Manipulation benchmark tasks for physical robotic systems: A proposed approach

A. C. Huamán Quispe

***Abstract*—One of the main goals in robotics research consists on creating intelligent systems with a high degree of autonomy, such that they can be used in home or industrial environments. While there has been important research done in the areas of perception, manipulation and control of robotic systems, little attention has been given to develop formal evaluation tools to benchmark the performance of these systems as a whole integrated unit. What capabilities are the minimum required for a system to be considered intelligent? How far are we still from a truly autonomous robot aide? Should requirements for home and industrial robots be the same? In this short paper, we propose a set of benchmark tasks to evaluate and compare the performance of manipulation policies applied in physical robotic systems.**

I. Introduction

One of the main objectives of robotics is the development of intelligent robot systems capable of a high degree of autonomy. Autonomy involves a system able to perform tasks safely on their own. For humans, it has been empirically shown that dexterity is a reliable predictor of independence [20]; that is, the more dexterous humans are, the more likely they can perform daily tasks on their own, without neither supervision nor extra help. Numerous dexterity tests are currently used in different fields, such as medicine, ergonomics, occupational therapy, among others. Tests that evaluate a person's dexterity help to select treatments, design training programs or determine the suitability of a person for a job that requires particular fine manipulation capabilities.

Going back to our discussion on robot autonomy, a few questions quickly arise:

- How autonomous our current robotic systems are?
- How far (or how close) are we from having a robot aide to be deployed at our homes to perform simple tasks?
- What should be considered a simple task?
- Given two robotic systems targeted to the same niche, how can their performances be compared?
- What are the most minimum tasks a robot should be able to perform?

All the questions above could be addressed, at least partially, if we had a consistent way to evaluate our robotic systems. During the last decades, we have seen noteworthy advances in diverse areas of robotics, such as perception, planning, control and manipulation. We consider that we have reached a state mature enough such that we can design a framework with which evaluate physical robot systems as a whole. Currently most systems are evaluated piece by piece (i.e. some perceptual systems might not consider the occlusion generated by a robot hand holding a tracked object, some grasping algorithms are tested assuming that accurate geometry information of the object is available).

Keywords: Robotics, Integrated Systems

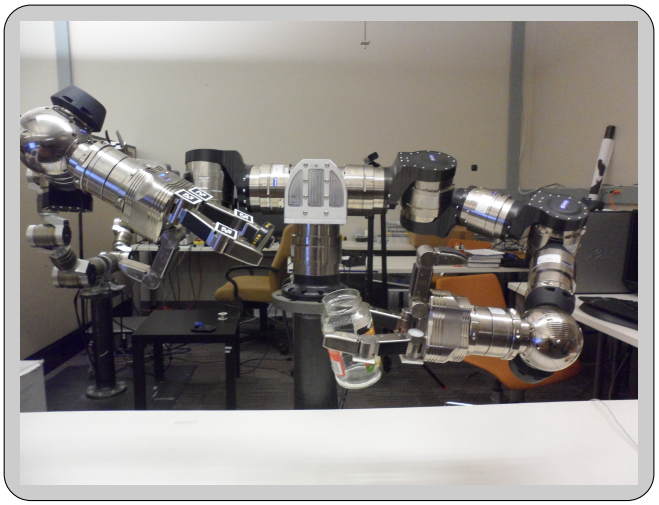

Fig. 1: One of the proposed bimanual manipulation benchmark tests: Opening a jar

Given that the fields of robotic applications is immense, trying to design a framework general enough for all possible robotic systems would be too ambitious and probably incomplete. Instead, we choose to focus on systems whose specific goal is to perform autonomously basic manipulation tasks, such as manipulators whose targeted niche is home environments. In order to design the framework to evaluate our systems, we revise the existing literature in dexterity tests targeted for humans and draw inspiration from them to select benchmark tasks that evaluate the following characteristics:

- **Physical fitness:** Actuated hardware (i.e. the arm(s) and the hand(s)).
- **Perceptual abilities:** Identify the object to manipulate (i.e. using visual or tactile perception).
- **Planning abilities:** How to create a policy to accomplish a specified goal task.
- **Awareness:** Is the robot capable of determining its success (or failure) after executing the task?

