A Taxonomy of Benchmark Tasks for Bimanual Manipulators

Author Names Omitted for Anonymous Review. Paper-ID 52

***Abstract*—This paper presents a taxonomy of benchmark manipulation tasks for bimanual service robots. Our contributions are threefold: (1) A review of relevant literature regarding manipulation tests in the robotics domain and related fields, such as physical therapy, assistive technologies and prosthetics. (2) Guidelines to design useful testing protocols to evaluate manipulation performance. (3) A proposed general taxonomy of benchmark manipulation tasks and sample tests per each class. We also present a discussion to highlight the importance of building standard norms to evaluate integral robotic systems by using metrics and scenarios that are relevant, simple and objective.**

I. Introduction

In 2013, the field of robotics turned 50. Through these decades, great advances have been made in diverse subfields of robotics, including robot perception, learning, and manipulation. Research on grasping and manipulation, in particular, has lead to impressive results such as robots that can use chopsticks [32], robots that can handle deformable objects [44], and robots that can use a variety of power tools [17]. While these and other results highlight remarkable achievements of individual research groups, it still remains an open question how much the field progressed as a whole. How much closer are we to advanced service robots that are capable of complex manipulation skills in real-world environments?

A common theme among recent workshops and symposia on grasping and manipulation, is that there exists an "increasing difficulty in comparing the practical applicability of all new algorithms and hardware" [3]. Among the reasons that have led to this impasse are (a) the lack of standardized benchmark tasks, (b) a lack of suitable metrics to evaluate the integral performance of a robot, (c) the lack of comparative data that allows us to broadly categorize robot performance, and (d) the lack of guidelines for a realistic experimental setup. Hence, the methodology used to perform and evaluate robot manipulation experiments greatly varies in different publications, often rendering a comparison infeasible. Under these circumstances it is difficult to document and evaluate the progress made in the development of robot manipulation skills over the years. Although during the last years, important efforts have been made to evaluate robotic performance (i.e. through robotic competitions [26, 57]), there is still progress to be made.

The challenge of evaluating motor skills is not unique to robotics. Medical professionals have faced similar challenges, and have as a result studied systematic approaches for assessing physical capacities of patients. Using standardized tests and measures, e.g., dexterity tests and functional capacity tests, physicians estimate and broadly categorize the physical capabilities of patients. These assessments allow the physician to investigate the existence and degree of impairments caused by an injury, accident or illness. It allows informed decisions about treatments, training programs, or the suitability of a person for a specific job that requires particularly fine manipulation skills. In addition, through the repeated application of these tests, a physician can monitor the rehabilitation progress of a patient and compare physical capacities among different individuals.

Fig. 1: How to meaningfully compare 4 different robots performing the same manipulation task?

In this paper, we lay groundwork to formally establish a taxonomy of benchmark tasks for bimanual manipulators. We review literature concerning dexterity and functional capacity tests developed for medical purposes and extract guidelines to define clear methodologies to evaluate manipulation tasks. We also analyze the existing efforts towards benchmarking in the robotics literature and determine the commonalities, advantages and shortcomings among them. Our goal is to discuss both the possibility of standardizing robot manipulation benchmarks, as well as the difficulties that are hampering these efforts. We (1) propose a taxonomy of benchmark manipulation tasks and its corresponding metrics, (2) justify the need for evaluation frameworks that consider the robot as a whole and not as the sum of its parts, (3) discuss the strong need of realistic assumptions for robot experiments.

Our work focuses on service robots, although some of the sections can be considered useful for industrial robot

evaluation as well. The rest of this paper is structured as follows: Section II presents a review of existing work in manipulation tests in both robotics and related fields. In this section we also introduce definitions to be used through the paper. Section III presents an in-depth discussion regarding the criteria needed to define realistic benchmark tasks. Section IV presents the benchmark taxonomy proposed and examples of representative tasks and a brief discussion of each type. In this section we also discuss the guidelines proposed for realistic testing. Finally, section V presents our insights and concluding remarks.

II. Revisiting Manipulation Tests

A. Background and term definition

Arguably, there is a deep relationship between human intelligence and dexterous manipulation. In [56], Williams presented experimental evidence showing that manual ability is highly correlated with the level of independence in elderly population. In this context, manual ability refers to the ability to perform actions involving object manipulation in order to achieve a goal. We will refer to this concept as hand function.

Definition 1 (Hand Function)

The ability to use the hand(s) in everyday activities, which involves dexterity, manipulation skills and task performance skills [36].

Hand function then refers to using the hand(s) to accomplish a useful task, such as writing, cutting meat, pouring a drink or opening a door. All the tasks mentioned require dexterity, defined loosely here as the ability to manipulate an object with the hand. Dexterity can further be described by two related terms:

Definition 2 (Manual Dexterity)

The ability to make skillful, well directed arm-hand movements in manipulating fairly large objects under speed conditions.

Definition 3 (Fine-Motor Dexterity)

The ability to make rapid, skillful, controlled movements of small objects where the fingers are primarily involved.

As it was mentioned in definition 1, hand function requires no one but a combination of multiple abilities such as motor, perceptual and cognitive skills, which constantly interact with each other.

Definition 4 (Manipulation Skills)

Motor Skills: Ability to actuate in the environment.
Perceptual Skills: Ability to gather information from the environment.
Cognitive Skills: Capacity to elaborate plans to achieve a goal, taking into account both motor and perceptual knowledge.

There has been acute interest in developing tests to evaluate the manipulation capabilities of both humans and robots. Regarding humans, testing is important for a variety of reasons, some of them highlighted in Table I. On the robotics field, evaluation is important in order to compare quantitatively different research approaches.

Designing a unique test to evaluate hand function for robots (and for humans) is hard. As it was explained, manipulating an object involves the interaction of diverse, well-defined capabilities. Should these components be evaluated per separate? What metrics should be used? How should the tests be selected?

In the following section we analyze the most relevant tests we found in literature related to hand evaluation from a purely human perspective. Afterwards, we review robotics literature regarding benchmarking and common manipulation tasks performed by physical robots. We finish this section summarizing the similarities, differences and lessons learned from both types of tests.

