Meeting Transcript
Grasping and Manipulation Performance Measures and Benchmarking
Held March 1, 2018

Notes:

1) This meeting covered the materials in the NIST presentation, NIST_Force_Grasp_Manipulate_Performance_Master_RHGM.pdf.
2) Related links:
https://www.nist.gov/programs-projects/performance-metrics-and-benchmarks-advance-state-robotic-grasping
https://www.nist.gov/programs-projects/performance-metrics-and-benchmarks-advance-state-robotic-assembly
3) The following transcript often paraphrased statements made.

Joe Falco:

The purpose of this presentation is to introduce to some of the work we are doing in the area of grasping and manipulation performance measures and benchmarking. And try to put together a coordinated effort under RHGM. There are a lot of people out there doing things – putting together benchmarks – and we think it would be great if we could all work together towards a common set of benchmarks. And this set of benchmarks could just be something you could pick and choose from depending on what you are doing, and it could be used for your own work, but also be compared with the results from others working on similar efforts.

We are giving you a snapshot of what NIST is doing to jumpstart talks including what others are doing, and identify overlaps and common interests. We are giving a quick overview of NIST benchmarks, including at the elemental level, which just includes a robot hand/gripper. NIST is primarily interested in industrial application space so our work is focused mostly on assembly at the task level. We also want to discuss parallel efforts in benchmarking, and hope to discuss steps forward towards a unified effort.

Our goal as NIST is to develop metrics, test methods, and artifacts with example datasets to characterize the performance of grasping and manipulation emphasis on deployment on manufacturing tasks. We think this will help provide the robotics community with unbiased measurement for both elemental characteristics and functional-level performance capabilities. In the short term, we think this will provide researchers and developers insight for improving their hardware and software designs. In the long term, we would like to see this work evolve into a set of specifications that will help match capabilities to end-user manufacturing needs.

For NIST, when we develop these test methods and benchmarking, we need to have various sensors, arms, and hands available for testing. NIST has a variety of hardware in these areas.

We have ten elemental gripper/hand tests that measure features starting from finger strength all the way to in-hand manipulation. A variety of sensors are using including single-axis load cells embedded within a split-cylinder artifact. We try to make these sensors easy to replicate and affordable. For example, we have also used force sensor resistors instead of load cells to reduce cost. We use motion low-cost, precalibrated motion capture systems for measuring in-hand manipulation and object pose estimation performance.

For example, touch sensitivity is a measure of the smallest, self-registered contact force exerted by the robotic finger on an object. This is particularly useful for imbuing robot hands with the ability to minimally disturb an object during part acquisition. It's an integration of raw sensor capability and the controllers of the hand. An example of using the split cylinder artifact is for grasp strength. Closing a hand around the cylinder provides an estimated of the maximum force the hand can impart on an object, and can be used to estimate the payload of a robotic hand. This could be used in a future technical specification to help make a decision regarding hands for your particular application. Finally, on a more complicated side, we do a measure of in-hand manipulation. This a measure of a robotic hand's ability to control the pose of an object. This can quantify range-of-motion, frequency response, and control accuracy.

Karl Van Wyk:

Functional-level performance testing is a departure from the tests we have just seen, and includes an arm, a gripper/hand, a vision system, and higher-level logic for completing a task. We are particularly interested in measuring performance of assembly operations including pick-place, insertions, fastening, meshing, harnessing. We think this will be useful for comparing systems as a whole, so the systems can be completely different. We also think that this could be used for component testing. So, if you are an algorithms designer and you are interested in determining whether some algorithms work better than others, you can swap these out during the tests to determine whether you see differences or not.

We are interested in task-level tests, and one of the first questions one should ask is how do you design the tests in a defensible way. We have decided to focus on assembly operations. This nice thing about this route, you can use existing design for assembly (DFA) research that methodically quantifies the performance of human when they perform various pick-place and insertion tasks. A snapshot of a series of available tables from existing DFA research that quantify human performance. Principal measure is completion time – handling time (object grasping and transport), and insertion time. DFA research can help guide the design space, so instead of being arbitrary for what you are picking for a test, you can look at these tables and sample around. Another nice thing from using this data is that you can forward calculated a theoretical human performance for a task without conducting human trials as a reference point of comparison against robot systems.

We envision two principal test modes: disassembly and assembly, quantifying how well a robot can take something apart, and how well it can put it back together. Perhaps once nice thing about disassembly is that you can also quantify the accuracy of part placement in a location after it has been removed from the assembly. Primary performance metrics govern speed and reliability. Speed is captured by completion time, and reliability by probability of success. There exist good tertiary metrics that measure contact forces during assembly processes, but these come at a significant added cost from instrumentation. Also, this metric likely only becomes relevant after performance in speed and reliability are good, so for now, we have left it off. We aim report metrics at two different levels of granularity – per-part/operation or as a whole. Reporting performance as a whole gives a good high-level impression of how well a system did, but researchers and practitioners will need to look at performance at the low-level as well to find out where things went well and where things went poorly.