The rest of this paper is organized as follows. Section II summarizes existing previous work regarding evaluation methods for robot manipulators, most of them focused on evaluating physical capabilities such as grasp stability and robustness and general taxonomies to differentiate grasp types in order to assess a robot's capabilities to perform them. Section III gives an overview of the current state of dexterity

tests administered to human subjects and will elaborate on the metrics that are used to measure dexterity. Section IV show the tasks we propose as benchmarks, categorized according to the characteristic measured, the complexity of the task being evaluated and the robot physical capabilities. Finally, section V presents a brief discussion of our approach.

II. Previous Work

Dexterity is one of the most important human skills. As such, a goal in robotics manipulation is to create robotic systems that exhibit similar, if not superior manipulation abilities than these showed by humans. *Grasping*, that is, the ability to hold an object securely within a hand, has historically been studied extensively. Regarding work on benchmarking grasping tests, we should mention 2 important areas:

A. Grasp Taxonomies

Perhaps the oldest attempt to classify grasps was done by Schlesinger [17], who proposed to classify grasps according to the shape of the object being manipulated (i.e. cylindrical, circular, prismatic). Later on, Napier presented in [15] a basic taxonomy in which grasps are divided in power and precision grasps, the former with a high emphasis on the object being held robustly and fully constrained, whereas the latter was applied generally in tasks where in-hand manipulation were required (i.e. holding a pen to write). In [8] Cutkosky presented a grasp taxonomy inspired on observations of machinists at work and also based on both of the taxonomies previously mentioned.

After these initial proposed taxonomies, a myriad of new classifications have been proposed. Most notably in the area of robotics, Feix et al [10] proposed a set of 33 different grasp types based on the existing classifications works and recently Bullock et al. [6] proposed an extended version of the Feix's taxonomy to acknowledge non-prehensile grasps.

B. Grasp Metrics

In order to manipulate objects, a robotic system must generate a set of candidate grasps, select the one that is most likely to be robust and then execute it. Diverse metrics are currently used to assess the likelihood of a grasp to be successful. One of the most popular metrics still in use is the metric proposed by Ferrari and Canny in [11]. The ϵ metric measures the maximum disturbance wrench that a given grasp configuration can afford to resist.

While the Ferrari-Canny metric is vastly used, there have been some authors that noted that this metric is not sufficient to guarantee a stable grasp to be executed on the real world [9]. Other authors have further proposed empirically-based metrics, such as the volume of the object enclosed by the hand and the alignment of the wrist with the object's principal axes [1]. A more detailed review can be found in [18].

From the studies above, it is evident that existing metrics, while informative, need to be improved in order to be used to compare manipulation performance. One of the main drawbacks of existing metrics is that they isolate the evaluation to only the geometry and simplified contact dynamics between the hand and the object being grasped. As it has been noted previously [8], the selection of a grasp is influenced by a diversity of factors such as:

- Hand geometry
- Object geometry
- Task to be performed with the object at hand

This is depicted graphically in Figure 2. Most existing metrics consider the first two criteria (hand and object geometry) but not the task. Arguably, this is a logical procedure since there exists a myriad number of possible different tasks, hence a proper, general characterization is hard to achieve.

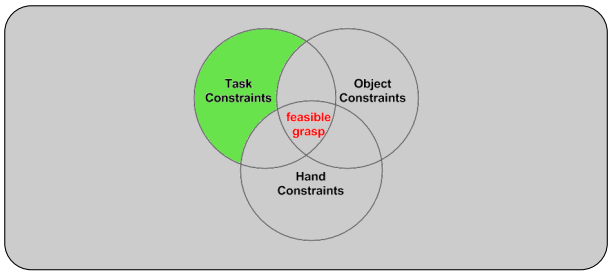

Fig. 2: Grasp selection based on diverse factors, including the nature of the task to be accomplished

III. Human Dexterity Evaluation

In this section we will present a brief overview of metrics used to evaluate human dexterity. Dexterity tests have a rich history and numerous alternatives exist, depending on the specific capabilities that need to be assessed. In particular, we are interested on dexterity evaluations that focus on measuring the capacity of human subjects to live independently, performing tasks with neither the need of external supervision nor additional help. Tasks of this nature are collectively known as *Activities of Daily Living* or ADL [3]