B. Manipulation Tests for Humans

There is not an universal definitive test for hand function, hence diverse manipulation tests have appeared during the last 200 years under different names, such as hand function tests, dexterity tests and motor assessment evaluations. We reviewed 17 representative tests (Table II). The procedure to select these tests was the following:

- Tests appearing in recent Physical Therapy surveys [36, 50] and whose original sources were available in English.
- Tests referenced by the ones above.
- Tests found through Google Scholar using the following search keywords: *manipulation, dexterity test, hand* and *function*.

Due to space constraints, we only give a general overview of some of the tests on Table II. A more detailed description of all our reviewed tests can be found in the supplementary material.

1) Manipulation Tests through the years: Pioneering studies on hand function focused on motor skills with emphasis on hand strength and joint mobility. Regarding the former, Mathiowetz et al. [42] analyzed 4 measures of strength: grip, palmar pinch, key pinch and tip pinch. Regarding the latter, in [29] the Kapandji Thumb Opposition scale is presented, which measures the range of motion of the hand by evaluating if the thumb is able to reach the other 4 fingers at their fingertips and at their joints.

TABLE I: Examples of subject types using the tests presented

Target population	What is being evaluated?
1. Young children	Normal development
2. Stroke patients	Gradual recovery of motor skills
3. Elderly population	Capability of living independently
4. Workers	Fine-dexterity skills necessary for assembly jobs

The justification to use motor-skill tests was the high correlation between these skills and hand function. However, it was noted that these tests do not involve interaction with objects, which is a fundamental part of manipulation. This fact inspired to the development of tests involving simple manipulation actions, such as pick-and-place operations. In [41] Mathiowetz proposed the Block and Box Test (BBT) to measure manual dexterity. The test consists of picking up wooden blocks from a bin and throwing them into another bin as fast as possible. Another test, used for selective evaluation of assembly workers, is the Purdue Pegboard Test (PPT) [40], which requires the use of a board with rows of holes. The task consists on placing thin pegs into the holes with the left arm, then the right arm and then both arms at the same time. The test also included a final bimanual assembly task, where after placing the peg, 3 small objects are placed on top of it (two washers and a collar). In this case, the test emphasizes fine-motor dexterity.

The tests mentioned above, and many others [4, 23], involve the manipulation of pegs or cubes. Given the extensive variety of objects to be potentially grasped, additional tests were proposed to evaluate the ability of the hand to perform different grasps for objects with different geometry (cylindrical, spherical, three-jaw). An example of these is the ARAT test [59], composed of 4 sections, 3 of them related to grasping, gripping and pinching objects such as a wood cube, a cricket ball, a small metallic tube and a marble.

So far, the test reviewed are informative. However, it has been questioned if they truly measure the ability of the subjects to use their hand functionally. Take as an example stroke patients. It has been observed that, although many of them performed poorly in fine-dexterity tests, they nonetheless are capable of performing ordinary everyday tasks. The challenges in motor skills are faced with a better use of their cognitive skills. Given this, tests featuring functional tasks were introduced. The Jebsen Hand Function Test (JHFT) [28] consists on 7 tasks chosen as representative of common activities of daily living such as picking up cans, scooping beans with a spoon (simulated feeding) and stacking checkers. All of these tasks are unimanual. However, most manipulation tasks involve the use of two arms. Successive tests took care of including both type of tasks. In [53], the Sollerman test was proposed, consists of 20 tasks from which 14 are bimanual. An interesting aspect of this test is that the tasks were chosen such that the grasps used were proportional to their frequency. Recently, the SHAP test [33], inspired heavily by the Sollerman test, has been presented. This test is particularly interesting because it present tasks for both dexterity and functional evaluation.

TABLE II: Reviewed manipulation tests

Test	Aspect measured
1. Kapanji Test [29]	Thumb Opposition
2. Hand Strength [42]	Grip and pinch strength
3. BBT [41]	Manual dexterity
4. MPT [47]	Finger dexterity Sensory assessment
5. NHPT [48]	Finger dexterity
6. FPT [4]	Finger dexterity
7. PPT [40]	Finger dexterity
8. MRMT [23]	Manual dexterity
9. GPT [9]	Visuomotor aptitude Motor speed
10. JHFT [28]	Hand function
11. SODA [55]	Hand function
12. WMFT [58]	Manual dexterity Hand function
12. CAHAI [6]	Hand function
13. ARAT [59]	Arm dexterity
14. Sollerman [53]	Hand function
15. TEMPA [16]	Hand function
16. SHAP [35]	Hand function
17. AMAT [31]	Hand function

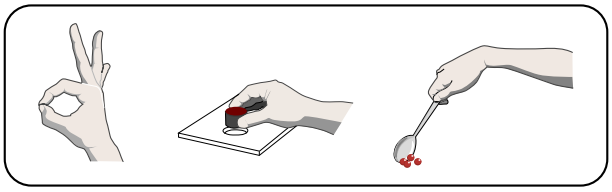

Fig. 2: Different samples of manipulation tests items used in occupational therapy: Left: Kapandji test. Middle: Peg-and-Board Generic Test. Right: Jebsen Hand Function Test.

2) Common aspects in human tests: While the tests presented are diverse, they share common characteristics, which we identify as guidelines that can be potentially adapted to design benchmark tasks for robots:

- **Cognitive simplicity:** Most tests have not an specific target population, hence they should be adequate for children, adults and elderly population.
- **Time-efficiency:** The test should be administered in a short time frame.
- **Relevance:** The tests should assess crucial manipulation skills.
- **Affordability:** Ideally, the objects used in the tests should be easy to obtain or easily built. Specifications should be made available.
- **Normalization:** Studies in healthy population must be performed such that comparisons and evaluations can be made.

C. Manipulation Tests for Robots

In this section we review manipulation tests from a robotics perspective. Similar to the case of human tests, there is not an evaluation tool unanimously embraced by the community. As a result, the research literature shows a myriad of robotic systems capable of executing heterogeneous tasks under varying assumptions. During the last years, though, there has been a growing interest in designing common tasks to compare the performance of different robot systems.

