, Gaithersburg, MD 20899
harry.scott@nist.gov, 301-975-3437

Stephen Balakirsky, NIST, 100 Bureau Drive, Stop 8230, Gaithersburg, MD 20899
stephen.balakirsky@nist.gov, 301-975-4791

Abstract:

The ability to comprehensively evaluate the quantitative and qualitative performance of an intelligent system is critical to accurately predicting how it will perform in various situations. The design of such evaluations is often as much of a research challenge as is the design of the intelligent systems themselves. Over the past decade, the Intelligent Systems Division (ISD), a part of the National Institute of Standards and Technology (NIST), has been at the forefront of assessing the performance of the various intelligent systems. This paper will give a broad overview of some of the evaluation efforts that have been pursued by ISD over the past few years, including performance evaluation of emergency response robots, sensor systems on unmanned ground vehicles, speech-to-speech translation systems, and the development of performance metrics for mixed-palletizing through the use of a simulation environment.

1.0 Introduction

As new technologies develop and mature, it becomes critical to provide both formative and summative assessments on their performance. Performance assessment events range in form from a few simple tests of key elements of the technology to highly complex and extensive evaluation exercises targeting specific levels and capabilities of the system under scrutiny. Typically, the more advanced the system, the more often performance evaluations are warranted and the more complex the evaluation planning.

Over the past decade, the Intelligent Systems Division (ISD), a part of the National Institute of Standards and Technology (NIST), has been at the forefront of assessing the performance of intelligent systems ranging from autonomous vehicles to urban search and rescue robots to speech translation and manufacturing systems. The evaluations have occurred in multiple environments including operationally-relevant field venues and simulated test environments. Evaluations range from examining the system as a whole to assessing very specific capabilities. In parallel, NIST has coordinated an annual Performance Metrics for Intelligent Systems (PerMIS) workshop (http://www.nist.gov/mel/isd/permis2010.cfm) to bring together colleagues in the field to explore challenges behind defining measures and methodologies of evaluating the performance of intelligent systems.

This paper gives a broad overview of some of the evaluation efforts that have been pursued by ISD over the past few years. Specifically, this paper is organized as follows:

- Section 2 describes a Department of Homeland Security (DHS)-funded effort to develop a comprehensive set of standard test methods and associated performance metrics to quantify key capabilities of emergency response robots.
- Section 3 describes an Army Research Lab (ARL)-funded effort to develop and execute algorithm and system technical evaluations leading to the definition and use of appropriate evaluation metrics, measurement methods and calibration methods to address the challenge of detecting, classifying, and tracking moving vehicles and people from an unmanned vehicle.
- Section 4 describes a Defense Advanced Research Projects Agency (DARPA)-funded effort to assess the performance of speech-to-speech translation systems. This was performed by bringing together US Military personnel and native foreign language speakers to immerse them in realistic environments where they would role-play relevant dialogues with the translation technologies to assess the performance of the systems as a whole.
- Section 5 describes a NIST-funded effort to develop performance metrics for mixed-palletizing through the use of a simulation environment. The goal of this effort is to demonstrate that performance metrics may be developed and initial system evaluations may be performed through the use of a low-cost open source simulation package.
- Section 6 concludes the paper.

2.0 DHS US&R Effort

In 2004, the Department of Homeland Security (DHS) asked the NIST to lead an effort to develop performance standards for robots that could assist responders in the very dangerous task of searching for victims after a major disaster, such as a building collapse or a hurricane. The DHS/NIST project seeks to aim technological progress in ways that expand robot capabilities for the benefit of emergency response applications. NIST has organized meetings to determine what responders need, organized tests to improve robots and works with groups to set standards so that soon rescue robots will be among the primary tools in an emergency situation. [1]

The standards under this project are being developed through a task group within ASTM International's Homeland Security Committee's Operational Equipment Subcommittee (E54.08). All standards being developed are based on requirements that members of the Federal Emergency Management Agency (FEMA) Urban Search and Rescue (US&R) Task Forces defined through a series of workshops hosted by NIST. The requirements were defined by teams that confront the most formidable disasters, but the results are intended to be useful to the entire range of the response community, from local departments on up.

