The Cluttered Environment Picking Benchmark (CEPB) for Advanced Warehouse Automation

Evaluating the Perception, Planning, Control, and Grasping of Manipulation Systems

By Salvatore D'Avella, Matteo Bianchi, Ashok M. Sundaram, Carlo Alberto Avizzano, Máximo A. Roa, and Paolo Tripicchio

Autonomous and reliable robotic grasping is a desirable functionality in robotic manipulation and is still an open problem. Standardized benchmarks are important tools for evaluating and comparing robotic grasping and manipulation systems among different research groups and also for sharing with the community the best practices to learn from errors. An ideal benchmarking protocol should encompass the different aspects underpinning grasp execution, including the mechatronic design of grippers, planning, perception, and control to give information on each aspect and the overall problem. This article gives an overview of the benchmarks, datasets, and competitions that have been proposed and adopted in the last few years and presents a novel benchmark with protocols for different tasks that evaluate both the single components of the system and the system as a whole, introducing an evaluation metric that allows for a fair comparison in highly cluttered scenes taking into account the difficulty of the clutter. A website dedicated to the benchmark containing information on the different tasks, maintaining the leaderboards, and serving as a contact point for the community is also provided.

Digital Object Identifier 10.1109/MRA.2023.3310861
Date of current version: 26 September 2023

INTRODUCTION

Since the first years of life, children learn by experience how to grasp objects of different shapes and in different scenarios. Thanks to that, for an adult human being, pick and place

becomes a mechanical movement, and it is quite easy to understand how to grab an object never seen before thanks to their own baggage of experience [1]. Nowadays, pick and place is one of the most repetitive tasks for human workers, and it is also one of the golden standard tasks used to assess the capabilities of manipulator robots. Even today, in many warehouses, a lot of human pickers stand in front of a shelf all the time and repeatedly pick objects from that shelf to place them into bins. It appears clear that the automation of pick and place actions is a hot topic for industries because it would allow increasing the throughput while lowering expenses. On the other hand, it is a challenging problem for the robotics research community. Even if it is a simple task for a human, depending on the boundary conditions, it could be difficult for a robot to pick and place objects, especially when they are in a cluttered environment. For this purpose, many technical issues have to be solved. So far, there are no autonomous robots able to face the unpredictability of complex industrial environments that can be envisioned in the near future. As happens when a topic becomes popular, many solutions have been proposed to tackle this issue. Each of them adopts its own workflow and performs validation tests on a different set of objects using distinct modalities and criteria. For this reason, a unified benchmark that provides the guidelines and the set of objects enabling reproducibility and comparison across different solutions represents an important step to advance the developments in the area.

Note that evaluating pick and place operations, and in general, manipulation tasks, is not easy, since the involved pipeline is often complex, encompassing vision, planning, control, actuation, sensors, and grasping. Disentangling each component is not trivial, and understanding which of them gives the most significant contribution to the overall system performance is of crucial importance to evaluate and, eventually, redesign and improve the overall system. In recent years, several protocols, benchmarks, and datasets have been proposed to provide a contribution to the community and offer tools for the evaluation of autonomous robotic platforms [2], [3], [4], but some of them did not use standardized objects or scenarios, making the experiments not reproducible or comparable; others evaluated just some parts of the complex system without taking into account the ensemble [5]; others, on the contrary, focused just on the whole system, neglecting the single components in favor of the completion of the task [6]; and others used stringent protocols that are not effective for the unpredictability and flexibility of the future industrial scenarios [7]. Furthermore, differently from research fields that can be precisely evaluated just on data and simulations, for robotics applications, it is important to test the system with real experiments on physical objects since simulations cannot be considered reliable. An easily reproducible protocol and a methodical benchmark that would allow engaging a vast community of robotic researchers for the comparative evaluation of the results to improve previously developed approaches has not been established yet.

This work proposes to bridge this gap by providing the following contributions:

- a comprehensive literature survey of existing benchmarks, challenges, and datasets employed in the different subproblems of the pick and place task
- a novel benchmark framework consisting of
  - a selected list of objects to be used in the tests
  - protocols for different tasks that evaluate both the single components (vision, planning, control, sensors, and grabbing) of the system and the system as a whole
  - an evaluation metric for each of the proposed tasks
- a novel photorealistic dataset developed to mimic the cluttered scenes of the proposed benchmark (Figure 1), thus containing both rigid and soft/deformable objects and even objects filled with liquid that present a complex dynamic, which can be useful as a tool for training deep learning-based vision pipelines
- the introduction of a complexity estimation algorithm validated on the photorealistic dataset
- a baseline for one of the tasks
- a website at http://cepbbenchmark.eu/ that provides some guidelines for the protocols and that allows continuous submissions and updated leaderboards.

The proposed protocols span from tasks on individual objects for evaluating targeted components of the system to heavily cluttered scenes. The experimental setup, the procedures, and the evaluation metrics have been designed aiming at reproducibility without constraining the scenario and allowing comparisons among research groups. For this purpose, the novel evaluation metric guarantees a fair comparison, leaving some flexibility due to the randomness of the scenarios, thus mimicking the unpredictability of the future industrial environments.

This work does not introduce yet another completely new object set to reinvent the wheel but proposes a selection of objects taken from adopted existing benchmarks. The objects in the proposed set have the objective of stressing the components of the manipulation pipeline separately and as a system. The objects not only present different sizes, shapes, and weights but also have diverse rigidity and texture properties that pose difficulties to the grasping and perception part. Having access to these

FIGURE 1. A possible scenario to face within the benchmark with the complete forty-object set.

objects is important for the experiments since it is not possible to have a general overview of the performance relying on simulations only. Through the website, it is possible to find the necessary information to get all the objects used for the benchmark. The website will serve as a contact point for other researchers who would like to contribute to establishing an active community. As the world of robotics is growing very fast, the benchmark has been designed to be modular, allowing it to be updated with new protocols reflecting new industrial challenges.

The remainder of the work is structured as follows: the "Literature" section reviews the existing benchmarks, competitions, and datasets highlighting their scope and limitations; the "Design Choices" section describes and motivates the objects chosen for the test; the "Guidelines" section introduces the guidelines to reproduce the experiments; the "Evaluation Metrics" section explains the novel procedure adopted for evaluating the performance of the system exploiting a complexity evaluation algorithm trained on the photorealistic dataset presented in the "Photorealistic Dataset" section; the "Baseline" section depicts the approach used for accomplishing one of the benchmark's tasks; and the "Conclusions" section summarizes the contribution of this work.

LITERATURE

A robotic system for pick and place in a cluttered environment typically consists of a manipulation arm, a vision system, and objects to be manipulated. Grasping is a complex problem since it is a multidisciplinary task that spans from the mechatronic design of grippers to higher-level domains like perception, planning, and control. The lack of common guidelines causes difficulties in quantitatively understanding the performances of different systems.

Many related works concentrated on only one of the aspects of the manipulation system. The Cornell Grasping dataset [8] and VisGraB [9] concentrate on the aspect of manipulation. In particular, the former provides data for the manipulation task representing antipodal grasps as rectangles aligned to the pose of the end-effector, focusing on two-finger grippers only. The latter puts attention more on the simulation providing open frameworks to compare object manipulation capabilities. Although many manipulation activities start from simulated environments since simulations can give access to unavailable platforms, are intrinsically safe, and facilitate the reproducibility of the experiment, they are not realistic and fully reliable for what concerns the control level and the interaction with the objects [10]. Furthermore, most simulations are designed for rigid bodies but cannot deal with deformable objects or liquids. Unfortunately, unlike perception algorithms or other disciplines like navigation and SLAM, robotic manipulation cannot be primarily evaluated just on digital data, but real experiments on physical objects are necessary for accurately understanding the system's performance [11].

The YCB benchmark [2] responds to the lack of a standardized set of physical objects, selecting a dataset of daily life items leveraging studies concerning the rehabilitation of the human upper limb. It also proposes an evaluation framework and several examples of task protocols where the grasping aspect of manipulation is preponderant over the vision or planning factors, as also suggested by the selected set of objects. In addition, the proposed setups do not consider cluttered environments of heterogeneous items that are the most critical scenario to face for Industry 4.0. The ACRV picking benchmark [7] is a recent work that presents a set of 42 physical objects and illustrates a detailed procedure on how to conduct the experiments to be reproducible. Even if it proposes a well-defined way to evaluate the complete robotic system, it specifies the placement of the objects to guarantee the reproducibility of the experiment.