As pointed by [25], most of the existing tests for robot manipulators can be classified in two types:

- **Component Benchmarking:** Tests that evaluate a single component of the robot system.
- **System Benchmarking:** Tests that evaluate the robot system as a whole, considering the interactions between its components.

In this section we only address tests of the second type.

In 1985, Collins proposed a test for industrial manipulators, commonly known as the Cranfield Assembly Benchmark [13], which consists on assembling 17 parts to form a mechanical pendulum. As it was noted in the paper, the goal of this test was *to examine the software and control features applied in the assembly process*. Given the structured nature of the assembly environment, perception was not considered as an element to be evaluated.

Home environments pose challenges far different from industrial settings: Each house is different, there exists many different household tasks, and each of them can potentially require a different set of skills from the robot. Therefore, it should not come as a surprise that designing a set of benchmark tasks is considered a daunting task. Nonetheless, through the years researchers have used common sense to choose tasks to evaluate their systems, such as window cleaning, fetching and pouring drinks and folding laundry [44], to name a few. Table III shows a sample of tasks currently performed by state-of-the-art bimanual manipulators, most of them cited in a recent survey [52]. Table IV presents the assumptions considered regarding world knowledge.

TABLE III: Common tasks presented in robots

Task	Robotic platform
Making coffee	Rollin Justin [34]
Capping a pen	Captain Crichton [14]
Picking up a human	RIBA [46]
Insert a plug into a socket	PR2 [43]

TABLE IV: Assumptions

Task	Assumption
Making coffee	3D model of objects is known
Capping a pen	3D model of object is known
Picking up a human	Constant human guidance
Insert a plug into a socket	Fiducial marker in plug

With the advent of both affordable robotic manipulators and RGB-D sensors, the interest in benchmark tasks to compare robot performance was spurred. In 2008, Grunwald et al. [25] presented a set of benchmark tasks for dual-hand manipulators, as part of the DEXMART project [51]. The tests were targeted to bimanual manipulators and were chosen considering a cafeteria environment, so tasks such as carrying a tray with two hands and clearing a table were proposed.

In order to foster research advancement, robotic competitions featuring manipulation tasks have notably flourished in the last decade. In 2010, DARPA organized the Autonomous Robotic Manipulation (ARM) program, a 4-year project aimed to develop software and hardware to enable a robot to autonomously manipulate, grasp and perform complicated tasks with humans providing only high-level supervision. More recently, the Amazon Picking Challenge (APC) [2] has been proposed and aims to evaluate bin-picking skills in an industrial environment. Multiple other competitions exists that involve manipulation as one of the evaluated aspects. On 2013, the DARPA Robotics Challenge (DRC) took place in Miami and 4 out of its 8 tasks directly involved manipulation. A more traditional competition, the Robocup@Home [57], has been running since 2005. Its main goal is to evaluate the progress of mobile manipulators in a home environment. Navigation, manipulation and human-robot interaction were among other features evaluated. Table V shows the details of the tasks involved in the competitions mentioned above.

1) Common aspects in robotic tests: The following are observations based on the literature mentioned above.

- **World-knowledge assumptions:** Different robots operate under different assumptions regarding available knowledge.
- **Isolated evaluations:** Since there is not a clear methodology towards manipulation benchmarking, each robot is evaluated using different metrics.
- **Lack of normative data:** Absence of a golden-standard to compare against. Given this, there is not a clear sense of how well a manipulator performs a task (in comparison with the state-of-the-art).
- **Risk of cherry-picking tasks:** As it was pointed by Choi et al. [11], many tasks (and the objects involved) can be selected with an unconscious bias towards the ones that are more suited to the particular robot being evaluated.

III. Guidelines to Design Manipulation Testing Methodologies

In order to define benchmark tasks, we must first define the methodology to follow. Efforts to establish standard practices for benchmarking manipulation have been shown by different authors [1, 7, 27, 37, 60]; however no clear answers have been delineated. The closest attempt, in our view, to ground clear protocols for manipulation benchmarks is the work presented in [27], in which the author suggests that benchmarking tests should be defined by two aspects: *Test Description*, which specifies the conditions under which the tests will be performed; and *Test Evaluation*, which describes how the task performance will be measured. Both of these aspects will be discussed in III-A and III-B.

A. Test Evaluation

In order to compare reported results in robot manipulation, well-defined evaluation metrics are needed. Evaluation metrics can be of qualitative or quantitative nature and can address different sub-topics of manipulation. Robot hands and arms are typically first assessed through their physical characteristics and low-level performance indicators, e.g., degrees-of-freedom, applicable forces, number of fingers, etc. However,

TABLE V: Manipulation tasks in recent robot competitions

Event	Task
DARPA ARM I	1. Grasp 12 objects: radio, rock, ball, flashlight, hammer, case, floodlight, shovel, screwdriver 2. Staple a stack of paper 3. Turn on flashlight 4. Open door (handle) 5. Unlock a door (key) 6. Drilling 7. Hang up a phone
DARPA ARM II	1. Change a tire 2. Open a bag, extract pliers and cut a wire
DRC	1. Carry and connect fire hose 2. Open series of doors 3. Drive and exit utility vehicle 4. Locate and close leaking valves
RoboCup@Home	1. Hand an object to a human 2. Grasp a cup and a bottle 3. Pick up objects from shelf 4. Pick up 5 unknown tabletop objects
DEXMART	1. Empty a trash bin 2. Open a screw cap 3. Solve a Rubick Cube 4. Pour water into a glass 4. Carry a box with two hands 5. Insert a battery into a drill
APC	1. Bin-picking a single object 2. Bin-picking a single object in light clutter

these performance indicators can not be used to infer better or worse manipulation capabilities, since this depends on the design of the hand, as well as the interplay of the components when put to test. In [24] and [15] *taxonomies* are used to evaluate the range of executable grasps. Given a grasp taxonomy and a specific robot hand, the set of executable grasps can be identified and compared to other robots. To assess the range of grasps in a more quantitative manner, Feix et al. [21] introduced a so-called anthropomorphism index which compares reachable fingertip poses between a human hand and a robot hand. An *anthropomorphism index* is a continuous variable and can therefore be used to assess even small changes in the hand. Yet, it only compares feasible hand shapes and does not include contacts with objects. In contrast to that, the *graspability map* introduced in [49] explicitly analyses the contacts between a hand an object. A graspability map represents the set of poses that might lead to a precision force closure grasp on an object. Comparing the size and quality of graspability maps for different robot hands can indicate which hand is more likely to produce stable precision grasps.

