Performance Measurement and Benchmarking Robotic Assembly with NIST Task Boards

I. Overview

The goals for the NIST task boards include facilitating performance measurement of robotic assembly and promoting benchmarking among robotic systems. The foundations for accomplishing these goals include a unified set of assembly tasks, objects, and test methods accompanied by specified performance metrics and statistical analyses. Furthermore, the test designs are specifically low-cost and internationally acquirable. For more details on current task board designs, adoption, publications, and pertinent downloads go to the NIST website on performance measurement for robotic assembly.

II. Design Philosophy

Since any assembly fundamentally consists of a series of assembly *operations*, the NIST assembly task boards are designed around these operations (e.g., simple insertions, threading, snap-fitting, meshing, routing). The process was begun with an analysis of assembly tasks through the lens of measurement science to uncover key metrics and potential test methods [1]. Unfortunately, the design space for assemblies are limitless, and therefore, converging on a relevant and attractive task board design is quite challenging. To narrow the possibilities, the task board designs are the intersection of three main criteria: 1) includes only real-world, standardized components that are low-cost and internationally acquirable; 2) targets components with differing human performance levels (completion time) based on existing design-for-assembly research [2]; and 3) components are reasonable for grasping and manipulation by existing robot systems. Clearly, reporting robot performance for any one task board cannot fully portray the performance for that system. However, repeated testing on these task boards and other benchmark tasks (e.g., YCB object set) can serve as a basis for system comparison, and, when conducted over a wide variety of tasks, can more accurately capture robot performance.

III. Test Method

Regardless of the task board, there exist two principal test modes – disassembly and assembly. The tests are intended for evaluating **integrated robot system performance,** including the perception and localization of the task board, components, and destination or source bin.

a.) Disassembly setup and process: To test a robot's disassembly capabilities, randomly place both the fully assembled task board and destination bin (for parts) within the dexterous workspace of the robot. At a minimum, the task board and bin planar configurations should change (two translational components and one rotational component). Once configured, there should be no human intervention (e.g., lead-through programming) and the robot system should remain autonomous. Following, the robot system should move to, grasp, disassemble, and transport all target components from task board to target destination bin. Components may be engaged in any order. Any type of manual interference, e.g., physical, teleoperative, or via remote input, by a human operator occurs, then the trial is considered void and the test must be reset to starting conditions. To encourage robust systems, neither the task board or bin should be rigidly

fixed or adhered to surfaces. If the board or bin are unintentionally displaced by the robot during an operation, the robot system must automatically adjust to compensate for the state change.

b.) Assembly setup and process: To test a robot's assembly capabilities, randomly place both the initialized task board and kit of parts for assembly within the dexterous workspace of the robot. At a minimum, the task board and kit planar configurations should change (two translational and one rotational components). Once configured, there should be no human intervention (e.g., lead-through programming) and the robot system should remain autonomous. Following, the robot system should move to, grasp, transport, and assemble all components from kit to task board. Components may be engaged in any order. If manual interference, through physical, teleoperative, remote input, or otherwise, between a human operator and robot occur, then the trial is considered void and should be reset. To encourage robust systems, the task board, kit, or components should not be rigidly fixed or adhered to surfaces. If the board, kit, or kit parts are unintentionally displaced by the robot during an operation, the robot system must automatically adjust to compensate for the state change.

c.) Significance testing: In order to instill confidence in the performance assessment of a robotic system and allow for the application of various statistical tests, many trials of disassembly or assembly must be conducted per task board. Unfortunately, the number of trials depends on the variability of robot performance, performance requirements, and cost per trial. However, it is recommended that at least 30 trials are conducted for good power of subsequent statistical tests (otherwise, increased likelihood of false-negatives will exist). Note, every conducted trial must consist of a new, random placement of task board, kit, and bin within the robot's dexterous workspace.

IV. Performance Metrics

The bottom line for the performance of any robot system for any task constitutes primarily of *speed* and *reliability*. For these task boards, speed is reflected by the **completion time** of a grasping operation, an assembly operation, and the entire task board. Reliability is reflected by the **probability of successfully** grasping an object and completing an assembly operation, and the **degree** to which a task board was completely disassembled or assembled. We acknowledge that there are other good tertiary metrics as well, e.g., exerted forces/torques; however, these come with a significant additional cost of test equipment and, in many applications, are only significant once speed and reliability requirements are met. Therefore, we currently exclude these metrics.

a.) Operation- and object-centric metrics: capturing performance per grasping and assembly operation per part improves granularity and insight on robot system capabilities. **Completion time** and **binary pass-fail** should be recorded for every move-grasp-transport sequence (per part) and every assembly operation sequence (per part) for the task board assembly mode. **Completion time** and **binary pass-fail** should be recorded for every move-disassemble sequence (per part) and every transport-place sequence (per part) for the task board disassembly mode.

b.) Mode- and board-centric metrics: capturing the degree to which a task board was successfully disassembled and assembled provides a more easily interpretable, holistic

viewpoint. **Completion time** should be recorded from start-to-finish for a robot system disassembling or assembling a task board. The **percentage** of parts successfully transported to their final destination during disassembly should be recorded. The **percentage** of parts successfully installed on the task board during assembly should be recorded.