Mnyusiwalla et al. [3] propose a bin-picking benchmark with a protocol, objects, and evaluation system for picking fruits and vegetables from a container and placing them in an ordered bin. The work uses just a limited number of items, not covering all possible manipulation difficulties, and it proposes 15 different scenarios, from very simple to more complex, focusing on a particular type of clutter with a multitude of the same object. Morgan et al. [12] introduce a benchmark for pick and place inspired by the clinical box and block tests used for the evaluation of the upper limb manipulation dexterity of physically impaired individuals. Even if the test is conducted in cluttered conditions, the items are only square bricks of the same size and varying colors. Furthermore, the considerations for the experiment are focused just on hand-shaped end-effectors. Recently, Bekiroglu et al. [4] presented a benchmarking protocol for the evaluation of grasp planning algorithms. They selected seven objects from the YCB dataset and described how to set up the workspace and where to place the objects. Even this work concentrated mostly on robotic hands and was not effective for industrial scenarios, since the placement of the objects was too stringent and did not consider cluttered cases.

In this discussion, robotics challenges should also be taken into account. In recent years they have been a decisive way to drive scientific progress. The DARPA Autonomous Robotic Manipulation Competition and the IROS Robotic Grasping and Manipulation Competition (RGMC) [13] push for manipulators with a high degree of autonomy able to grasp and manipulate a wide range of object geometries in unstructured environments across diverse application spaces. In particular, the IROS RGMC was also held during the pandemic era due to the SARS-CoV-2 virus in the online version with the Open Cloud Robot Table Organization Challenge cloud-based benchmark [14], where the participants uploaded their solutions to a remote server that executed the code on remote robot setups. Another example is the Amazon Picking Challenge (APC) [6]. It was one of the most visible events in the robotic scenario. It was an annual competition from 2015 to 2017 that tried to strengthen the ties between industry and the research community to realize an autonomous robotic platform for picking objects in a cluttered environment.

Even though the competitions have the merit to spur the advance of the research, concerning benchmarking and

repeatability, they are limited to the participants, use objects and setups that are hard to replicate by other researchers not involved in the challenge, and evaluate the solutions only at the system level, making it difficult to understand which component of the complex system contributed most to the success or the failure of the task.

The literature review denotes that a lot of work has been done to facilitate the comparison of the performance of manipulation systems at different levels thanks to datasets, competitions, and benchmarks. To summarize, many contributions focus on a single aspect (i.e., perception, planning, control, or holding/grasping of objects) of the solution, not considering that the orchestration of the single components would also be important for a sweet melody; others, instead, evaluate the solutions only at the system level, losing a finer understanding of which component contributes most to the success or failure of the system. If YCB tried to standardize the set of objects, at least for the manipulation aspect, much more needs to be done, especially for what concerns the practical industrial field and more complex cluttered scenarios. Note that most of the works consider the difficulty of the objects in absolute value during their selection and sometimes in the scoring phase give more points for the more challenging ones. Such a decision can lead to unprecise evaluations and encourage the use of a particular type of gripper over others since the difficulty is not in correlation to the typology of the end-effector used by the robotic system and its perception pipeline or to the arrangements of the object in the scene.

It is worth noticing that if the aforementioned works have gained a lot of popularity in the research community, the protocols they proposed have been rarely adopted by other research groups, excluding from this computation the manuscripts produced by the same group that proposed that benchmark and without taking into account the competitions for the reasons already discussed. We found that only one or two articles used each of the benchmarks proposed in the literature at the moment of writing. On the other hand, the vast majority of articles that cited YCB [2] and ACRV [7] only used the standardized set of objects for their experiments but not the protocols, demonstrating that the effort in standardizing the object set is perceived as important in the research community. The motivations behind the poor adoption of the protocols are multiple and diverse for each of the works. Most of the protocols proposed in [2] refer to general tasks that are not suitable for the specificity of the experimental validation required by the research groups. The selected objects in [3], [4], and [5] are not representative of the complexity of real-world tasks without offering a challenging test bench that could help to advance the research in the field. In addition, for [3] and [5], the set of objects is different from the widely used objects proposed in [2] and [7] and too specific for that task. Moreover, [4] and [7] decided to constrain the positions of the objects to allow a fair comparison in settings that probably are not representative of a big audience. The lack of a website with a leaderboard or a way to attract interested people working in the same field to build a community in [3], [4], and [5] can also be a big limitation.

The presented benchmark framework proposes objects mainly coming from already adopted datasets leveraging previous research. The characteristics of the items are taken into account during the selection process, evaluating the difficulty that different gripper typologies and vision systems may have in grabbing and perceiving them in cluttered situations, respectively. The protocol guarantees repeatability and comparability but leaves some degree of randomness to emulate the unpredictability of industrial environments thanks to the evaluation metric that considers the difficulty of the clutter. The benchmark proposes several protocols that test the manipulation system at each level, requiring the users to report the characteristics of the adopted solution and the causes of each failure to better exploit the purpose of a benchmarking system in favor of the research community and drive progress. The focus of the benchmark proposed in this work is specifically on bin-picking in cluttered environments, which is a hot topic for the industrial/logicist sector but also for the research community in several aspects. However, this work should be intended as the starting point to build a globally shared community oriented to the evaluation of industrial protocols enriching the benchmark with protocols reflecting the new challenges of the future. For example, in the future, it could be interesting to add other sensing modalities, e.g., tactile sensing, or to introduce human–robot collaboration scenarios to our protocols.

DESIGN CHOICES

The set of objects has been chosen to pose difficulties to the end-effector and the perception system. In such a way, the hardware design of the gripper and the software level built upon it can be evaluated. This benchmark mostly uses the objects already present in existing datasets for the different subproblems that constitute the manipulation task.

A. GRIPPERS

The most used grippers in manipulation tasks, especially when they are related to industrial objectives, are parallel-jaw grippers, suction grippers, and, for the last few years, soft grippers like the universal jamming gripper or pneumatic soft grippers. Magnetic grippers are also simple to actuate, but they are not effective for the tasks discussed in this work, since they can pick only ferromagnetic objects. Therefore, magnetic grippers are not investigated in the following. More elaborate solutions like multifingered hands are not commonly employed in these kinds of tasks due to their hardware complexity and/or the software effort needed to control them. However, their interest is also increasing in the industrial field, especially in cobotics applications. They come with a wider range of hardware characteristics: anthropomorphic and not anthropomorphic, with a different number of fingers that can be completely actuated or underactuated in the form of rigid or soft (continuous or articulated) devices. Each of these solutions comes with specific control strategies for the execution of grasping and

manipulation actions, resulting in a significant task dependence of the performance that can be achieved [15]. For such a reason, it is very challenging to identify a unique benchmarking framework that can be valid for all state-of-the-art technologies. There is an open debate on how to evaluate all the pipelines correctly in autonomous manipulation systems endowed with hands as end-effectors. Dealing with this issue is out of the scope of this work.

B. PERCEPTION

Vision systems and perception algorithms are related to the typology to which the gripper belongs. Procedural algorithms inspect the scene to find geometrical features relevant to the type of gripper employed in the system. In particular, antipodal grasping points are suitable for parallel-jaw grippers, planar surfaces are the best for suction grippers, and edges or corners are convenient for universal jamming grippers. Compared with industrial grippers, artificial hands come with a wider range of hardware characteristics, and thus, a rule of thumb to search for the best grasping point does not exist.