The above performance metrics mainly target the mechanical design and hardware properties of a robot manipulator. Grasp planning algorithms, in contrast, are compared using a different set of metrics. A common approach is to evaluate the grasp quality using specific grasp quality metrics [8], such as the ϵ-metric, which measures the total and maximum finger force [22]. Yet, as was noticed in recent publications [5], grasp quality measures are typically calculated in simulations and often do not reflect the grasp executed by a robot. Hence, the most prominent approach for evaluations is to analyse the success rate in an object lifting task. Typically, a set of representative objects is grasped, then lifted, and the number of successful trials is documented. While such lifting tests are more informative, they still do not provide complete information about the degree of stability of the grasp. In [30] Kim et al. suggested using both visual inspection, as well as interactive, physical inspection to evaluate the success of grasps. In interactive inspection, a human subject touched the object while it was grabbed by the robot, applied physical perturbations (jiggling it), and then evaluated its stability. In a similar vein, Morales et al. [45] used shaking movements after grasping in order to estimate the stability of a grasp.

While grasp stability is *necessary* for many tasks, it is not a *sufficient* condition for manipulation tasks. Many tasks that go beyond pick-and-place require additional dexterous capabilities. Benchmarking of manipulation skills has therefore moved towards *functional tests*, such as opening a series of doors, removing a screw cap, or inserting a battery into a drill. Functional tests can evaluated by assessing whether the complete task was successfully achieved and by counting the success of individual sub-tasks. Table V contains a list of functional tasks that have recently been used in major competitions and projects on grasping and manipulation. The simple interpretation of achieved results and the embedding in real-world scenarios, makes functional tests particularly appealing for robot competitions. In addition, functional tests do not come with an inherent assumption about the robot hardware. In contrast to the Kapanji test, grasp quality measures or taxonomic evaluations of dexterity, functional tests do not assume a specific morphology and can be performed with a wide range of different manipulators, e.g. jamming grippers, deformable hands, or anthropomorphic hands.

B. Test Description

Comparison of robot experiments can only be reasonably performed, if the involved tests are conducted under the same or sufficiently similar conditions. In grasping and manipulation, reported experimental results can be based on a wide variety of assumptions. Assumptions are often made with respect to the involved perception system, the amount of prior knowledge about the object, physical properties of the object, robot hardware, lighting conditions, the inherent uncertainty of the system, deformable vs. rigid objects, the location of objects, the used software system, the update frequency of the system, or the degree of variation in the scene during the task. As a result, it is often challenging to compare achieved results with reported values in the literature, or even replicate a result reported in a different paper. To overcome this challenge, we can design benchmarks in such a way as to minimize the variance in the inherent assumptions.

One approach to do this, is by clearly specifying *reasonable* assumptions. For example, as done in recent competitions, e.g. the Amazon Picking Challenge, specific benchmark objects along with their 3D models can be provided. Given a restricted set of objects to manipulate, it is also possible to provide a shared software framework for object detection. This would

reduce the effect of perception on the manipulation benchmark. However, it also bears the risk of *over-specialization*; the design of very specific, brittle solutions that do not generalize to new situations.

Another approach to minimize variability is to use a standard robot hardware and software platform. For example, various publication on manipulation use the PR2 robot in conjunction with ROS for manipulation. This allows different research teams to share algorithms at the code level, thereby facilitating benchmarking different manipulation methods. However, standardizing the robot platform also limits the range of possible research questions that can be addressed. In the case of the PR2, for example, no in-hand manipulation can be studied.

In this paper, we propose two different measures to ensure a fair comparison of grasping and manipulation methods. The first measure we propose is the inclusion of **inherent stochasticity** into grasping benchmarks. Instead of specifying a set of conditions under which an experiment is performed, e.g. a set of locations and orientations of the object at the start of the experiment, we can design our benchmarks such that the conditions are determined by the stochasticity in the task. In the case of grasping and lifting an object, for example, we can introduce stochasticity by repeatedly dropping the object into a bin and then performing the task again. The configuration of the object in consecutive trials will depend on the resting position after dropping. As a result the variation in pose will be imposed by the task rather than a human expert. The second measure we propose for fair comparisons of manipulation capabilities, is the focus on longterm **multi-step** evaluations. Instead of executing a task only one time after which the human tester resets the environment, we run the experiment in a loop without human intervention. In the case of the above lifting example, we can have the robot repeatedly lift an object from a bin and throw it into a second bin in alternation. Since no human intervention is allowed, these tests capture the robot's ability to repeatedly deal with stochasticity inherent to these tasks.

IV. Taxonomy of Benchmark Tasks

Section II and III provided an overview of the current status of benchmarking as well as insights to specify relevant test descriptions and test evaluations. In this section we present our proposed taxonomy of benchmark tasks for robot manipulators. Several criteria was considered in order to design the high-level taxonomy and the sample tasks that will be presented. In the following lines we will summarize these considerations, which we consider vital for a proper methodology of benchmark design:

1) **Hardware-agnostic:** Tests should not be designed with a specific platform in mind. In order for benchmarks to be widely accepted, they should be realizable under minimum hardware capabilities assumptions.
2) **Flexibility:** A lesson learned from the review of human tests is that there is not an unique "right way" to solve a task. The robot should be allowed to apply a strategy that better suits its particular situation (i.e. a robot equipped with a gripper might have to adopt a different approach to grasping an object than a robot with a 3-fingered hand).
3) **Time efficiency:** Benchmarking is not a goal by itself. Rather, it should be seen as a diagnostic tool to be applied regularly, to make sure our systems are comparable (or within reasonable distance) to the state-of-the-art. Accordingly, since benchmarks are a sidestep tool, they should be selected such that they can be evaluated in a rapid manner, without the need of a complex setup or highly constrained rules.
4) **Relevant metrics:** A metric is only useful if it can be compared against a standard value. Tasks should be selected such that the evaluation can be objective, numerically expressed and important for both the researcher and, eventually, for the end-user. From the discussion in III-A we believe that (1) Task completion time and (2) Success rate are the two objective, informative metrics that can more easily be used.
5) **Statistically relevant:** Evaluation should consider a minimum number of attempts to be considered valid. This has already been seen in the ARM project evaluations, in which 5 trials were performed and the average results were evaluated.
6) **Realistic assumptions:** While a Laboratory environment is not the same as a real-home space, we should stress the importance of avoiding to rely on assumptions that in no way will exist in the real world. For instance, assumption of markers placed in objects or off-board visual sensors are very unlikely to occur, so their use might not translate into a realistic evaluation of capabilities.
7) **User-focused tests:** Service robots will be deployed at human homes, so it is reasonable to take into account feedback from the end-users to evaluate our systems. Studies in assistive technologies have shown that the main objective performance measures used by humans with assistive robotic arms are: Their capacity to perform activities of daily living and time to task completion [54]

Based on the reasons exposed, we designed the taxonomy of benchmark tests shown in Figure 3. Details of each level in the classification follows:

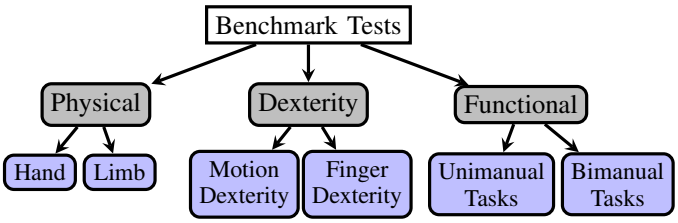

Fig. 3: Taxonomy of Benchmark Tasks

A. Physical Tests

The tests in this section measure fundamental motor skills in a robot system, requiring very limited or null perceptual and cognitive skills. The purpose of this battery of tests is to evaluate the potential of the robot's hardware, which is related to its ability to solve a task.

We further divide these tests in two types explained below. Examples of representative tasks are shown in Table VI .

1) Hand Tests: Tests that measure exclusively the hand physical capabilities. Given that a hand itself is often a complex system on its own, this decision seems reasonable.

An important observation for this section is that these tests effectively evaluate the hand potential, rather than the hand's actual ability to manipulate, which requires the additional evaluation of both perception and cognitive skills jointly with the motor skills.

An example of a hand test commonly found in robotics literature is the use of grasp taxonomies for hand evaluation [15].

2) Limb Tests: Tests that measure capabilities of both hand and arm interacting together. Evaluation of simple online control is tested, such as capability to follow a workspace trajectory and capability to follow an end-effector pose. These abilities can be considered basic building blocks for the tests in the following sections.

TABLE VI: Sample Physical Tasks

Sublevel	Examples
Hand	1. Maximum finger aperture 2. Maximum payload when picking up a high-friction object 3. Perform static grasps from existing taxonomies. 4. Perform static grasps of benchmark objects [11, 39]
Limb	1. Position the palm on a table surface 2. Follow a 2D curve on the table (i.e. circle) 3. Point at diverse objects on a table

B. Dexterity Tests

This type of tests evaluate motor skills, moderate perceptual skills and basic cognitive skills to solve simple tasks involving:

- Pick-up an object from a table
- Pick-and-place and object in an uncluttered planar surface
- Object in-hand reconfiguration
- Hand-eye coordination.
- Collision avoidance as principal constraint.
- Manipulation of objects of simple geometry.

The further subdivision is explained below:

1) Manual Dexterity Test: This type of tests measure the robot's ability to manipulate objects mainly for transport operations, in which in-hand manipulation is not required. In [18], Feix et al. found out - after analyzing video data of daily activities of 2 household workers and 2 machinists - that for many cases, power grasps were more vastly used instead of precision grasps, even for small objects. Hence, the geometry of the object play less a factor in the grasp selection when the task is simply pick and place. In other words, for this type of test tasks, an accurate 3D model should not be required.

Another interesting result from [19] is the fact that most of the objects (used by the subjects in the video) presented characteristics that made them easy to grasp: Object mass was normally less than 500 grams and the grasps required less than 7 cm in width. This information is corroborated by previous studies reporting the most common characteristics of manipulated objects [11, 39]

2) Fine-Motor Dexterity Test: In contrast to the manual tests, the fine-motor dexterity tests evaluate in-hand manipulation: The ability to modify an object configuration without using arm's movement. Fine-motor skils are challenging since they usually require more sophisticated sensing capabilities, such as tactile. Also, perceptual errors, which can be more or less tolerated and corrected when manipulating regular objects, can affect more adversely in this case.

The following table shows some sample tasks. In the manual examples, *object* is a generic term referring to objects of generic geometry (cylinder, cube, sphere) and weight no bigger than 500 g.

TABLE VII: Sample Dexterity Tests

Sublevel	Examples
Manual	Pick-and-place an object on a clear table Pick an object from a table and place it on a cupboard Pick an object from a box and set it at an adjacent box
Finger	Unscrew a bottle using fingers only Rotate a chopstick Grab a short cylinder from the table and rotate it such that its supporting face ends up facing upwards.

C. Functional Tests

The tests in this section require the full interaction of motor, perceptual and cognitive skills in order to solve the task at hand, which presents the following characteristics:

- Functional tasks usually require more than one step to be accomplished, hence task planning at the cognitive level should be addressed.
- Task-specific constraints must be considered. Of particular importance are object's pose constraints.
- When faced with a household task, humans possess knowledge from previous experience. In particular, object information is usually available at some abstract level. Given this, and following the inspiration of [26], it is acceptable to assume that the robot have previous available knowledge regarding the object and how it can be used to solve the task.
- In a similar manner to the previous consideration: Information should be available but not overly specific: Tests should evaluate reasonable generalization of abstract knowledge (i.e. ability to grasp similar objects).