Something we like to emphasize is the use of statistical analyses on the returned robot performance data. We think it's very important to instill confidence in performance assessment and reduce the likelihood of seeing a difference or improvement when it really is not significant. Regardless of the data type, attribute, ordinal, or continuous, there exist statistical tests for more indicating significant differences. We indicate some of the tests we use, and many are available in Matlab, R, or other packages.

Let's look at an example test around peg-in-hole. This is a very inexpensive test that you can 3D print, and the desire is to gauge robot performance at very basic insertions. One thing to notice is that the actual parameters behind the test design are backed by human data. For example, the spacing of 35 cm is the area where humans primarily conduct their primary assembly processes, namely, insertion, threading, outside of that, you are doing mostly coarse grasping and transport of parts. The design layout is triangular, which facilitates cyclical testing for more repetitions. One thing that we have added is the application of offline-generated offsets to the starting point of the insertion for the robot systems, and were generated from Gaussian distributions. They are simulations of error applied to the initial alignment of the peg and the hole. We have added this to help test the robustness of insertion strategies with respect to alignment errors. We looked at two different systems, a position controlled arm with a 16 degree of freedom and actuation robotic hand retrofitted with 6-axis force/torque transducers at the fingertips. The hand is controlling the peg with force control. The second system is an impedance controlled arm with a stiff, pneumatic parallel gripper. Exploiting the compliance of the arm, you can engage various planar search techniques for seating the peg into the hole.

The supplied video reveals what it looks like when these systems perform the insertion task. Keep in mind, that they are intentionally offset prior to their first insertion attempt, so they often times have to engage some sort of search or alignment strategy to perform the insertion. The two primary measures of performance were time to complete a grasp-transport-insert operation and probability of successfully completing an insertion. For the time metric, you can start by testing for autocorrelation to ensure that the performance values are sufficiently independent. This independence is imparted by the intentional offsetting of systems by random samples of Gaussian distributions. Independence is a requirement of statistical tests. As you can see, all three variations of System 2 exhibited statistically significant differences in the distribution of time performance with respect to System 1. Although, only some cases of mean and variance of time from System 2 are significant with respect to System 1. For example, System 2 *Random*, the mean insertion time of 15.62 seconds, which is close to System 1's 18.31 seconds, are not yet statistically significant even though they look different. The same is true for the probability of success measure, where System 1 failed a couple of times when compared to System 2, but it is not yet statistically significant based on the current number of repetitions.

That was a fairly simple test, and now we are trying to instill more variation in the tests. Our design paradigm is behind a set of themed task boards. The task boards focus on a set of assembly facets. We now have a task board that focuses on peg insertions, threading, meshing, and connecting plugs, with two other task boards that focus on wire harnessing, and pulley seating and routing. Again, these are designed with reference to DFA; we are not randomly picking parts. We also try to ensure that the overall cost for replicating these task boards are relatively low and that the task boards can be replicated internationally with internationally acquirable components. We also ensure that the components are real, existing components, so we are moving away from 3D printed objects. Task Board #1 is now a board that has been physically created, and focuses on simple insertions, nut threading, gear meshing, and connecting plugs. Again, the design is an intersection of pulling from DFA tables, using real, low-cost components, and is internationally replicable. A variation of this design was at the IROS 2017 grasping and manipulation competition so we could see how well teams from around the world could perform this task. This is just one avenue for proliferating the tests, and a video will be shown later. Around 20 of these task boards have been created and distributed to various research entities and companies, and we are hoping to see reports on their findings soon. Along with the task board, we are also created "kits" that is just a flat tray with contours of the objects for insertion. The objects are placed on their respective contours at the beginning of the test to control their initial conditions. You can generate as many different kits as you like for testing. CAD models, test setups, design, datasets on all the previously discussed elemental and functional level tests are available at the two links provided.

Joe Falco:

We keep calling ourselves NIST, which stands for National Institute of Standards and Technology. Our goal is standards for technology. Even though hands/grippers are still in the early stages of maturity, that does not mean that we shouldn't look ahead at standards for aiding development. As we mentioned today, one approach we are taking include NIST special publications, which are documents that we are making publicly available to the community, and in this case we have been working under IEEE RHGM technical committee. One document covers terminology, so we can all understand each other. The other is a test method document, geared at the elemental grasp test methods and benchmarking work. One thing we will do is circulate these documents, and we would really appreciate input on the content. We also have several publications out there on this work for your reference.

One of the ways we promote this work is through competitions. While the benchmarking methods allow you go test in great detail in a laboratory setting, we like to introduce them to the research community with competitions. We recently started off with IROS 2016 grasping and manipulation competition that had some manufacturing tasks. The IROS 2017 grasping and manipulation competition had a dedicated manufacturing track. We are also working with Japan on the World Robot Summit (WRS) Industrial Robotics track 2018 trials, and again, in 2020 for the finals. We might be doing an IROS 2019 competition as well.

The IROS 2017 grasping and manipulation competition, manufacturing track had two tasks. One was the already described task board, and the other was a full gear assembly that naturally included various insertions, fitting, and meshing. The parts were quite challenging, I think we would have like to see more perception. Most robots were lead-through programmed. However, as you will see at WRS, one of the goals is to promote full robot autonomy in an agile manufacturing environment.