Complementing the standards definition process is a series of field exercises in which FEMA US&R Task Force members deploy robots at FEMA training sites. Some of these sites are shown in Figure 1. These exercises allow responders to explore the potential of robots, understand their strengths and limitations, further refine their performance expectations and requirements, and develop concepts of operation. At the response robot exercises, test methods are tried out by the robot developers and the responders. To date, over sixty different models of robots – wall-climbers, ground, aerial, and underwater -- have taken part in the exercises. The diversity of robots serves to underscore the range of operational roles that robots will play.

Figure 1: Sample Sites Used in the DHS Effort

Because the robots will need such a wide spectrum of capabilities, the test methods under development emphasize quantifying performance of a particular capability along this spectrum and are not typically pass/fail. The performance required will depend on the role a search team would want the robot to play. For instance, one of the test methods that has become a standard is used to evaluate the visual acuity of the robot (shown in Figure 2). Typically, the robot is remotely controlled by an operator who uses a control station that displays views of what the robot's onboard cameras see. In this test, the operator sits in front of the control station and sees a view of standard eye charts relayed back from the robot's camera. The smallest line that the operator can successfully read is used to define the robot system's visual acuity. The test covers both near-field and far-field vision and is conducted under different lighting conditions, including darkness. For a robot that is to assist in evaluating structural stability, seeing very small features, perhaps at a distance (e.g., examining a crack from ceiling to floor

level) with no ambient lighting, is crucial. Therefore, if making a purchasing decision, a task force that will use the robot to assist the structural engineer will expect very high far field visual acuity under darkness. On the other hand, if a robot is expected to primarily be used to transport lumber or victims along a roadway, the visual acuity requirements would not be as stringent.

Figure 2: The DHS Visual Acuity Test

Other test methods under development measure the maximum distance at which a robot can effectively be controlled wirelessly, the power requirements (measuring the battery life), mobility over a range of terrain types, situational awareness when navigating an unknown environment, audio capabilities (can the robot's onboard microphone assist in locating victims?), and the manipulation capabilities of the robot (e.g., using its arm and a gripper to open doors or aim a sensor through a small hole). In all of these areas, the challenge that NIST and its partners in the standards process face is abstracting real-world complexities into simplified, repeatable, and easily reproducible test procedures and supporting artifacts.

Looking further into the future, robots will adopt more advanced capabilities, including producing maps of their environment as they explore and assistive autonomy features, such as independently navigating portions of their route. NIST has been infusing some of these more futuristic capabilities into the project, by featuring selected ones at the response robot exercises.

3.0 ARL Perception/Performance Evaluation Effort

The Army Research Laboratory (ARL) Robotics Collaborative Technology Alliance (CTA) conducted a multi-year effort to determine performance of robotic vehicle perception systems with a specific emphasis on human detection and tracking needed to enable safe operation around people and other moving objects. NIST developed and deployed test, measurement and analysis methods for this CTA

effort [2]. The CTA conducted several experiments for assessment and evaluation of multiple algorithms for real-time detection of pedestrians in Laser Detection and Ranging (LADAR) and video sensor data taken from a moving platform.

In these assessments, the robot vehicle was typically equipped with two pairs of stereo cameras, multiple scanning LADARs and line-scan lasers. The vehicle was driven by an operator, or driven autonomously, through routes of several hundred meters. Test runs included various configurations of moving pedestrians, fixed and moving mannequins, and various other fixed objects including other vehicles and foliage. In addition to the complexity of the environment, the variables included multiple robot vehicle speeds (30 km/h or 15 km/h) and pedestrian speeds (1.5 m/s or 3.0 m/s). A spectrum of environments and pedestrian behaviors ranged from relatively simple (straight roadway, few occlusions, simple pedestrian paths) to complex (NIST site with multiple structures and buildings and terrain types, many occluding objects, complex pedestrian behaviors). The more complex environments were intended to provide Military Operations in Urban Terrain, or MOUT, characteristics.