In these tests we include tasks involving the use of one and both arms, as a good fraction of household tasks involve the interaction of 2 limbs actuating either on the same object or on two objects interacting on the same task.

V. Discussion

In this paper we have reviewed existing methodologies to evaluate the performance of manipulators. Examples from medicine and robotics literature were presented and used to design a high-level taxonomy of benchmarking tasks. There

TABLE VIII: Sample Functional Tests

Sublevel	Examples
Unimanual	Open a door (handle/knob) Plug in a power plug Press level on an electric kettle Pick up a glass from a full dish rack Push an emergency button Pour a liquid in a wide container Stir slowly in a pot Spray from a bottle Stack cans Spoon beans (simulated feeding)
Bimanual	Cut a piece of Play-Doh on a table Rotate a steering wheel 45 degrees Empty a trash can (turn it upside down) Pick up a tray with a glass on it and transport it Open a jar with screw lid Grab an open tetrapak and pour liquid in a cup for 2s.

are still many questions that need to be addressed. We formulate some of these here:

A. Which benchmark level?

Our proposed taxonomy presents 3 main levels, from basic physical/control assessment, moving up to one-step basic tasks until reaching a functional level, in which only tasks are evaluated. Should a robot system be skillful in all these? Which level is the most important?

Service robots are conceived as helpful assistants that can carry out useful tasks for humans [12]. Hence, functional evaluation should be prioritized. Nonetheless, the other two levels (Physical and Dexterity) are still useful to test basic capabilities needed in order to tackle the functional tasks.

B. Functional vs. Component-wise

Benchmarks that address the physical level are typically difficult to generalize among different hand morphologies and actuator technologies. Specific tests however, can be helpful for specific subsets of manipulators, e.g. special purpose tests for compliant hands. As in the case of the anthropomorphism index or the graspability index, such test can provide quantitative feedback about even small changes in manipulation capabilities. In contrast, functional tests are less specific to a robot platform but offer a larger range for comparison. Last but not least, functional evaluation encourages synergy.

C. Norm Standards

Benchmarking implies comparison and evaluation of progress through time. Hence, it is necessary that we, as a community, make a joined effort to provide our results in an open, public manner and using metrics that are relevant and feasible to measure across different platforms.

While our main standard should be, theoretically, human performance, we consider nonetheless important relative evaluation: Comparison of our platform performance with respect to the current status in our field. This allows us to have a realistic perspective of our robot's strenghts and shortcomings. Towards this goal, we are currently evaluating some of the benchmarks proposed in this paper - the subject of an upcoming paper - and intend to propose them as an initial step towards creating a bank of results in the community.

D. Generality vs. Prioritized goals

One of the main reasons why benchmark tasks are so hard to agree upon is because different service robots have different goals, even if the final aim the same: Shared autonomy to operate in a household.

While it would be ideal for a robot to be able to perform all imaginable tasks in a house, we are not still there. We believe that a reasonable approach to benchmarking is one similar to the one proposed by COPM [38]. According to this approach, the goal of the evaluation is for the subject not to be able to perform every single task in a chart; rather, the subject gets to select which tasks have more priority for their particular situation and then focus on them. In humans, most - if not all - of the tasks selected per individual usually overlap (being the most common cooking and fetching objects), so even when not two people select all the same tasks, there exists enough commonalities that allow global evaluation.

E. Objective vs. Subjective Metrics

The two metrics proposed for benchmarking, namely completion time and success rate, evaluate the system as a whole. Besides being measurements which can be straightforwardly calculated, they are also advantageous because they are objective - do not depend on the evaluator's point of view.

We recognize that for service robots, there exists other metrics which can be potentially relevant, such as naturalness of movement. While important, this type of metrics are mostly subjective, or at least there are different ways of interpreting them. By selecting simply defined metrics, such as the ones we propose in this paper, we can guarantee that the evaluation is as general and applicable as possible.

F. Relationship to other taxonomies

In the manipulation field there exists many taxonomies related to the one presented here and which address more specific details of functional tasks. Taxonomies involving manipulation actions [10] can be used to further describe the tasks presented here. Taxonomies regarding grasp types [20] can be used to guide the physical tests and hand design. Taxonomies of objects of daily living [11, 39] are specially important, since they provide us with a subset of most-common objects and their characteristics, assisting us in shaping tasks that focus on useful, relevant items.

VI. Conclusion

In this paper we have presented a taxonomy of benchmark tasks for robotics manipulators. Using as inspiration the lessons learned from robotics and related fields we analyzed and proposed basic concepts as groundwork to formally define a standard manipulation benchmark methodology. We truly believe that this will only be possible if we, as a community, join efforts and work together to make scientific progress towards our shared goals.

REFERENCES

[1] Dexterous manipulation for manufacturing applications workshop. www.nist.gov/el/isd/upload/NIST-IR-7940.pdf, 2013.

[2] The Amazon Picking Challenge. http://amazonpickingchallenge.org/, 2014.

[3] Workshop on Autonomous Grasping and Manipulation: An Open Challenge. http://grasping-challenge.org/, 2014.

[4] D. Aaron and C. Jansen. Development of the Functional Dexterity Test (FDT): Construction, validity, reliability, and normative data. *Journal of Hand Therapy*, 16(1):12–21, 2003.

[5] R. Balasubramanian, L. Xu, P. Brook, J. Smith, and Y. 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.

[6] S. Barreca, C. Gowland, P. Stratford, M. Huijbregts, J. Griffiths, W. Torresin, M. Dunkley, P. Miller, and L. Masters. Development of the Chedoke Arm and Hand Activity Inventory: theoretical constructs, item generation, and selection. *Topics in Stroke Rehabilitation*, 11(4):31–42, 2004.