C. OBJECTS

Forty objects coming from existing benchmarks and datasets have been chosen. The standardized set of objects of the YCB dataset [2] is not enough to test all the components of the manipulation pipeline, and therefore, it has been enriched with other objects of the ACRV picking benchmark from the APC [7] and T-LESS [16]. Changes to the objects have been applied only to very few of them to guarantee easy availability worldwide to buy them while maintaining the same original properties. They present different levels of difficulty. Indeed, they can vary in size, shape, and weight, and have diverse surface materials and texture properties. There are objects with reflective, perforated, or symmetric surfaces that are challenging for the vision; others have deformable surfaces or strong orientation constraints and shift their centers of mass when manipulated. All these problems are accentuated in the clutter because accurate segmentation and stable grasp are more difficult. Table 1 lists all the objects, and for each one of them, it shows the original dataset to which the object belongs, the level of difficulty assigned for the different grippers, and the final score that will be useful for the performance evaluation. The difficulties have been assigned through a consensus protocol disseminating questionnaires among several colleagues. They were invited to select a difficulty score among three possible levels of difficulty (easy, medium, and hard) for each of the forty objects concerning a generic pick and place task using the aforementioned grippers (parallel-jaw, suction, and soft) without considering issues related to the vision pipeline. They had to assign an average score trying to imagine multiple scenarios in which a single object could be found in diverse configurations ranging from the simplest pose to pick to the most challenging configuration. A total of 68 questionnaires should have been collected according to the known formula (1)

$$\text{sample\_size} = \frac{\frac{z^2 p(1-p)}{e^2}}{1 + \frac{z^2 p(1-p)}{e^2 N}} \tag{1}$$

for computing the required sample size to have a confidence level z of 90% and an error margin e of 10%, considering that the population of the Robotic and Automation Society, N, is about 15,000 members. In particular, 75 questionnaires have been received from colleagues who have different expertise levels on such topics. Analyzing the data, the distributions for each of the forty objects and gripper categories follow a unimodal profile having the mode being at least 60% of the total. The total score for each object is the average of the difficulties of each gripper and the vision system. The difficulty can assume only three different discrete values that go from 1 to 3, resembling the linguistic variables "easy," "medium," and "hard," respectively.

The objects are subdivided into four subsets of ten elements each (see Table 2). For every subset, the mean difficulty should be the same for all the gripper typologies considered in Table 1, thus preventing the selection of a particular gripper in favor of the others, relying on the nature of the objects.

Table 1 reports the difficulties concerning just the vision and the most used industrial grippers, but even other end-effectors like anthropomorphic hands or other soft grippers can be used for this benchmark. For these other grippers, each item is considered as having a "medium" difficulty.

GUIDELINES

Considering the vast robotic applications and the increasing attention in always wider fields, it would be overbearing to cover all the possible areas of interest and remain relevant forever. However, we provide several industrial-oriented task protocols that are meant to examine most of the recent needs of such an industrial revolution (Industry 4.0) and beyond ranging from flexible automation and generalizable grasping to cluttered environments.

The benchmark has a modular design and is organized in stages that can have intermediate phases. In principle, stages are meant to represent an industrial relevant task, which is identified by the final phase test, while the intermediate phases of each stage aim at evaluating a specific subproblem of the manipulation task before getting to the final phase test, which puts all the intermediate skills together for different objectives. The user can apply for each stage independently and even for a specific intermediate phase using one of the subsets or the full dataset. Therefore, the website has a leaderboard for every component of the stages, separating the results per subset. However, stages are presented with an increasing level of complexity and should be addressed following that order and also completing the intermediate phases to have a clear picture

TABLE 1. Dataset for the Cluttered Environment Picking Benchmark (CEPB).

				DIFFICULTY				
	OBJECT	ORIGINAL DATASET	GENERALIZATION	PARALLEL JAW	SUCTION	SOFT GRIPPERS (I.E., UJG)	VISION	TOTAL
1	Cheez-It cracker box	YCB obj#1		1	1	3	1	1.5
2	French's mustard bottle	YCB obj#9		3	2	3	1	2.25
3	Tomato soup can	YCB obj#10		2	1	3	1	1.75
4	Scissors	YCB obj#35		3	2	1	2	2
5	Foam brick	YCB obj#57		1	3	1	1	1.5
6	Small clamp	YCB obj#46		3	3	2	2	2.5
7	StarKist tuna fish can	YCB obj#7		1	1	1	2	1.25
8	Plastic banana	YCB obj#11		2	2	1	1	1.5
9	Meat can	YCB obj#5		1	2	3	1	1.75
10	Mug	YCB obj#31		2	2	1	1	1.5
11	Padlock	YCB obj#38	X	2	3	1	2	2
12	Baseball	YCB obj#51	X	3	2	3	1	2.25
13	Bowl	YCB obj#25	X	3	1	3	1	2
14	Sleeve	T-LESS obj#13	X	1	2	1	2	1.5
15	Coca-Cola bottle, half	CEPB	X	3	3	3	3	3
16	ICRA duckie	APC/ACRV obj#1	X	2	3	2	1	2
17	Elmers school glue	APC/ACRV obj#20	X	1	1	2	1	1.25
18	Dice	YCB obj#58	X	3	1	1	2	1.75
19	Eggs plush puppies	APC/ACRV obj#13	X	2	3	3	1	2.25
20	Scotch duct tape	APC/ACRV obj#16	X	1	2	2	1	1.5
21	Haribo golden bears	CEPB		3	1	3	1	2
22	Plint board*	T-LESS obj#14		1	2	1	3	1.75
23	Flat screwdriver	YCB obj#43		3	3	2	1	2.25
24	Clamping plate*	T-LESS obj#15		1	2	1	3	1.75
25	Coca-Cola bottle, full	CEPB		2	2	3	2	2.25
26	Kong duck dog toy	APC		2	3	3	1	2.25
27	Spoon	YCB obj#27		3	3	2	2	2.5
28	Plastic strawberry	YCB obj#12		3	3	2	1	2.25
29	Paper towels	CEPB		3	3	3	1	2.5
30	Stabilo OHPen	CEPB		2	1	3	2	2
31	Plastic white cup	APC/ACRV obj#30		3	3	3	2	2.75
32	Wine glass	YCB obj#30		3	3	2	3	2.75
33	Key	YCB obj#38		3	3	2	2	2.5
34	Nail	YCB obj#40		3	3	2	3	2.75
35	Adjustable wrench	YCB obj#44		3	3	3	1	2.5
36	T-shirt	YCB obj#70		2	3	3	2	2.5
37	Rolodex jumbo pencil cup	APC/ACRV obj#3		2	2	2	3	2.25
38	Glove	APC/ACRV obj#22		2	3	3	2	2.25
39	Laugh Out Loud joke book	APC		3	1	2	1	1.75
40	Pringles chips can	YCB obj#8		1	1	3	1	1.5
41	Timer	YCB obj#71						
42	Clear box	YCB obj#63						
43	Clear box	IKEA obj#SAMLA(301.029.74)						

The objects with the generalization mark should be used in stage 2. The difficulty values are easy, medium, and hard. The total difficulty varies from a minimum of 1 to a maximum of 3 and is computed as the average among the difficulties for each gripper and the vision system.
*These two objects belonging to the T-LESS dataset have been replaced with similar objects that can be easily bought.

of the system's performance. Each of the proposed task protocols should be repeated a minimum of three consecutive times to be meaningful while balancing experimental time. This number has been chosen considering the most recent research approaches discussed in the literature review (the "Literature" section). Such a compromise allows for testing the robustness of the system and collecting important information that can enable a deeper understanding of its properties.

The user can conveniently place the robotic platform by choosing a desired placement since the success of the experimental tests will depend on the reachable workspace of the arm in each task. The working area must be divided into two parts, which hereafter will be called working table A for the pick phase and working table B for the placement phase. However, they must not necessarily be two physically separated tables but could also be two clearly distinct regions of the same table. The placement is considered successful if the target object is anywhere on working table B everywhere unless otherwise specified in the task protocol. A box (obj#42 or obj#43 in Table 1), which contains the objects, is laid down on working table A, whose dimensions can be arbitrarily chosen. Obj#42 and obj#43 differ just in their dimensions. Indeed, obj#42 is used for the intermediate tasks in which at most ten objects are involved, while obj#43, which is larger, is used for the final tasks where the items are employed altogether. The system must not have any prior knowledge of the position of the objects. It should only know the employed subset. At the end of each task, the user(s) should report the time (in seconds) the robotic system employed to complete the task, possibly giving the average time spent for the perception, grasp planning, and execution. Grasping and placing multiple items at the same time is considered an error unless specified by the task protocol. Furthermore, any external interventions are inadmissible after the robot has started moving. Therefore, dropped objects cannot be reintroduced.