Borrowing terms from Occupational Therapy literature we start by defining dexterity as the *voluntary movement used to manipulate objects during a specific task*. This concept is a slightly modified version of the concept given in [23] in order to consider the following aspects also involved in manipulation:

- *Reasoning:* Voluntary movement implies planning actions in order to achieve the desired outcomes.
- *Object manipulation:* Depending on the task, the manipulation action can involve only the hand (i.e. rotating a pencil between the fingers) or the combined action of both arm and hand.
- *Task description:* Depending on the goal, tasks can involve the use of one or two hands to manipulate one or more objects.

Different tasks require different levels of dexterity. Two types of dexterity are widely recognized:

- **Motion dexterity:** Refers to the capability of handling objects with a hand (i.e. picking up a mug). It normally involves the interaction of arm and hand together.
- **Finger dexterity:** As described by Fleishman et al. [12] is the ability to make rapid, skillful and controlled manipulative movements of small objects, using primarily the

fingers (i.e. writing, rotating an object between fingers, slipping an object from the fingertips to the palm).

There are diverse dexterity tests that evaluate the two basic types of dexterity mentioned above (for more details, the interested reader is directed to [23]). Most of these tests consists on a series of tasks to be performed by the subject in the presence of a certified Occupational Therapist. The metrics used are generally based on the following 3 principles:

- Time required to complete the whole task.
- Number of sub-tasks accomplished within a predefined slot of time.
- Quality of movement during the task execution.

It is worth noticing that the last metric require more training in order to be properly measured (in comparison with the first two metrics, which only require a numeric calculation).

A few examples of dexterity tasks - either for motion or finger dexterity- are presented in tables I and II:

TABLE I: Sample Motion Dexterity Evaluation Tasks

Task	Metric	Test
Lift can	Time of completion	WMFT [22]
Pick up and transport blocks	Number of blocks transported	BBT [14]
Stack checkers	Time of completion	JHFT [13]
Grasp a light ball	Time of completion	SHAP [2]

TABLE II: Sample Finger Dexterity Evaluation Tasks

Task	Metric	Test
Flip cards	Time of completion	WMFT [22]
Do up five buttons	Time of completion	CHEDOKEE[2]
Insert pins in holes using tweezers	Number of pins	CSPDT [4]
Place and screw screws	Number of screwed objects	CSPDT [4]
Assemble a pin, washer and collar	Number of assembled structures	PPT [19]

IV. Proposed Tests

A. General Considerations

In the previous section we have reviewed dexterity tests applied to humans. While our final goal is to provide robots with human and even superhuman capabilities, we are aware that there are milestones to reach before such comparisons can be possible. A few observations are noted below:

- Motion and finger dexterity are two building blocks that enable a robot to perform basic manipulation, i.e. pick up an object and transporting it without any particular constraint.
- Autonomy requires a robot to perform tasks that require some level of reasoning. Hence, in order to evaluate the capabilities of a robot, the whole robotic system, that is, the integration of perception, planning and control, should be evaluated.
- From the human studies seen in section III we conclude that a metric should involve an observable measurement, such as time of completion or quality of movement. Metrics that involve calculations on simplified assumptions for the manipulation problem cannot always be trusted.
- A good share of everyday tasks involve the use of both limbs (left and right arm), so a complete evaluation should provide not only uni-manual but also bimanual scenarios to test dexterity.
- The eye-hand relation is highly important for the completion of a manipulation task. In the absence of visual perception, tactile perception is normally used, rendering a slower pace to accomplish the task..

B. Benchmark Test Types

Depending on the level of complexity of the system being evaluated, different tests are presented. We propose 5 broad types which will be explained in the following subsections. It should be noted that at this time, the tests proposed in this short paper are purposely vague in terms of implementation details (i.e. characteristics of the objects to be used for testing). Currently, we are working on the implementation of the proposed tests in two different physical robotic platforms (one of them shown in Fig. 1) in order to have experimental data on which standards for the testing objects can be defined.

TYPE 1: BASIC CAPABILITIES

The goal of this set of tests is to measure basic capabilities of the manipulator that do not necessarily involve an object, but nonetheless are crucial skills needed in order to perform the manipulation tasks on the following levels.