Key to assessing the perception algorithms is independent collection of ground truth data. An Ultra WideBand (UWB) system employed by NIST provided position tracking of the moving and stationary humans, the robot, and other objects. Improved performance of the CTA tracking and recognition algorithms has called for improvements in the ground truth solution. Ground truth accuracy demands have increased from over a meter to the current capability of about one quarter meter, with further improvements being pursued. Processing techniques were developed and implemented to produce higher quality tracking solutions than those provided by the raw data captured by the UWB. These processing elements include several filter and interpolation algorithms, and an algorithm for finding the correspondence between the ground truth data and the CTA tracking data from the multiple perception algorithms.

The tracking system uses state-of-the-art, UWB radio receivers posted around the perimeter of the test environment to track multiple static and dynamic targets with badge-size or smaller transmitters (shown in **Error! Reference source not found.**). For the CTA experiments, it was used to track vehicles and personnel throughout areas over 80 000 m^2 (19.8 acres) with an average accuracy of approximately 20 cm (8 inches) with an update rate of approximately 50 Hz, which supports tracking vehicles at highway speeds. Some structures, including those with concrete walls, present transmission problems. Additional receivers are placed to mitigate these situations. The total number of dynamic and static transmitter tags used simultaneously thus far has been approximately 15 dynamic tags and 30 static tags marking obstacles and known fiducial points to check accuracy.

(a) (b) (c)

Figure 3: (a) UWB receiver, 1 W and 30 mW transmitter tags, (b) a receiver deployed on mast and centered over known fiducial marker, (c) badge tags attached to helmets to track personnel in scenarios.

Figure 4 shows a plot of the tracking results for a ground truth system coverage and accuracy test on the NIST Center Drive. course. Green and orange plots show the vehicle path and the other plots show pedestrian tracks.

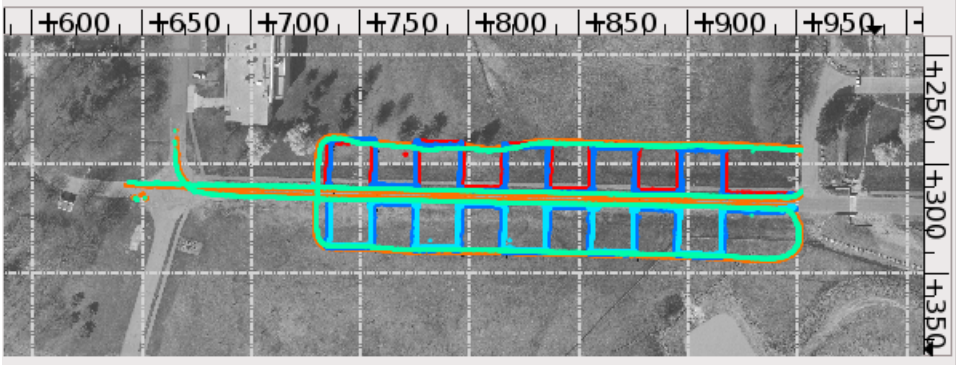

Figure 4. A calibration run with two transmitter tags mounted to a vehicle and two tags on each of two pedestrians to check coverage.

Data visualization is important for verifying the integrity of both the ground-truth data and the outputs of the CTA algorithms prior to, and during, the data collection. We developed an interactive viewer, CTAviewer, for this purpose (

Figure 5). The viewer uses various open source libraries and runs natively on Linux, Windows and Mac OS X. The viewer is used for displaying both the detection data from multiple perception algorithms and the corresponding ground-truth data. Individual datum can be toggled on or off by clicking on tag IDs or tracking IDs in the window. A slider control is especially valuable, allowing the user to move back and forth in time to see the detection plot at any chosen instant. This is used often to replay a run by moving the slider from left to right at a convenient rate while observing the detections as they occur in the data.