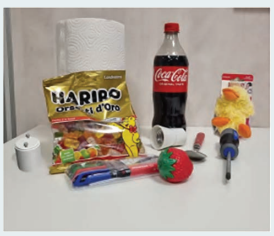
The following paragraphs introduce the experimental stages that can be executed separately per subset.

A. STAGE 1

PHASE 1 (VISION): *VISION OF NONSEQUENTIAL INDIVIDUAL OBJECTS*

The items are placed in the middle of the large clear box (obj#43) on working table A, and the vision system should be able to detect each of them in isolation, regardless of their configuration. Every object should be posed in different configurations that could vary depending on the object's geometry and its characteristics. Figure 2 gives some visual hints. The website documents for each object the configurations that should be tested in this stage. In this way, each object has to be recognized multiple times. The score is determined by the number of objects correctly detected in each of their configurations. The detection can be 2D or 3D and is considered correct if the intersection over union between the prediction and the (manually) labeled instance is greater than 0.75, matching the class of the item instance. When fails occur, the user(s) should report in which displacement it was not able to identify the identity of the object.

TABLE 2. Subsets for the CEPB benchmark.



	SUBSET 1	SUBSET 2	SUBSET 3	SUBSET 4
1	Cheez-It cracker box	Padlock	Plastic strawberry	Plastic white cup
2	Tomato soup can	Bowl	Haribo	Wine glass
3	Plastic banana	Sleeve	Flat screwdriver	Key
4	StarKist tuna fish can	Coca-Cola bottle, half	Clamping plate	Adjustable wrench
5	Scissors	Baseball	Kong duck dog toy	Dice
6	Foam brick	ICRA duckie	Coca-Cola bottle, full	Rolodex jumbo pencil cup
7	Meat can	Dice	Spoon	Nail
8	Small clamp	Elmers washable no-run school glue	Paper towel	Gloves
9	Mug	Kygen Squeakin' Eggs plush	Highlighters	Laugh Out Loud joke book
10	French's mustard bottle	Scotch duct tape	Plint board	Pringles chips can

PHASE 2 (PICKING):
PICK AND PLACE OF NONSEQUENTIAL INDIVIDUAL OBJECTS

The items are placed in five different positions in the large clear box (obj#43) on working table A, and the system should be able to pick the target, regardless of its pose. The five positions are the middle and the corners of the box to check if the manipulator has sufficient dexterity in all the workspace. The object's configurations are those of the previous phase. Figure 3 clarifies the concept, providing a visual example using the same objects depicted in Figure 2. The storage system for the place of the target and its position in the workspace are arbitrary choices. They do not affect the evaluation of the system performance and, therefore, can be designed considering the details of the robotics setup.

When failures occur, the system user(s) should describe the cause of the error. In particular, they should report if the failure has happened because the path planning was not correctly computed or if the hardware was not able to follow the path, if the end-effector could not grasp the target, or if the object was dropped before the placing. If the item falls out of the working table or is broken, it should be reported as well, and the object cannot be reintroduced in the pool for a further trial. The score is given by the number of objects correctly picked and placed in the storage system for each position and configuration.

FINAL PHASE TEST (CLUTTER):
PICK AND PLACE OF NONSEQUENTIAL OBJECTS IN A CLUTTERED ENVIRONMENT

The protocol proposes the bin-picking task, where the robot should grasp a single arbitrary instance of the specified item. The objects used in the intermediate phases of this stage should be used in their subsets in a final test that assesses the performance of the overall system taking into account vision, path planning, and holding of objects at the same time. The final stage should be repeated three consecutive times. When using the individual subsets, the ten objects should be placed in the larger clear box (obj#43) for shaking and then thrown gently in the smaller clear box (obj#42) on working table A. On the contrary, when using the full subset, the objects should be placed in the smaller clear box (obj#42) per subset, shaken, and thrown gently in the bigger clear box (obj#43) on working table A. The score is given by the number of objects correctly picked and placed in the storage system. When failures occur, the user should report the causes of the failures, as it is stated in the aforementioned intermediate phases (1 and 2), not specifying, of course, the configuration of the objects since they are determined by the randomness of the shake. The user should also report if the robotics system is blocked for some reason, trying to pick the same target without any progress. In this case, after ten attempts, the test is considered concluded, and the score counts just the objects correctly placed in the storage system until that moment.

B. STAGE 2

For systems based on neural networks, it is strategic to evaluate the generalization capacities of the perception part, which is fundamental for computing the correct pose of the end-effector to grab the target. For this reason, such systems can be trained with all the available objects in the dataset except the ones labeled as generalization items (see Table 1). In this stage, the system should be able to pick objects that it has not encountered during the training phase.

PHASE 1 (UNKNOWN PICKING):
PICK AND PLACE OF NONSEQUENTIAL UNKNOWN INDIVIDUAL OBJECTS

This test is similar to picking, but the targets are all the generalization items. When failures occur, the

FIGURE 2. An illustrative example of the orientation configurations for three objects with different geometry and texture properties to assess the performance of the vision pipeline in phase 1 of stage 1. (a) The Cheez-It cracker box has more configurations with respect to the other two due to its symmetry along all the three principal axes (x, y, z) and its diversified texture, (b) the tomato soup can has just four configurations due to its cylindrical symmetry, and (c) the Kong duck dog toy can only be placed laying on the floor of the box since it has only two different stable poses on a table but presents different visual appearances depending on the side it is lying on.