Task	Metric	Test
Put robot's forearm on a table	Time of completion	[21]
Put robot's hand on top of a table	Time of completion	[21]
Grip strength using dynamometer	Maximum force applied	[21]
Move arm across a table, brushing the surface	Time of completion	[21]

TYPE 2: MOTION DEXTERITY

In a similar manner as their human counterparts, these tests were selected to test basic arm/hand coordination to accomplish simple tasks upon which more complex goals can be built.

Task	Metric	Test
Pick up a light can	Time of completion	[22]
Pick up a pencil	Time of completion	[22]
Pick up a key	Time of completion	[22]
Stack checkers	Time of completion	[13]
Turn key in lock	Time of completion	[22]

TYPE 3: FINGER DEXTERITY

This set of tasks aim to evaluate the system's ability to perform manipulation mainly involving the fingers.

Task	Metric	Test
Put asymmetrical pegs on board	Time of completion	[5]
Assemble simple structures	Number of assemblies	[19]
Place pins using tweezers	Number of pins	[7]

Observe that all the tasks defined for Motion and Finger Dexterity are unimanual tasks that - other than the completion of the task itself- do not have any real-world effective goal.

The next two types of dexterity levels refer to *functional dexterity*, that is refers to the capability of a robotic system of successfully completing a given task by using their dexterity capabilities in conjunction with perception, planning and control.

TYPE 4: UNIMANUAL FUNCTIONAL DEXTERITY
These tasks are relatively more complex than the previous levels but are yet realizable with only one hand. The details of these tasks are explained in the original test citations.

Task	Metric	Original test
Spooning 5 kidney beans into a bowl	Time of completion	[13]
Pour a glass of water	Time of completion	[2]
Simulated feeding	Time of completion	[13]

TYPE 5: BIMANUAL FUNCTIONAL DEXTERITY

In the Occupational Therapy literature, it has been noted that most of the ADL activities we perform daily involve the use of both of our extremities, hence we consider it highly important to address manipulation capabilities for tasks that require both hands to operate. The following table show some proposed tasks:

Task	Metric	Original test
Opening a screw-topped jar	Completion time	[16]
Sharpening a pencil	Completion time	[16]
Fastening a snap	Completion time	[16]
Taking a cap off a bottle	Completion time	[16]
Squeezing toothpaste on a toothbrush	Completion time	[16]

V. Discussion

In this paper we have presented a set of suggested benchmark tasks for physical robotic systems targeted for manipulation in home environments. We are strongly convinced that in order to improve the capabilities our robots have, a set of metrics are needed, which will enable researchers to be able to compare their results in a more straightforward manner.

The second reason why we believe a set of benchmark tests is needed is because we consider that, at this point in time, we have available systems that must be tested as a whole, with metrics that evaluate the complete performance of all the sub-parts together. Perception, planning and control must be evaluated together, the same way as humans are. If we wish to provide our robots with human capabilities, our metrics should be at least as rigorous as the ones used for our human counterparts.

As we mentioned in the first sections of this paper, we don't intend to present the ultimate set of benchmark tests or to define an specific type of policy to solve the manipulation problem. Our main intention is to start the discussion and encourage fellow roboticists to contribute with their own points of views regarding benchmarking and metrics for physical systems. Only with collaboration and with an open dialogue between researchers in the robotics community we believe that we can keep improving the manipulation capabilities of our current systems.

References

[1] Ravi Balasubramanian, Ling Xu, Peter D Brook, Joshua R Smith, and Yoky Matsuoka. Physical human interactive guidance: Identifying grasping principles from human-planned grasps. In *The Human Hand as an Inspiration for Robot Hand Development*, pages 477–500. Springer, 2014.

[2] Susan Barreca, Carolyn Kelly Gowland, Paul Stratford, Maria Huijbregts, Jeremy Griffiths, Wendy Torresin, Magen Dunkley, Patricia Miller, and Lisa Masters. Development of the chedoke arm and hand activity inventory: theoretical constructs, item generation, and selection. *Top Stroke Rehabil*, 11(4):31–42, 2004.