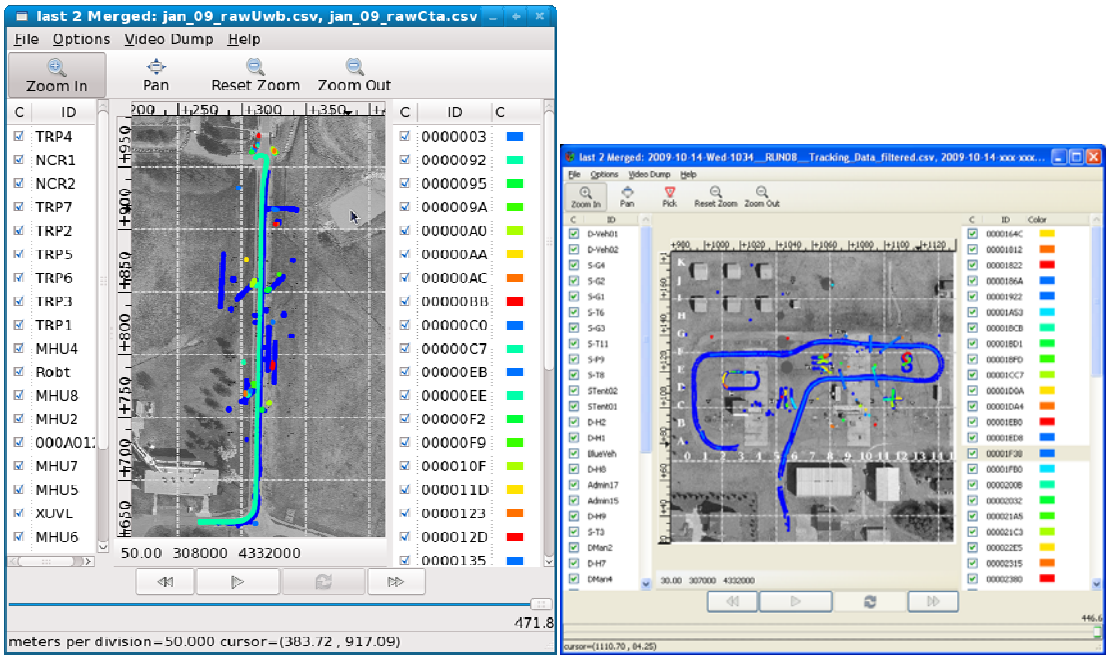

Figure 5: CTAviewer screenshots of NIST Center Drive (l) and the more complex site (r) showing both detection data and ground-truth. The left panel lists the tag ID and the right panel lists the tracking ID.

Post processing of the collected data results in a spreadsheet for each perception algorithm with metrics for analysis. A record is formed for each algorithm-reported entity believed to be a human. Each algorithm assigns an identifier to an entity on the course classified by the algorithm to be a human. All information related to that algorithm identification is condensed to a single record. This record may hold information from many cycles of the algorithm. Post processing determines whether that entity is, in truth, a human or mannequin (true positive), another known course entity not human or mannequin (misclassification), or an unknown course feature with no associated ground truth (false positive). Distinctions are also made between moving and stationary entities and various classes of nonhuman entities (e.g., barrels, cones, crates). Field notes describe test conditions under which the data were collected, absolute and relative positioning of the robot platform and detected entities recorded at the time detections first occurred for an identification, time and cycle number indicators of the persistence of detection, and the accuracy of the algorithm classification decision.

The described advances in measurement technology improve the assessment process markedly. The ground truth precision provides an objective evaluation of the results reported by the algorithms. It makes possible the exact tracking of moving entities on the course, essential given the planned

assessment of the "detection and tracking" purposes of the algorithms. This was previously not possible. The CTA viewer has proven to be not only a useful tool in visual analytics, but has also provided an instant check during the conduct of the experiment as to whether or not data are being collected and whether systems are in good calibration.