Here is the 2018 proposed assembly part for WRS, Industrial robots track. It is a pulley assembly with a flexible belt and small parts which should prove to be quite challenging. We will be sending all of you a link to some promotional materials regarding this competition.

We are aware of some related work including the Yale-CMU-Berkeley (YCB) object benchmarks. The Advanced Robotics for Manufacturing (ARM) Institute has interest in performance measurement and benchmarking. Berkeley is leading an open discussion about robot grasping benchmarks and metrics. We also have Adam (Norton) from UMass Lowell - NERVE center and he will talk about the work they are doing. We would like to follow up with a discussion of unifying our efforts. We would be more than happy to post some regular meetings under the umbrella of IEEE RHGM. We can do this in the format we have today, people can present the work they are doing. We can determine the appropriate frequency, perhaps quarterly. We can even break up into sub-focus areas that could meet independently. We can have a yearly face-to-face meeting at an IEEE robotics conference as well. We would like to use this group to create a consensus on the benchmarking work within the robotics community. Working publications could be used as precursor to standards efforts. At this time, we would like to open the floor to anyone who would like to talk about their work.

Berk Calli

I would like to pre-announce a special issue geared towards benchmarks for robotic manipulation. We are proposing it together with Maximo Roa and Aaron Dollar. We propose benchmarking protocols along with a baseline result. We are still in the proposal phase of it, but I just wanted to pre-announce it. Of course, we will be announcing it in Robotics Worldwide as well.

Hyungpil Moon

I suggest a test for quantifying how well a robot hand can deal with tools. If you look at the IROS 2017 and WRS competition – those tests are suitable for robots with parallel gripper end-effectors, not for humanlike advanced hands. I wonder, what would be a better test for those advanced hands.

Joe Falco:

I guess you would have to design a test that forces the use of tools. If you look at the IROS 2017 competition, I suppose that most of the tasks did not require the use of tools. Yes, I think that a functional test that requires the use of tools could be useful.

Karl Van Wyk:

What you'll find with the assembly tests, the use of tools are allowed, so you could have a robot system that tries to turn a nut a hundred times to try to thread it, or if you could have a robot system with a hand that has the ability to do the task with a tool, you could automate the process more. I think, you would naturally see a performance gain in this regard. So, the current tests to some extent, already address this issue.

Hyungpil Moon

Another question: for the peg-in-hole test, how can we tell how much is attributed to the hardware, the arm or the hand, or the algorithm in terms of the performance?

Karl Van Wyk:

We employed three different search algorithms for System 2 in the peg-in-hole test, but used the same hardware and control, otherwise. We conducted the test for all three and, in this way, you can see the performance change due to the algorithm itself. The issue that arises comes from changing too many aspects at once for a robot system, which will obscure how much any one component contributed to the overall performance change. These details are discussed in the paper.

Lionel Birglen:

The last year, we have been extensively using your tests for evaluating grasps, namely, the slip resistance one and grasp strength. For our hands, your other tests did not really make any practical sense, but those two very interesting for us. I do have numerous comments about them. For instance, you can get widely different numbers depending on how you use those tests. The slip resistance depends on the velocity of the linear stage, and also the stiffness of the spring. For grasp strength, we have huge differences, and we compared them, using a 6-axis load cell inside the split-cylinder. Actually, we ended up designing our own grasp artifact that we get consistently better results. Maybe we can discuss this in more detail later. We are using both tests a lot in the last year.

Joe Falco:

This is exactly the kind of input we are looking for from the community to improve the tests. We would be willing to work with you if you could share your experience and data on the matter.

Lionel Birglen:

Sure, we are looking to publish a few articles, but will mostly cover the design details of the hand, and not so much the fine tuning of the tests themselves. We would be happy to share all the data we have regarding the tests themselves.

Berk Calli

What are the locations and availability of the test stations?

Adam Norton:

I am from UMass Lowell, and I help direct a test facility, called the NERVE center. New England Robotics Validation Experimentation center. We replicate and design a lot of different test methods and benchmarks, working with organizations like NIST. We are currently building up our capabilities for testing robotic manipulation, and help iterate and proliferate existing test methods. Our facilities are not just for our own internal research, but for others as well. We are currently still working on the model for allowing outside use, but are hoping to get up and running soon. We are building up all the tests over the next few months, and are in the process of acquiring various robotic equipment and sensor systems.

Joe Falco:

We do want to reiterate that the designs for the tests are intentionally low-cost. For instance, you can replicate Task Board #1 for around $200 USD. The website contains documentation on assembly and where to acquire the parts.

Adam Norton:

We have recently replicated this task board, and it is simple to do. We are also not doing anything proprietary in terms up setting up a test facility, so are happy to share lessons learned, experience with any of those who are interested in setting up their own testing site.

Joe Falco:

This concludes the meeting. We will be sending out publicity material to WRS 2018. We will also distribute meeting minutes to everyone, and also to our website for those who could not make it. We will also distribute our internal documents, and appreciate an evaluation of what we have done so far. We will be putting another meeting together in the upcoming months, and are looking forward to future discussions.