[3] Claude Bouchard, Roy J Shephard, and Thomas Stephens. Physical activity, fitness, and health. In *Health promotion and physical activity: joint meeting, Cologne*, pages 37–49, 1994.

[4] AM Boyle and JC Santelli. Assessing psychomotor skills: the role of the crawford small parts dexterity test as a screening instrument. *Journal of dental education*, 50(3):176–179, 1986.

[5] PJ Bryden and EA Roy. A new method of administering the grooved pegboard test: performance as a function of handedness and sex. *Brain and Cognition*, 58(3):258–268, 2005.

[6] Ian M Bullock, Thomas Feix, and Aaron M Dollar. Finding small, versatile sets of human grasps to span common objects. In *2013 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1060–1067. IEEE, 2013.

[7] EN Corlett, G Salvendy, and WD Seymour. Selecting operators for fine manual tasks: A study of the o'connor finger dexterity test and the purdue pegboard. *Occupational Psychology*, 1971.

[8] Mark R Cutkosky. On grasp choice, grasp models, and the design of hands for manufacturing tasks. *Robotics and Automation, IEEE Transactions on*, 5(3):269–279, 1989.

[9] Rosen Diankov. *Automated construction of robotic manipulation programs*. PhD thesis, Citeseer, 2010.

[10] Thomas Feix, Roland Pawlik, Heinz-Bodo Schmiedmayer, Javier Romero, and Danica Kragic. A comprehensive grasp taxonomy. In *Robotics, Science and Systems: Workshop on Understanding the Human Hand for Advancing Robotic Manipulation*, pages 2–3, 2009.

[11] Carlo Ferrari and John Canny. Planning optimal grasps. In *Robotics and Automation, 1992. Proceedings., 1992 IEEE International Conference on*, pages 2290–2295. IEEE, 1992.

[12] Edwin A Fleishman and Gaylord D Ellison. A factor analysis of fine manipulative tests. *Journal of Applied Psychology*, 46(2):96, 1962.

[13] Robert H Jebsen, NEAL Taylor, RB Trieschmann, MJ Trotter, and LA Howard. An objective and standardized test of hand function. *Archives of physical medicine and rehabilitation*, 50(6):311, 1969.

[14] Virgil Mathiowetz, Gloria Volland, Nancy Kashman, and Karen Weber. Adult norms for the box and block test of manual dexterity. *Am J Occup Ther*, 39(6):386–391, 1985.

[15] John R Napier. The prehensile movements of the human hand. *Journal of bone and joint surgery*, 38(4):902–913, 1956.

[16] Massimo Penta, Jean-Louis Thonnard, and Luigi Tesio. Abilhand: a rasch-built measure of manual ability. *Archives of physical medicine and rehabilitation*, 79(9):1038–1042, 1998.

[17] Ing G Schlesinger. Der mechanische aufbau der künstlichen glieder. In *Ersatzglieder und Arbeitshilfen*, pages 321–661. Springer, 1919.

[18] Raúl Suárez, Jordi Cornella, and Máximo Roa Garzón. *Grasp quality measures*. Institut d'Organització i Control de Sistemes Industrials, 2006.

[19] Joseph Tiffin and ESTON J Asher. The purdue pegboard: Norms and studies of reliability and validity. *Journal of applied psychology*, 32(3):234, 1948.

[20] Mark E Williams, Nortin M Hadler, and Jo Anne L Earp. Manual ability as a marker of dependency in geriatric women. *Journal of chronic diseases*, 35(2):115–122, 1982.

[21] Steven L Wolf, Deborah E Lecraw, Lisa A Barton, and Brigitte B Jann. Forced use of hemiplegic upper extremities to reverse the effect of learned nonuse among chronic stroke and head-injured patients. *Experimental neurology*, 104(2):125–132, 1989.

[22] Steven L Wolf, Paul A Thompson, David M Morris, Dorian K Rose, Carolee J Winstein, Edward Taub, Carol Giuliani, and Sonya L Pearson. The excite trial: attributes of the wolf motor function test in patients with subacute stroke. *Neurorehabilitation and Neural Repair*, 19(3):194–205, 2005.

[23] Katie E Yancosek and Dana Howell. A narrative review of dexterity assessments. *Journal of Hand Therapy*, 22(3):258–270, 2009.