We expect to continue to use the described capabilities in future CTA work. We are continuing research in improving the processing and analysis algorithms and software, in extending the visualization capabilities, and in enhancing tracking in difficult environments. Further, we are applying these capabilities to other projects.

4.0 DARPA TRANSTAC Effort[1]

One of the most difficult challenges that military personnel face when operating in foreign countries is clear and successful communication with the local population. To address this issue, the Defense Advanced Research Projects Agency (DARPA) is funding academic institutions and industrial organizations through the Spoken Language Communication and Translation System for Tactical Use (TRANSTAC) program. The goal of the TRANSTAC program is to demonstrate capabilities to rapidly develop and field two-way, speech-to-speech translation systems that enable speakers of different languages to communicate with one another in real-world tactical situations without an interpreter. Evaluations of these technologies are a significant part of the program and DARPA has asked NIST to lead this effort [3].

All of the TRANSTAC systems work in the following way. When the English speaker speaks into the system, the Automatic Speech Recognition (ASR) component of the TRANSTAC system analyzes the speech to recognize what was said and generates a textual transcription of the speech. The Machine Translation (MT) component of the TRANSTAC system next translates that text file from the source language to the target foreign language. Finally, the Text-To-Speech (TTS) component of the TRANSTAC system converts the textual target language translation into speech, which is then spoken to the foreign language speaker. This same process happens when the foreign language speaker speaks and the system translates from the foreign language into English.

To evaluate the performance of these translation systems, the evaluation team implemented the System, Component, and Operationally-Relevant Evaluation (SCORE) framework [4] which has been developed at NIST over the past three years to provide formative evaluations of advanced technologies that are still under development. Using SCORE, the evaluation team produced an evaluation design to capture both quantitative technical performance and qualitative utility assessments of the TRANSTAC systems. NIST implemented a multi-faceted testing methodology which included scenarios performed by representative live speakers using the translation systems (shown in Figure 6), and a separate evaluation using pre-recorded utterances, which we refer to as an *offline* evaluation.

[1] The views, opinions, and/or findings contained in this article/presentation are those of the author/presenter and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense.

Figure 6: Military personnel and native foreign language speakers testing the TRANSTAC systems

Scenarios were developed to focus on capturing the utility and usability for the end users of the various platforms. Both the English and foreign language speakers were given realistic and relevant character motivations which they used to produce spontaneous conversations, using the translation from the TRANSTAC systems to communicate with each other.

For the offline evaluation, prerecorded utterances were fed into the TRANSTAC systems, first in audio format to test the systems' performance using ASR followed by MT and then in accurately transcribed text format (in effect, with perfect ASR) to test the systems' MT performance in isolation. In contrast to the live evaluations, the offline evaluation gives each system exactly the same inputs, enabling a true apples-to-apples comparison.

A variety of metrics were used to gain a comprehensive understanding of the capabilities of the TRANSTAC systems. The metrics included:

1. **High-Level Concept Transfer:** A count of the number of utterances that were properly translated from one language to the other, as judged by a panel of bilingual judges. This number is divided by the time it took to get through the utterances to produce a high-level concept transfer rate.
2. **Likert Judgment:** A judgment of the semantic adequacy of the translations, scored one at a time by a panel of bilingual judges. A numerical scoring range was used where +3 is completely adequate, +1 is tending adequate, -1 is tending inadequate, and -3 is inadequate.
3. **Low-Level Concept Transfer:** A quantitative measure of the transfer of the low-level elements of meaning in each utterance. In this context, a low-level concept is a specific content word (or words) in an utterance. For example, the phrase "The house is down the street from the mosque." is one high-level concept, but is made up of three low-level concepts (house, down the street, mosque). A panel of bilingual judges provide these assessments and the scores are averaged. [5]
4. **Automated Metrics:** A suite of well-accepted automated metrics were used. For speech recognition, we calculated Word-Error-Rate (WER). For machine translation, we calculated BLEU [6] and METEOR [7] using four reference translations.
5. **TTS Evaluation:** To assess the performance of a TTS component, human judges listened to the audio outputs of the TTS evaluation and compared them to the text string of what was fed into the TTS

engine. They then gave a Likert score from 1 to 5 (five being the best) to indicate how understandable the audio file was in comparison to what was fed into it.