V. Data Analyses and Benchmarking

After many trials have been conducted as previously stated, the collected performance data is ready for statistical analysis. The metrics suggested herein fall into one of two categories – attribute or continuous data.

a) Attribute (pass-fail) data:

This data type can be analyzed in two different ways. First, the theoretical upper bound probability for successfully inserting a component (PS) is calculated given a confidence level (CL), the number of successes (m), and the number of independent trials (n). Given the binomial cumulative distribution function,

$$F(m-1; n, PS) = \sum_{i=0}^{m-1} \binom{n}{i} PS^i (1-PS)^{n-i} \geq CL,$$

where the PS is its minimum value to some precision while still satisfying the above inequality. More information regarding this calculation can be found in [3]. The most significant quality about this calculation is that it promotes conducting many trials to instill more confidence in the assessment of robot reliability.

The second analysis involves calculating whether there exists a statistically significant difference between any two calculations of PS (e.g., comparing PS values between two different robot systems). In reality, any two calculations of PS can be different (even marginally different), and the next line of inquiry should be whether the observed difference is statistically significant. This assessment helps reduce the likelihood of false-positive or false-negative assessments, the existence of which only obscures one's ability to see true improvement or change. One algorithm for conducting this assessment includes the Kolmogorov-Conover algorithm [4]. This algorithm also applies to ordinal data as well. Implementations of this algorithm can be obtained from NIST's website on performance data analytics.

b) Continuous Data:

This data type can be analyzed in three different ways. Since trials are conducted with random placements of task board, kits, and bins, the returned performance data is likely independent, satisfying an underlying assumption of these statistical tests.

The first test involves an analysis of the *distribution* of performance data acquired over many trials. A distributional analysis can serve as a first-line indicator for whether there exists a

significant difference between any two sets of data. There exist many different algorithms for performing such a test, but one popular, non-parametric method is the two-sample Kolmogorov-Smirnov test. Implementations of this algorithm exist in many code bases including Matlab, R, and can also be obtained from NIST's website on performance data analytics.

The second test involves an analysis of variance (ANOVA) between two sets of data. An ANOVA test will indicate whether the variance of data in one set is significantly different from that of another set. A smaller variance indicates better precision in robot performance, a desirable trait. This test is a precursor to means testing, and many algorithms exist for conducting this test including the Levene test. Several test statistics for this test exist. However, without any assumption of the underlying distribution of the data, the data based on medians (Brown-Forsythe statistic) yields relatively good robustness and power. Implementations of this algorithm exist including Matlab, R, and can also be obtained from NIST's website on performance data analytics.

The third test involves means testing that indicates whether the mean of data in one set is significantly different from that of another set. Average calculations yield insight on the expected performance of a robot system. There are many algorithms for conducting means testing, including the Student's t-test. Again, implementations of t-test algorithms exist in Matlab, R, and can also be obtained from NIST's website on performance data analytics.

It is recommended that the above data analysis methods be applied to robot performance data when benchmarking systems. Ultimately, these in-depth analyses will help mitigate the issuance of false statements regarding robot system capabilities, and help guide the advancement of such systems with more accurate comparative feedback. Refer to [5] and [6] for example applications of these algorithms on robot performance data.

References

[1] Shneier, Michael O., et al. "Measuring and Representing the Performance of Manufacturing Assembly Robots." *NIST Interagency/Internal Report (NISTIR)-8090* (2015).

[2] Boothroyd, Geoffrey, Peter Dewhurst, and Winston Anthony Knight. *Product Design for Manufacture and Assembly, revised and expanded*. CRC press, 1994.

[3] Gilliam, David, et al. "Pass-fail testing: Statistical requirements and interpretations." *Journal of research of the National Institute of Standards and Technology* 114.3 (2009): 195.

[4] Conover, William J. "A Kolmogorov goodness-of-fit test for discontinuous distributions." *Journal of the American Statistical Association* 67.339 (1972): 591-596.

[5] Van Wyk, Karl, and Jeremy A. Marvel. "Strategies for Improving and Evaluating Robot Registration Performance." *IEEE Transactions on Automation Science and Engineering* (2017).

[6] Van Wyk, Karl, et al. “Comparative Peg-in-Hole Testing of a Force-based Manipulation Controlled Robotic Hand.” *IEEE Transactions on Robotics (2018)*, to appear.