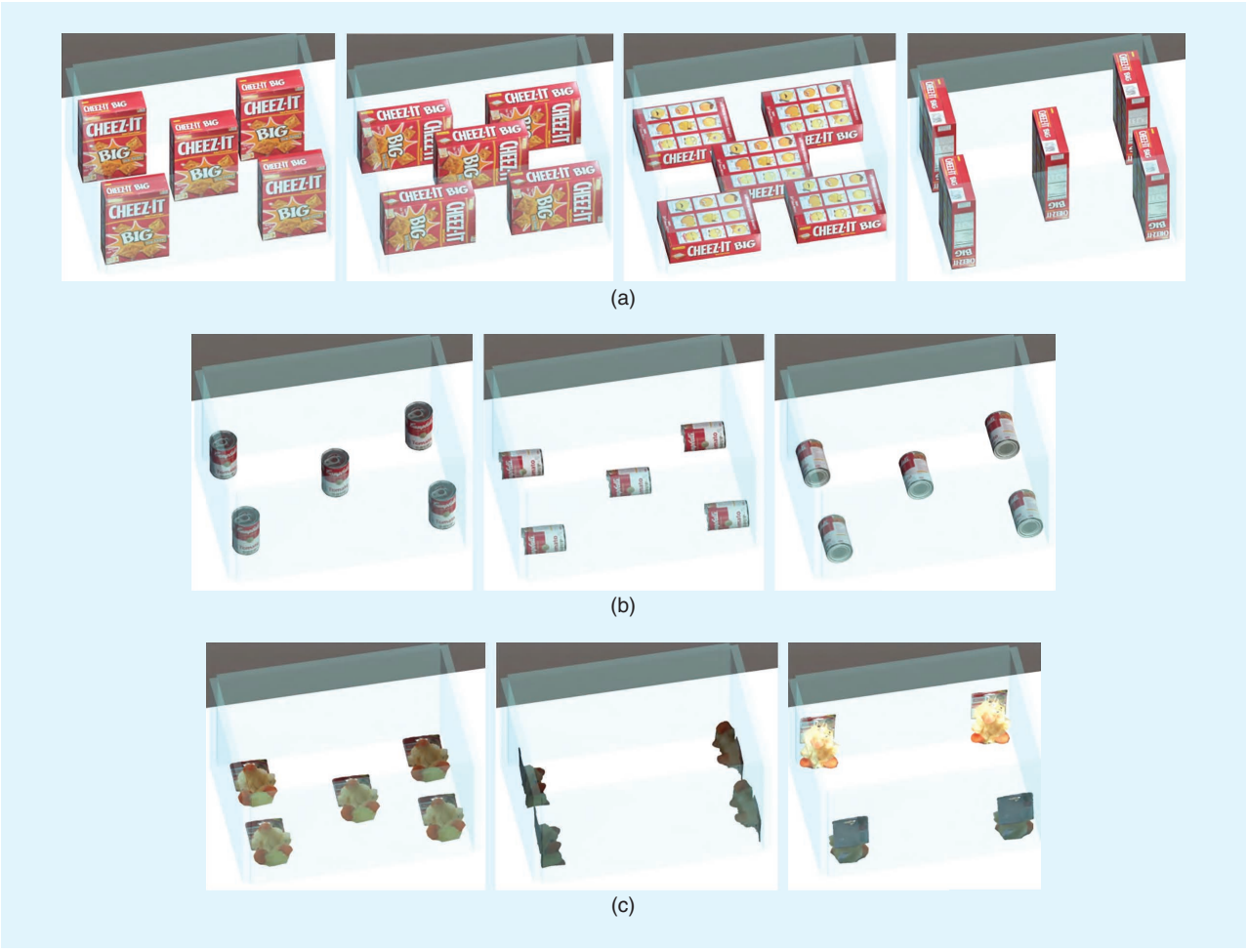


FIGURE 3. (a)–(c) An illustrative example of the position configurations in the box for the same objects presented in Figure 2 to assess the dexterity of the manipulator in phase 2 of stage 1. The Cheez-It cracker box (a) has more possible configurations than the other two, but to assess the system dexterity, it is sufficient to test fewer orientation configurations. The same consideration applies to all objects. It could happen, like in the case of the Kong duck dog toy (c), that many objects that could not stand isolated in the center of the box can instead lean on the corners.

system user(s) should also report if it happens because the vision part could not detect the correct pose of the end-effector.

FINAL PHASE TEST (UNKNOWN CLUTTER): *PICK AND PLACE OF NONSEQUENTIAL UNKNOWN OBJECTS IN A CLUTTERED ENVIRONMENT*

This test is similar to clutter, but the clutter is composed only of the objects of subset 2 (generalization items).

C. STAGE 3

Until now, the stages presented before do not consider a precise order for picking the objects, but every robotic system can decide depending on their needs. This stage considers a possible application of the manipulation system in an industrial environment in which the robot interacts with other devices like Programmable Logic Controllers, receiving information on which target to pick at each time. This stage has no intermediate phases but only the final stage test.

FINAL PHASE TEST (SEQUENTIAL): *PICK AND PLACE OF SEQUENTIAL OBJECTS IN A CLUTTERED ENVIRONMENT*

This test is similar to the clutter test, but the sequence in which objects have to be picked and placed is predefined. The website also reports the sequence order for all the subsets. When failures occur, the system user(s) should also report if this happened because the sequence was not respected.

The benchmark is meant to be open to the research community. The organization of the task protocols in stages allows the benchmark to be modular and flexible. Anyone interested in contributing to proposing new task protocols should respect the general organization of the other task protocols articulated in the intermediate phase(s) and the final phase and can share their work by contacting the organizers through the website http://cepbbenchmark.eu/.

EVALUATION METRICS

The score of the intermediate phases of each stage (the ones that present the objects in isolation) is computed

algorithmically, performing the weighted sum of the total difficulty of the successfully picked objects. N being the total number of objects employed in the experiment, d_i being the difficulty of the ith item (see last column of Table 1), and d_i^* being the difficulty of the ith item that has been successfully picked and placed, the score can be formulated as follows:

$$s_I = \sum_{i=1}^{N} \frac{d_i^*}{d_i} \tag{2}$$

The score for the final stage tests (the ones in a cluttered environment) takes into account the involved objects and the difficulty of the clutter. Many related works compute the evaluation metric at the system level as $R = n^*/N$ considering just the amount of objects picked and placed correctly n^* over the total of the N objects. Only a few other benchmarks weigh the score by taking into account the difficulty of the objects, but none of them deal with the level of clutter in the scene. The weight of each item in the evaluation metric corresponds to its "total" difficulty reported in the last column of Table 1. The complexity of the clutter is measured by analyzing the occlusion percentage of the scene that depends on the random placement of the objects after the shaking of the container. This randomness allows for generating scenes always diverse from each other that can reproduce the unpredictability of industrial environments.

The evaluation metric is formulated as follows:

$$s_F = s_I \times \sum_{i=1}^{N} d_i(1+\gamma) \tag{3}$$

where s_I is computed as formulated in (2) and represents the score that can be obtained in a scenario in which all the objects are isolated and well separated from each other, while γ depends on the clutter percentage (from 0.0 to 1.0) denoting the difficulty introduced by the disorder. The factor γ is computed by summing up the outputs of the sigmoid function $(1/2)(1/1 + e^{-30(x+0.25)})$, which, for each object employed in the test, takes as input the surface percentage x occluded in the clutter. Roughly speaking, the sigmoid function has been designed to start accounting for the additional difficulty of an object introduced by the clutter when its surface is occluded more than 10% and saturate the contribution of the disorder complexity up to 0.5 when the occluded surface is greater or equal to 50%. The algorithm used to find out the factor γ is shown in the form of pseudocode in Algorithm 1, where the 3D cuboid information from the photorealistic dataset is assumed to be known. It is worth noticing that at the moment of the performance evaluation, the occluded surfaces are computed using the polygons derived from the manual segmentation provided by the users, as also shown in the "Baseline" section. The segmentation consists of multipoint polygons that should shape the silhouette of the objects, but the objective is to employ a neural network to automate the process. According to the latest results reported by the benchmark for 6D pose estimation (BOP) benchmark [17], state-of-the-art solutions cannot yet achieve a reliable performance in reconstructing the 6D pose of the objects in cluttered scenes. Indeed, the statistics shown by the BOP benchmark reported that the scores suffered from a huge drop even at low levels of occlusion, as demonstrated by the 30% gap of difference in performance obtained in LINEMODE and Occluded-LINEMODE that provides the same objects but partially occluded. Estimating the 6D pose of objects is an active field with important practical implications, and after 2018, other works [18], [19] have been published, showing a margin of improvement for several aspects. Therefore, the authors believe that in the near future, such methods can be employed for the proposed benchmark to automatically detect the occlusion percentage of cluttered scenes in the evaluation metric, but in the meanwhile, manual segmentation guarantees more accurate measurements.

Of course, the time required by the robotic system to complete the task can affect the score. This benchmark establishes two score categories: one with an unlimited amount of time at disposal for those who want to concentrate mostly on the accuracy of the manipulation task moving the robot at a moderate speed and the other with elapsing time for those who want to take into account time constraints moving the robot at high speed. For the latter category, the score is computed as follows:

$$s_F(t) = \frac{s_F \times (\text{max\_time} - t)}{\text{max\_time}} \tag{4}$$

where t is the elapsed time, and max_time is the estimated maximum time which is $40\ \text{s} \times 2N$, thus giving two possibilities of grasping per object.

ALGORITHM 1: Algorithm used to compute the clutter percentage of the starting scene and the scene difficulty.

Data: 3D bounding box for each object *cuboid*, objects difficulties, *difficulties*
Result: γ, scene_difficulty
objects = objects used in image *I*;
cuboids = 8 *x*, *y*, *z* coordinates of the objects in *I*;
polygons = 2D polygons derived from *cuboids*;
for *each obj in objects* **do**
  $cuboid_{obj}$ = pose corresponding to *obj*;
  $polygon_{obj}$ = 2D polygon corresponding to *obj*;
  $max\_depth_{obj} = max(cuboid_{obj}(z))$;
  compute the **union** $union_{other}$ of the other *polygons* for which $max\_depth_{obj} \geq max\_depth_{other}$;
  compute the intersection $occlusion_{obj}$ between $polygon_{obj}$ and $union_{other}$;
  $occluded\_area_{obj} = \frac{area(occlusion_{obj})}{area(polygon_{obj})}$;
  $\gamma_{obj} = sigmoid(occluded\_area_{obj}) \times 0.5$;
end
$\gamma = \frac{\sum_{obj \in objects} \gamma_{obj}}{\#objects} \times 100$;
$scene\_difficulty = \sum_{obj \in objects} difficulty_{obj} \times (1+\gamma)$;

The proposed score has been tested using the generated photorealistic dataset on about 25,000 synthetic scenes, measuring for each subset the mean and the variance of the scene difficulty, as reported in Figure 4.