6. **Surveys/Semi-Structured Interviews:** After each live scenario, the military personnel and the foreign language speakers filled out a detailed survey asking them about their experiences with the TRANSTAC systems. In addition, semi-structured interviews were performed with all of the participants in which questions such as "What did you like?, What didn't you like? and What would you change?" were explored.

5.0 USARSim/MOAST Effort

Stacking objects onto pallets is the most widely used methods of bulk shipping, accounting for over 60% of the volume of goods shipped worldwide. One example of this problem set is the distribution of packed grocery items to various retailers. The shipment may be decomposed by class of goods (e.g. milk or cookie of brand x) and arranged by workers such that each class is on its own pallet. However, for some vendors and retailers, a full pallet of a class of goods would exceed their total demand. To solve this problem, various commercial logistics solutions allow products to be shipped in mixed pallet loads, where multiple classes of products are grouped onto a single pallet. Most of these solutions use heuristic approaches or formulate the problem as a mixed integer linear program to solve the manufacturer's bin packing problem. However the heuristics used in these problems are statistical, and there is no way to know if a pallet can be created at all. In addition, there are no industry-wide standards or metrics that dictate what comprises a "good" pallet, nor an accepted way to present the information required to formulate a pallet representation. Roughly speaking, a metric for palletizing is a quantitative measure of some aspect of any of the following:

- one package that is part of a stack on a pallet
- the entire collection of packages in a stack on a pallet
- a set of stacked pallets
- the process of building stack(s) of packages on pallet(s).

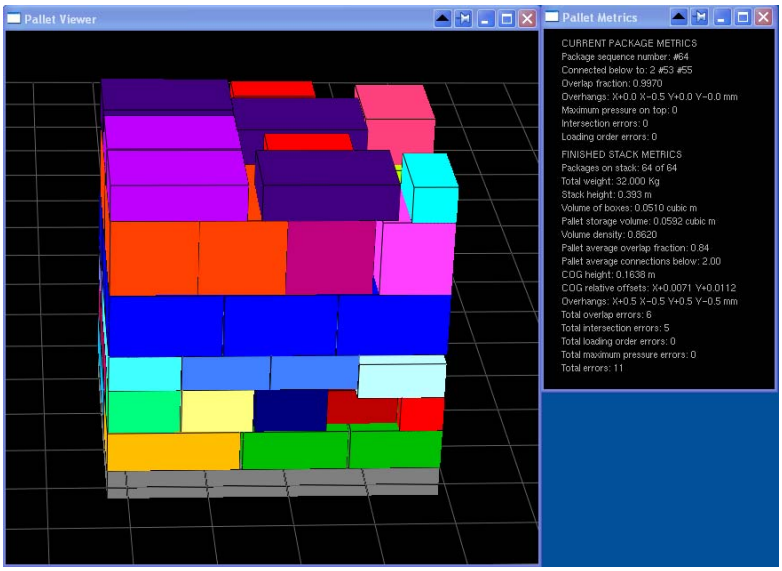

Figure 7: Screen shot from NIST's pallet viewer software. Potential metrics are computed as the pallet is constructed and their values are displayed in the window on the right of the figure.

The authors decomposed the above metric areas into the three distinct phases of static simulation, dynamic simulation, and scaled operation. For all phases of the evaluation, the system under test was required to produce a pallet build plan that conforms to NIST's XML-based pallet build schema. In the static simulation phase, a newly created pallet quality evaluation simulator known as Pallet Viewer was utilized to judge the quality of the proposed finished product. This simulation judges metrics aimed at both individual packages that comprise a pallet stack as well as the overall pallet. As shown in Figure 7, the Pallet Viewer utility displays a 3D color view of a pallet and the as-planned stack of packages on it. In addition, the Pallet Viewer currently calculates and displays 6 metrics for the individual packages and 15 metrics for the as-planned stack. As our understanding of the metrics improves, metrics may be added or removed from this simulation. Detailed information on the currently evaluated metrics may be found in Balakirsky et. al [8].