[7] R. Bischoff, T. Guhl, A. Wendel, F. Khatami, H. Bruyninckx, B. Siciliano, G. Pegman, M. Hägele, E. Prassler, T. Zimmermann, et al. eurobotics-shaping the future of european robotics. In *41st International Symposium on Robotics (ISR) and 6th German Conference on Robotics (ROBOTIK)*, pages 1–8, 2010.

[8] Jeannette Bohg, Antonio Morales, Tamim Asfour, and Danica Kragic. Data-driven grasp synthesis—a survey. *IEEE Transactions on Robotis*, 30, 2013.

[9] P. Bryden and E. 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.

[10] I. Bullock, R. Ma, and A. Dollar. A hand-centric classification of human and robot dexterous manipulation. *IEEE Transactions on Haptics*, 6(2):129–144, 2013.

[11] Y.S. Choi, T. Deyle, T. Chen, J. Glass, and C. Kemp. A list of household objects for robotic retrieval prioritized by people with ALS. In *IEEE International Conference on Rehabilitation Robotics*, pages 510–517, 2009.

[12] H.I. Christensen, T. Batzinger, K. Bekris, K. Bohringer, J. Bordogna, G. Bradski, O. Brock, J. Burnstein, T. Fuhlbrigge, R. Eastman, et al. A Roadmap for US Robotics: From Internet to Robotics. 2013.

[13] K. Collins, A.J. Palmer, and K. Rathmill. The development of a european benchmark for the comparison of assembly robot programming systems. In *Robot Technology and Applications*, pages 187–199. Springer, 1985.

[14] N. Dantam, H. Ben Amor, H.I. Christensen, and M. Stilman. Online multi-camera registration for bimanual workspace trajectories. In *14th IEEE-RAS International Conference on Humanoid Robots*, 2014.

[15] R. Deimel and O. Brock. A novel type of compliant, underactuated robotic hand for dexterous grasping. *Robotics: Science and Systems*, pages 1687–1692, 2014.

[16] J. Desrosiers, R. Hébert, E. Dutil, and G. Bravo. Development and reliability of an upper extremity function test for the elderly: the TEMPA. *Canadian Journal of Occupational Therapy*, 60(1):9–16, 1993.

[17] Maurice Fallon, Scott Kuindersma, Sisir Karumanchi, Matthew Antone, Toby Schneider, Hongkai Dai, Claudia Perez D'Arpino, Robin Deits, Matt DiCicco, Dehann Fourie, et al. An architecture for online affordance-based perception and whole-body planning. Technical report, Massachusetts Institute of Technology, 2014.

[18] T. Feix, I. Bullock, and A. Dollar. Analysis of human grasping behavior: Correlating tasks, objects and grasps. *IEEE Transactions on Haptics*, 7:430–441, 2014.

[19] T. Feix, I. Bullock, and A. Dollar. Analysis of human grasping behavior: Object characteristics and grasp type. *IEEE Transactions on Haptics*, 7:311–323, 2014.

[20] T. Feix, R. Pawlik, H. Schmiedmayer, J. Romero, and D. Kragic. A comprehensive grasp taxonomy. In *Robotics, Science and Systems: Workshop on Understanding the Human Hand for Advancing Robotic Manipulation*, pages 2–3, 2009.

[21] T. Feix, J. Romero, C.H. Ek, H. Schmiedmayer, and D. Kragic. A metric for comparing the anthropomorphic motion capability of artificial hands. *IEEE Transactions on Robotics*, 29(1):82–93, 2013.

[22] C. Ferrari and J. Canny. Planning optimal grasps. In *IEEE International Conference on Robotics and Automation*, pages 2290–2295, 1992.

[23] D.S. Gloss and M.G. Wardle. Use of the Minnesota rate of manipulation test for disability evaluation. *Perceptual and Motor Skills*, 55(2):527–532, 1982.

[24] Markus Grebenstein. The awiwi hand: An artificial hand for the dlr hand arm system. In *Approaching Human Performance*, pages 65–130. Springer, 2014.

[25] G. Grunwald, C. Borst, and J. et al. Zöllner. Benchmarking dexterous dual-arm/hand robotic manipulation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, Workshop on Performance Evaluation and Benchmarking, Nice, France*, 2008.

[26] D. Hackett, J. Pippine, A. Watson, C. Sullivan, and G. Pratt. An overview of the DARPA autonomous robotic manipulation (ARM) program. *Journal of the Robotics Society of Japan*, 31(4):326–329, 2013.

[27] I. Iossifidis, G. Lawitzky, S. Knoop, and R. Zöllner. Towards Benchmarking of Domestic Robotic Assistants. In *Advances in Human-Robot Interaction*, pages 403–414. Springer, 2005.

[28] R. Jebsen, N. Taylor, R.B. Trieschmann, M.J. Trotter, and L.A. Howard. An objective and standardized test of hand function. *Archives of physical medicine and rehabilitation*, 50(6):311, 1969.

[29] A. Kapandji. Clinical test of apposition and counter-apposition of the thumb. *Annales de chirurgie de la main: organe officiel des societes de chirurgie de la main*, 5(1):67–73, 1985.

[30] J. Kim, K. Iwamoto, J. Kuffner, Y. Ota, and N. Pollard. Physically based grasp quality evaluation under pose uncertainty. 29, 2013.

[31] B. Kopp, A. Kunkel, H. Flor, T. Platz, U. Rose, K. Mauritz, K. Gresser, K. McCulloch, and E. Taub. The Arm Motor Ability Test: reliability, validity, and sensitivity to change of an instrument for assessing disabilities in activities of daily living. *Archives of physical medicine and rehabilitation*, 78(6):615–620, 1997.

[32] T. Koshizaki and R. Masuda. Control of a meal assistance robot capable of using chopsticks. In *41st International Symposium on Robotics (ISR) and 6th German Conference on Robotics (ROBOTIK)*, pages 1–6. VDE, 2010.

[33] P. Kyberd, A. Murgia, M. Gasson, T. Tjerks, C. Metcalf, P. Chappell, K. Warwick, S. Lawson, and T. Barnhill. Case studies to demonstrate the range of applications of the Southampton Hand Assessment Procedure. *The British Journal of Occupational Therapy*, 72(5):212–218, 2009.