PHOTOREALISTIC DATASET

The benchmark proposed in this work is also supported by a synthetic dataset. The peculiarity of the dataset is the huge amount of clutter in many of the scenes compared with existing datasets with a similar scope. Furthermore, it provides three different views of the same scene, seen under three different lighting conditions, and seven high dynamic range imaging maps for domain randomization purposes. The dataset is generated synthetically using 3D computer-aided design (CAD) models of the selected objects. The scenes are photorealistic images of objects inside a clear box generated through the Unity 3D engine with the support of Flex, a position-based physical simulation library. Thanks to this peculiar physical simulation, the clutters contain rigid, soft, and deformable objects, and the interaction among the different objects is properly resolved. In addition, by assuming some simplifications, even items filled with liquid, thus having a complex internal dynamic, are considered, resulting in realistic rendering. For each subset, 10,000 scenes have been generated for a total of 50,000 (also considering the full dataset), and for each scene, the dataset provides the RGB and depth image along with the object-oriented bounding box of the involved objects in screen space coordinates and in the world (camera) coordinates as a list of eight points enclosing the objects and the segmentation and normal images. Such information is provided in a YAML file that also contains the transformation (translation and rotation) between the camera and the ground truth pose of the objects. Figure 5 gives some examples.

The synthetic dataset can be useful for training vision algorithms based on deep learning techniques and has been used for designing and testing the goodness of the evaluation metric.

BASELINE

We also provide a baseline for stage 1, the final phase test (clutter), with subset 1. For the experiments, we used a recent multimodal grasp planning framework for hybrid grippers [20]. The approach exploits only geometrical information (3D bounding box). The grasp pipeline leverages on Deep Object Pose Estimation (DOPE) [18] for extracting the bounding boxes and the 6D poses for the objects involved in the scene to compute two-finger grasps and suction grasps. DOPE is a model-based approach that only uses an RGB image as input. First, it estimates the belief maps of 2D key points of all the objects in the image coordinate system and then the 6D pose of each object instance with a standard perspective-n-point (PnP) algorithm on the peaks extracted from these belief maps. The final step uses the detected projected vertices of the bounding box, the camera intrinsic parameters, and the object dimensions to recover the final translation and rotation of the object with respect to the camera. Starting from the 3D bounding box, the upward faces, i.e., the faces that point toward the camera, are computed for each object. Then, the parallel-jaw and suction grasps are synthesized using geometrical computations and are refined and filtered to get more precise and feasible grasps. Finally, a scoring mechanism returns the target object and the best grasp modality, depending on the arrangement of the objects in the scene. Figure 6 depicts

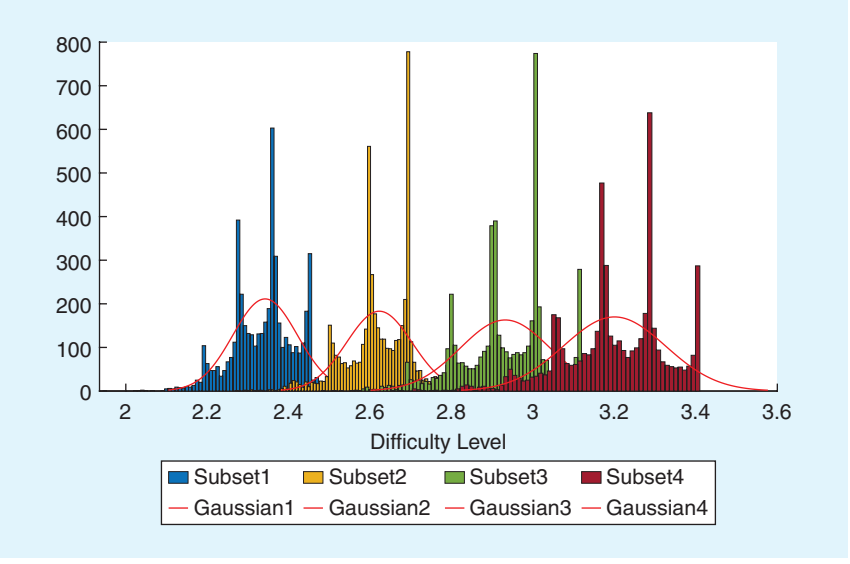


FIGURE 4. The difficulty distributions for the four subsets computed from 5,000 photorealistic generated scenes each.

FIGURE 5. Some information provided for each scene: the scenes seen by the (a) left, (b) middle, and (c) right cameras under directional lighting conditions; the scenes seen by the middle camera as RGB images (d) with spot lighting and (e) with point lighting, also showing the objects' bounding boxes; (f) the depth image; (g) the normal image; and (h) the segmentation mask.

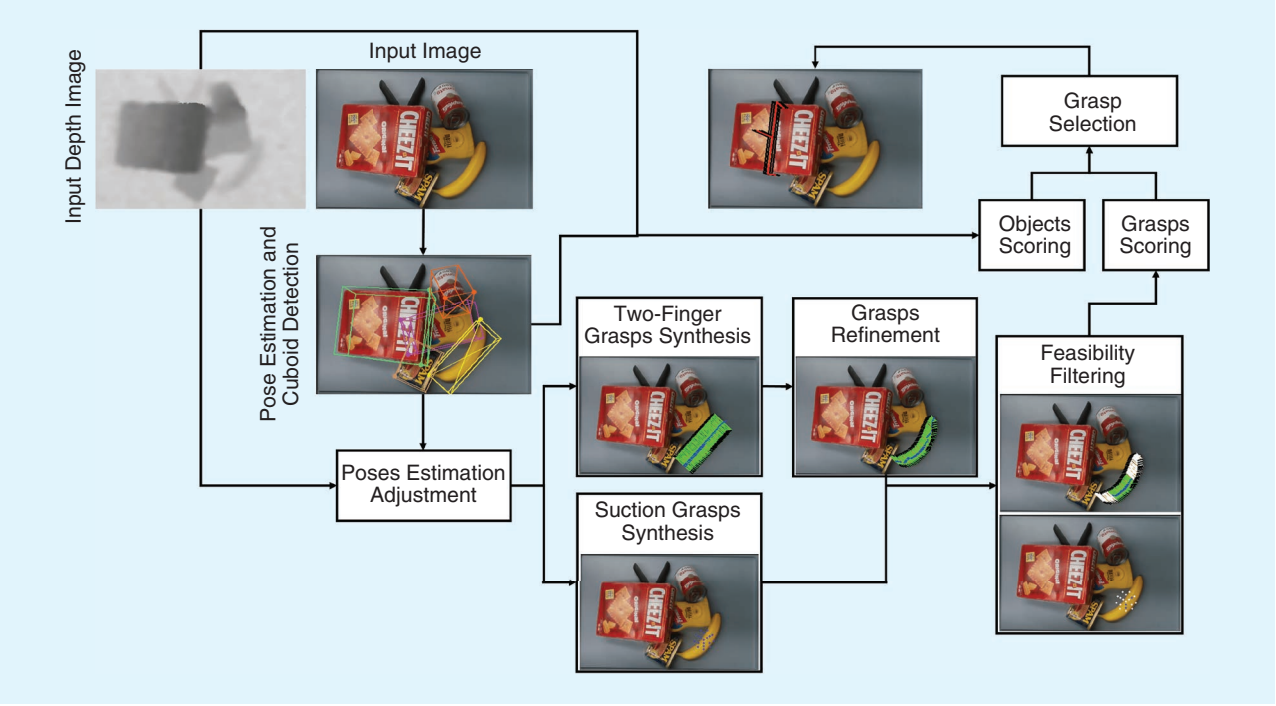


FIGURE 6. A block diagram of the grasping pipeline for two-finger and suction grasp modes. The diagram shows the modularity of the design and flow of information. The input images (RGB and depth) are fed to the DOPE backbone network for pose estimation and a pose estimation adjustment exploiting the depth image is applied. Then, the grasps are computed, and the two-finger grasps are refined using the available CAD models of the object. Grasps are filtered based on some feasibility criteria, and after a scoring stage, the best grasp candidate is selected.

the approach schematically. The experiments have been performed using the DLR Hybrid Compliant Gripper (HCG) (see Figure 7) mounted as an end-effector of a DLR Light Weight Robot (LWR). The HCG has eight degrees of freedom (DoF), with each finger equipped with a suction cup at the fingertip providing three grasping modalities: 1) two-finger grasp, 2) single suction grasp, and 3) double suction grasp. Each finger has three DoF (one-DoF distal interphalangeal and two-DoF metacarpophalangeal joint), and the finger's stiffness can be controlled independently of the position. The fingers are mounted on a base that provides an additional DoF per finger to tilt them away from the palm, enhancing the grasp span up to 260 mm. The maximum object weight for a pinch grasp is about 1.5 kg (for friction coefficients above 0.75) and about 500 g for one suction cup. Figure 7 shows the adopted setup where a Realsense D435 RGB-D camera looking down at the objects has been employed for the vision system.