Figure 8: Pallet under construction in USARSim.

The second phase of the evaluation process involved the dynamic construction of pallets in simulation and judging of the process of building the pallets. For this effort, the Unified System for Automation and Robot Simulation (USARSim) [9] was utilized. This test aimed to determine if dynamic aspects of the pallet construction were valid. For example, the schema calls for approach points for the delivery of each package along with the package's final resting position to be computed. The static simulation is able to judge the quality of the final resting position, while a dynamic simulation is required to determine if the approach points will safely deliver the package to the desired location. Ground truth from the simulation was utilized to construct an "as-built" file for each pallet. This as-built file was then fed into the Pallet Viewer software for comparison with the desired build plan to determine if a stable build solution was achieved.

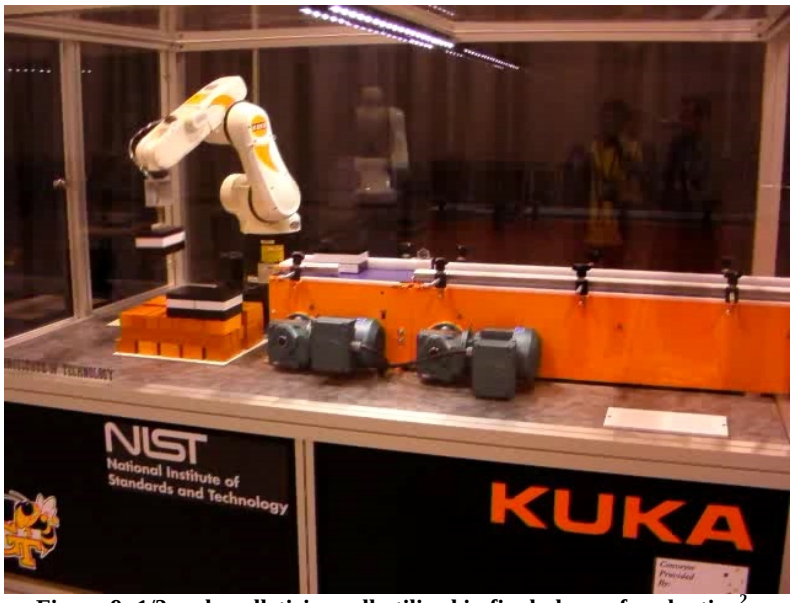

Figure 9: 1/3 scale palletizing cell utilized in final phase of evaluation[2].

The final piece of the evaluation tied together all aspects of the pallet build process by constructing actual pallets on a 1/3 scale palletizing cell. This allowed human observers to judge the quality of the pallet construction as well as the final completed pallet. Future work in this area will allow us to compare human judgment of pallet quality with our automatically generated metrics. This will allow for the determination of a set of metrics that will accurately predict the quality of the mixed pallets.

This evaluation represents an ongoing effort. Research teams are formulating new approaches to the NP-hard mixed pallet problem, and will be competing against each other at the IEEE Robot Challenge that is part of the International Conference on Robotics and Automation. More information on the challenge and on how to become involved may be found at http://www.vma-competition.com.

6.0 Conclusion

As described in this paper, it is the authors' firm belief that the design of an effective performance evaluation is as much of a research challenge as the development of the technology itself. This paper describes four evaluation efforts that are ongoing at NIST which have developed approaches to characterize the performance of very different types of intelligent systems, including search and rescue robots, robotic vehicle perception systems, speech translation systems, and manufacturing mixed palletizing. There are many other performance evaluation efforts which are on-going in the Intelligent

[2] Certain commercial products and software are identified in this paper in order to explain our research. Such identification does not imply recommendation or endorsement by NIST, nor does it imply that the products and software identified are necessarily the best available for the purpose.