[34] D. Leidner, C. Borst, and G. Hirzinger. Things are made for what they are: Solving manipulation tasks by using functional object classes. In *12th IEEE-RAS International Conference on Humanoid Robots*, pages 429–435, 2012.

[35] C. M Light, P. H Chappell, and P. Kyberd. Establishing a standardized clinical assessment tool of pathologic and prosthetic hand function: normative data, reliability, and validity. *Archives of physical medicine and rehabilitation*, 83(6):776–783, 2002.

[36] S. Lin, J. Chang, P. Chen, and H. Mao. Hand function measures for burn patients: A literature review. *Burns (Journal of the International Society for Burn Injuries)*, 39(1):16–23, 2013.

[37] R. Madhavan, R. Lakaemper, and T. Kalmár-Nagy. Benchmarking and standardization of intelligent robotic systems. In *IEEE International Conference on Advanced Robotics*, pages 1–7, 2009.

[38] E. Månsson Lexell, S. Iwarsson, and J. Lexell. The complexity of daily occupations in multiple sclerosis. *Scandinavian journal of occupational therapy*, 13(4):241–248, 2006.

[39] K. Matheus and A. Dollar. Benchmarking grasping and manipulation: properties of the objects of daily living. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5020–5027, 2010.

[40] V. Mathiowetz, S. Rogers, M. Dowe-Keval, L. Donahoe, and C. Rennells. The purdue pegboard: norms for 14-to 19-year-olds. *American Journal of Occupational Therapy*, 40(3):174–179, 1986.

[41] V. Mathiowetz, G. Volland, N. Kashman, and K. Weber. Adult norms for the box and block test of manual dexterity. *American Journal Occupational Therapy*, 39(6):386–391, 1985.

[42] V. Mathiowetz, K. Weber, G. Volland, and N. Kashman. Reliability and validity of grip and pinch strength evaluations. *The Journal of Hand Surgery*, 9(2):222–226, 1984.

[43] W. Meeussen, M. Wise, S. Glaser, S. Chitta, C. McGann, P. Mihelich, E. Marder-Eppstein, M. Muja, V. Eruhimov, T. Foote, et al. Autonomous door opening and plugging in with a personal robot. In *IEEE International Conference on Robotics and Automation*, pages 729–736, 2010.

[44] S. Miller, J. Van Den Berg, M. Fritz, T. Darrell, K. Goldberg, and P. Abbeel. A geometric approach to robotic laundry folding. *The International Journal of Robotics Research*, 31(2):249–267, 2012.

[45] Antonio Morales, E Chinellato, PJ Sanz, AP Del Pobil, and Andrew H Fagg. Learning to predict grasp reliability for a multifinger robot hand by using visual features. *AISC proceedings*, 2004.

[46] T. Mukai, S. Hirano, H. Nakashima, Y. Kato, Y. Sakaida, S. Guo, and S. Hosoe. Development of a nursing-care assistant robot riba that can lift a human in its arms. In *Intelligent Robots and Systems (IROS), IEEE/RSJ International Conference on*, pages 5996–6001. IEEE, 2010.

[47] C. Ng, D. Ho, and SP Chow. The Moberg pickup test: results of testing with a standard protocol. *Journal of Hand Therapy*, 12(4):309–312, 1999.

[48] J. Poole, P. Burtner, T. Torres, C. McMullen, A. Markham, M.L. Marcum, J.B. Anderson, and C. Qualls. Measuring dexterity in children using the nine-hole peg test. *Journal of Hand Therapy*, 18(3):348–351, 2005.

[49] M. Roa, K. Hertkorn, F. Zacharias, C. Borst, and G. Hirzinger. Graspability map: A tool for evaluating grasp capabilities. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1768–1774, 2011.

[50] K. Schoneveld, H. Wittink, and T. Takken. Clinimetric evaluation of measurement tools used in hand therapy to assess activity and participation. *Journal of Hand Therapy*, 22(3):221–236, 2009.

[51] B. Siciliano. *Advanced bimanual manipulation: results from the DEXMART project*, volume 80. Springer, 2012.

[52] C. Smith, Y. Karayiannidis, L. Nalpantidis, X. Gratal, P. Qi, D. Dimarogonas, and D. Kragic. Dual arm manipulation: A survey. *Robotics and Autonomous Systems*, 60(10):1340–1353, 2012.

[53] C. Sollerman and A. Ejeskär. Sollerman hand function test: a standardised method and its use in tetraplegic patients. *Scandinavian Journal of Plastic and Reconstructive Surgery and Hand Surgery*, 29(2):167–176, 1995.

[54] K. Tsui, D. Feil-Seifer, M. J Matarić, and H. Yanco. Performance evaluation methods for assistive robotic technology. In *Performance Evaluation and Benchmarking of Intelligent Systems*, pages 41–66. Springer, 2009.

[55] W. van Lankveld, P. van't Pad Bosch, J. Bakker, S. Terwindt, and M. Franssen. Sequential occupational dexterity assessment (SODA): A new test to measure hand disability. *Journal of Hand Therapy*, 9(1):27–32, 1996.

[56] M. Williams, N. Hadler, and J.A. Earp. Manual ability as a marker of dependency in geriatric women. *Journal of chronic diseases*, 35(2):115–122, 1982.

[57] T. Wisspeintner, T. Van Der Zant, L. Iocchi, and S. Schiffer. Robocup@Home: Scientific competition and benchmarking for domestic service robots. *Interaction Studies*, 10(3):392–426, 2009.

[58] S. Wolf, P. Thompson, Dorian K. Morris, D., C. Winstein, E. Taub, C. Giuliani, and S. 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.

[59] N. Yozbatiran, L. Der-Yeghiaian, and S. Cramer. A standardized approach to performing the action research arm test. *Neurorehabilitation and Neural Repair*, 22(1):78–90, 2008.

[60] M. Zillich. My Robot is Smarter than Your Robot-On the Need for a Total Turing Test for Robots. *Revisiting Turing and his Test: Comprehensiveness, Qualia and the Real World*, page 12.