The average object pose estimation time required by the selected vision module DOPE is 1.1 s. The average time for the grasp planner to return the grasping pose and modality for the next target is 1.8 s. Five, eight, and six objects have been successfully grasped in three trials. The main problems were related to the 6D pose estimation network, which in heavy clutter scenes with the transparent bin does not detect some objects at all or makes the wrong orientation estimation, and to the collisions with the bin walls. The environmental constraints of the bin walls are considered in the space filtering check for the two-finger grasps, which leaves few feasible grasps for final selection. In addition, to a lesser extent, some grasping poses were not kinematically feasible, and the robot was not able to grasp the object after moving in the pregrasping pose, especially in the corners of the box. The approach used for accomplishing the task was always to try to grasp the highest object, which is less occluded. However, the task was never completely finished since at some point, after removing the top objects, the pose estimation network detected the same wrong pose more than ten times, violating the benchmark constraint to proceed further. Figure 8(a) shows the initial scenes of the three trials used in the scoring phase to

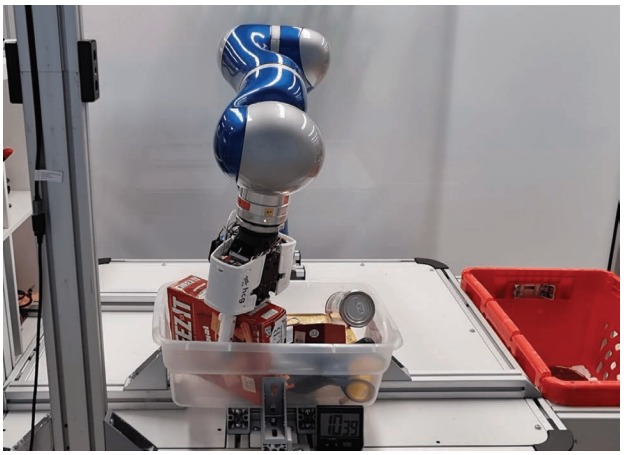


FIGURE 7. The setup used for the experiments.

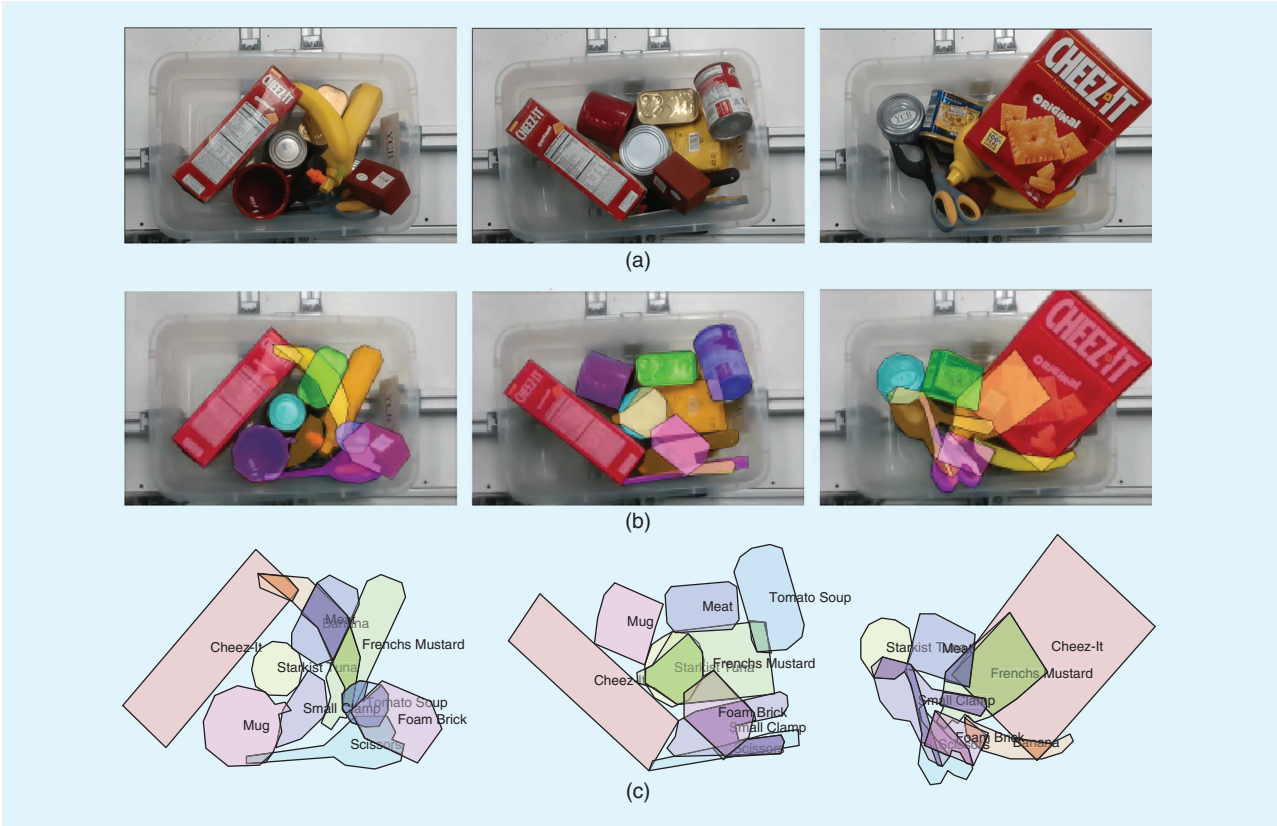


FIGURE 8. (a) The initial images of the three trials. (b) The annotated images and (c) the segmentations obtained by the annotations.

compute the scene difficulty and the clutter percentage. The scores achieved for each trial were 0.95, 1.60, and 1.14, over maxima of 1.80, 1.72, and 1.97, respectively, getting the final score of 1.22 out of 1.83.

For the sake of completeness, Figure 8(b) depicts the images annotated manually for computing the score. It is sufficient to segment each object providing its 2D polygon (line 7 of Algorithm 1) and the corresponding occluded portion taking into account the depth arrangement (line 10 of Algorithm 1). Figure 8(c) shows the data processed by the proposed algorithm. Any annotation tool can be used, but the authors suggest CVAT (https://www.cvat.ai/), exporting the annotation in the coco format.

DISSEMINATION OF RESULTS

The users interested in competing using this benchmark for any of the proposed protocols should register on the website http://cepbbenchmark.eu/ and submit their solution. During the submission phase, the user is required to report other relevant data that are important to more deeply understand the strengths and weaknesses of the system and to have a detailed overview of the proposed solution. They are

- a model of the robot (if it is not a commercial one, a brief description of the main features will be appreciated)
- the position of the robot with respect to the working table
- the typology of the gripper
- the hardware and software details of the vision system (i.e., RGB or depth-only camera, point cloud, or neural network)
- the displacement of the components of the vision system with respect to the manipulator
- the subset number
- the grasping strategy
- the motion planning algorithm
- the grasping synthesis algorithm
- the time spent for the vision and decision-making, planning, and execution (optional if the system competes for the category without time)
- the overall time spent (optional if the system competes for the category without time)
- an image of the clutter at the beginning of the test (only for final stage tests)
- the score.

A video at its normal speed of the entire test case (including the shaking for the clutter tests) is also required as validity checking.

Different leaderboards exist for each phase of each stage, with and without considering the time. When the proof of the video is validated, the results will be uploaded to the leaderboard corresponding to the category for which the participant has applied. The website also reports some visual clues for a better understanding of how to conduct the tests.