Systems Division which could not be included in this paper due to space limitations. To find out more about these, please contact the authors.

Bibliography

[1] E. Messina, "Robots to the Rescue," *Crisis Response Journal*, vol. 5, no. 3, pp. 42-43, 2009.

[2] Bodt B., R. Camden, H. Scott, A. Jacoff, Hong T., T. Chang, R. Norcross, T. Downs, and A. Virts, "Performance Measurements for Evaluating Static and Dynamic Multiple Human Detection and Tracking Systems in Unstructured Environments," in *Proceedings of the 2009 Performance Metrics for Intelligent Systems Conference* Gaithersburg, MD: 2009.

[3] C. Schlenoff, B. Weiss, M. Steves, G. Sanders, F. Proctor, and A. Virts, "Evaluating Speech Translation Systems: Applying SCORE to TRANSTAC Technologies," in *Proceedings of the 2009 Performance Metrics for Intelligent Systems (PerMIS) Conference* 2009.

[4] C. Schlenoff, "Applying the Systems, Component and Operationally-Relevant Evaluations (SCORE) Framework to Evaluate Advanced Military Technologies," *ITEA Journal of Test and Evaluation*, vol. 31, no. 1 Feb.2010.

[5] G. Sanders, S. Bronsart, S. Condon, and C. Schlenoff, "Odds of Successful Transfer of Low-Level Concepts: A Key Metric for Bidirectional Speech-to-Speech Machine Translation in DARPA's TRANSTAC Program," in *Proceedings of the LREC 2008 Conference* Morocco: 2008.

[6] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)* Philadelphia, PA: 2002, pp. 311-318.

[7] S. Condon, J. Phillips, C. Doran, J. Aberdeen, D. Parvaz, B. Oshika, G. Sanders, and C. Schlenoff, "Applying Automated Metrics to Speech Translation Dialogs," in *Proceedings of the LREC 2008 Conference* Morocco: 2008.

Craig Schlenoff Biography

Craig Schlenoff is the Acting Group Leader of the Knowledge Systems Group in the Intelligent Systems Division at the National Institute of Standards and Technology. His research includes performance evaluation techniques applied to autonomous systems and manufacturing as well as research in knowledge representation/ontologies. He previously served as the program manager for the Process Engineering Program at NIST and the Director of Ontologies at VerticalNet. He leads numerous million-dollar projects, dealing with performance evaluation of advanced military technologies. He received his

Bachelors degree from the University of Maryland and his Masters degree from Rensselaer Polytechnic Institute, both in mechanical engineering.

Harry Scott Biography

Harry Scott is currently Acting Group Leader of the Machine Systems Group of the NIST Manufacturing Engineering Laboratory's Intelligent Systems Division (ISD). He is program manager for ISD's support of an Army Research Laboratory effort to characterize performance of autonomous vehicles, most recently with respect to perception and pedestrian detection capabilities, and for a Federal Highway Administration Exploratory Advance Research Program studying driver visibility requirements for driving curved roads at night. He applies his expertise in intelligent systems, autonomous vehicles, geomatics, architectures, and GPS, inertial, ultra wide band and other measurement systems, to develop ground truth and performance measures for these and other programs.

Stephen Balakirsky Biography

Dr. Balakirsky is the Project Manager for Advanced Simulation in the Intelligent Systems Division at the National Institute of Standards and Technology. He has over 20 years of experience in multiple areas of robotic systems with his current research focusing on robotic simulation, multi-agent behaviors, world modeling, and robotic performance evaluation. Dr. Balakirsky is the principal investigator of the IEEE Virtual Manufacturing and Automation Competition, and on the executive committee of the RoboCup Federation. He received his Doctor of Engineering degree from the University of Bremen and his Master's and Bachelor's degrees from the University of Maryland, College Park.