CONCLUSION

Evaluating and comparing robotic grasping and manipulation systems among different research groups to share the best practices with the community and learn from errors requires standardized benchmarks. After reviewing the existing

benchmarks, datasets, and competitions published in the last years, the proposed work presents a novel benchmark that evaluates the single components of perception, planning, control, and grasping of the manipulation system and the system as a whole with its protocols for different tasks. In addition, the benchmark exploits a new evaluation metric that takes into account the difficulty of the clutter in the scene depending on the objects' arrangement allowing for a fair comparison without constraining the objects' placement in fixed positions. The task protocols are industrial-oriented and meant to be modular to follow the needs spread out with Industry 4.0 covering flexible automation and generalizable grasping. In addition, a website dedicated to the benchmark contains information on the different tasks, maintains the leaderboards, and serves as a contact point for the community. A baseline approach that exploits geometrical computation for synthesizing grasping points starting from 3D bounding box information provided by a neural network demonstrates the complexity of the benchmark.

Since the robotic field is evolving very fast and is subject to cross-fertilization with other fields of research, it would be overbearing to cover all the possible areas of interest and remain relevant forever. However, the benchmark is open to new opportunities coming from the research community and to add new popular tasks and emerging testing procedures following the baseline proposed in this work.

ACKNOWLEDGMENT

This work was partially supported by the EU Horizon 2020 research and innovation program under Grants 871237 (Sophia), 101017274 (Darko), and 101070136 (Intelliman), by the Italian Ministry of Education and Research in the framework of the FoReLab project (Departments of Excellence), and the Spanish Government through the Project CaRo, PID2020-114819GB-I00.

AUTHORS

Salvatore D'Avella, Department of Excellence in Robotics & AI, Mechanical Intelligence Institute, Scuola Superiore Sant'Anna, 56017 Pisa, Italy. E-mail: salvatore.davella@santannapisa.it.

Matteo Bianchi, Enrico Piaggio Center of Research and the Information Engineering Department, University of Pisa, 56122 Pisa, Italy. E-mail: matteo.bianchi@centropiaggio.unipi.it.

Ashok M. Sundaram, German Aerospace Center, Institute of Robotics and Mechatronics, 82234 Wessling, Germany. E-mail: ashok.meenakshisundaram@dlr.de.

Carlo Alberto Avizzano, Department of Excellence in Robotics & AI, Mechanical Intelligence Institute, Scuola Superiore Sant'Anna, 56017 Pisa, Italy. E-mail: carlo@sssup.it.

Máximo A. Roa, German Aerospace Center, Institute of Robotics and Mechatronics, 82234 Wessling, Germany. E-mail: maximo.roa@dlr.de.

Paolo Tripicchio, Department of Excellence in Robotics & AI, Mechanical Intelligence Institute, Scuola Superiore Sant'Anna, 56017 Pisa, Italy. E-mail: p.tripicchio@santannapisa.it.

REFERENCES

[1] J. Konczak, M. Borutta, H. Topka, and J. Dichgans, "The development of goal-directed reaching in infants: Hand trajectory formation and joint torque control," *Exp. Brain Res.*, vol. 106, no. 1, pp. 156–168, Sep. 1995, doi: 10.1007/BF00241365.

[2] B. Çalli, A. Walsman, A. Singh, S. S. Srinivasa, P. Abbeel, and A. M. Dollar, "Benchmarking in manipulation research: The YCB object and model set and benchmarking protocols," 2015, *arXiv:1502.03143*.

[3] H. Mnyusiwalla et al., "A bin-picking benchmark for systematic evaluation of robotic pick-and-place systems," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 1389–1396, Apr. 2020, doi: 10.1109/LRA.2020.2965076.

[4] Y. Bekiroglu et al., "Benchmarking protocol for grasp planning algorithms," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 315–322, Apr. 2020, doi: 10.1109/LRA.2019.2956411.

[5] P. Sotiropoulos et al., "A benchmarking framework for systematic evaluation of compliant under-actuated soft end effectors in an industrial context," in *Proc. IEEE-RAS 18th Int. Conf. Humanoid Robots (Humanoids)*, 2018, pp. 280–283, doi: 10.1109/HUMANOIDS.2018.8624924.

[6] N. Correll et al., "Analysis and observations from the first Amazon picking challenge," *IEEE Trans. Autom. Sci. Eng.*, vol. 15, no. 1, pp. 172–188, Jan. 2018, doi: 10.1109/TASE.2016.2600527.

[7] J. Leitner et al., "The ACRV picking benchmark: A robotic shelf picking benchmark to foster reproducible research," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2017, pp. 4705–4712, doi: 10.1109/ICRA.2017.7989545.

[8] Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from RGBD images: Learning using a new rectangle representation," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2011, pp. 3304–3311, doi: 10.1109/ICRA.2011.5980145.

[9] G. Kootstra, M. Popović, J. A. Jørgensen, D. Kragic, H. G. Petersen, and N. Krüger, "VisGraB: A benchmark for vision-based grasping," *Paladyn*, vol. 3, no. 2, pp. 54–62, Jun 2012, doi: 10.2478/s13230-012-0020-5.

[10] J. Collins, D. Howard, and J. Leitner, "Quantifying the reality gap in robotic manipulation tasks," in *Proc. Int. Conf. Robot. Automat. (ICRA)*, 2019, pp. 6706–6712, doi: 10.1109/ICRA.2019.8793591.

[11] Y. Huang, M. Bianchi, M. Liarokapis, and Y. Sun, "Recent data sets on object manipulation: A survey," *Big Data*, vol. 4, no. 4, pp. 197–216, Dec. 2016, doi: 10.1089/big.2016.0042.

[12] A. Morgan et al., "Benchmarking cluttered robot pick-and-place manipulation with the box and blocks test," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 454–461, Apr. 2020, doi: 10.1109/LRA.2019.2961053.

[13] Y. Sun, J. Falco, M. A. Roa, and B. Calli, "Research challenges and progress in robotic grasping and manipulation competitions," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 874–881, Apr. 2022, doi: 10.1109/LRA.2021.3129134.

[14] Z. Liu et al., "OCRTOC: A cloud-based competition and benchmark for robotic grasping and manipulation," *IEEE Robot. Autom. Lett.*, vol. 7, no. 1, pp. 486–493, Jan. 2022, doi: 10.1109/LRA.2021.3129136.

[15] C. Piazza, G. Grioli, M. G. Catalano, and A. Bicchi, "A century of robotic hands," in *Proc. Annu. Rev. Contr., Robot., Auton. Syst.*, 2019, vol. 2, pp. 1–32, doi: 10.1146/annurev-control-060117-105003.

[16] T. Hodan, P. Haluza, Š. Obdržálek, J. Matas, M. Lourakis, and X. Zabulis, "T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects," in *Proc. IEEE Winter Conf. Appl. Comput. Vision (WACV)*, 2017, pp. 880–888, doi: 10.1109/WACV.2017.103.

[17] T. Hodan et al., "Bop: Benchmark for 6D object pose estimation," in *Proc. Eur. Conf. Comput. Vision (ECCV)*, 2018, pp. 19–34, doi: 10.1007/978-3-030-01249-6_2.

[18] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, "Deep object pose estimation for semantic robotic grasping of household objects," 2018, *arXiv:1809.10790*.

[19] A. Remus, S. D'Avella, F. Di Felice, P. Tripicchio, and C. A. Avizzano, "i2c-net: Using instance-level neural networks for monocular category-level 6D pose estimation," *IEEE Robot. Autom. Lett.*, vol. 8, no. 3, pp. 1515–1522, Mar. 2023, doi: 10.1109/LRA.2023.3240362.

[20] S. D'Avella, A. Sundaram, W. Friedl, P. Tripicchio, and M. Roa, "Multimodal grasp planner for hybrid grippers in cluttered scenes," *IEEE Robot. Autom. Lett.*, vol. 8, no. 4, pp. 2030–2037, Apr. 2023, doi: 10.1109/LRA.2023.3247221.

Open Access funding provided by 'Scuola Superiore "S.Anna" di Studi Universitari e di Perfezionamento' within the CRUI CARE Agreement