Journal of Intelligent & Robotic Systems (2021) 102: 56
https://doi.org/10.1007/s10846-021-01410-5

SHORT PAPER

A Benchmarking Platform for Learning-Based Grasp Synthesis Methodologies

Jacques Janse van Vuuren[1] · Liqiong Tang[1] · Ibrahim Al-Bahadly[1] · Khalid Mahmood Arif[1]

Received: 9 November 2020 / Accepted: 28 April 2021 / Published online: 3 June 2021

Abstract

Benchmarking is a common practice employed to quantify the performance of various approaches toward the same task. By maintaining a consistent test environment, the inherent behaviours between methods may be distinguished—which is key to progressing a research field. In robotic manipulation research there is a current lack of standardisation, making it challenging to fairly assess and compare various approaches throughout literature. This paper proposes new criteria in conjunction with a benchmarking platform to measure the effectiveness of a grasping pipeline. The proposed benchmarking template offers a testing platform for 2-fingered, vision-based grasp synthesis methodologies. A prototype system was constructed. The prototype was shown to serve as a suitable benchmarking platform for the deployment of various grasp synthesis methodologies. 4000 trials were conducted to evaluate the differing approaches. Results showed that the proposed metrics provide useful insights into the quality of grasp poses produced by a grasp synthesis methodology. Moreover, such metrics provide more comprehensive insights into grasp outcome than traditional methods used to quantify performance of a methodology and present a fair baseline for comparison between different approaches.

Keywords Robots · Robotic learning · Robot control · Robotics and automation · Benchmarking · Performance evaluation

1 Introduction

Reproducible benchmarks, protocols and metrics play a vital role in progressing data-driven research by providing quantifiable insights into the effectiveness of an approach. By objectifying such elements, a fair basis of performance evaluation and comparison may be established, leading the research community to build upon methods that work well. This was particularly evident in computer vision. Large-scale datasets such as Pascal VOC [1, 2], ImageNet [3] and COCO [4] for instance, advanced image recognition considerably by exposing gaps in the state-of-the-art—clarifying the direction of community effort.

Reproducibility is particularly difficult in the field of robotic manipulation, as it is not always possible for researchers to experiment under the same conditions. Many works consequently select for themselves a set of objects, hardware, tasks, or objectives that they consider representative of the desired functionality. Unfortunately, this hinders any direct comparison to other works and makes it difficult to assess relative performance. This issue is well-established, long-standing and stems in-part from variations in experimental protocols, methodological assumptions, research goals and differences in physical hardware, e.g., sensors, lighting, robotic arms, grippers and tested objects [5–7]. Though this is broadly the case, a few subsets of the manipulation community have converged on standardised sets of test objects. Some have also adopted the grasp rate metric, which is determined through physical trials and defined by grasp success.

To accompany standardised object sets and the common grasp success metric, this work proposes a low-cost, self-contained benchmarking platform, experimental protocols and new object-agnostic metrics that quantify the quality of a grasp trial. By maintaining consistent test conditions and employing comprehensive evaluation criteria, a grasping pipeline may be placed with respect to other approaches and the advantages and disadvantages of a specific method may be recognised.

The benchmarking template offered in this paper is industrially focused and aimed at 2-fingered, force-closure grasp

✉ Jacques Janse van Vuuren
j.jansevanvuuren@massey.ac.nz

[1] Department of Mechanical and Electrical Engineering, SF&AT, Massey University, Palmerston North, New Zealand

synthesis methodologies that utilise machine learning (ML). The platform is largely comprised of a conveyor system, vision enclosure and robotic manipulator—Fig. 1. Two RGB cameras are employed, offset by 180°. The conveyor rests on a load-cell subsystem, which is capable of weighing objects and centre of gravity (COG) acquisition. A beam sensor is used to automatically situate assessed objects near the centre of the vision system. Construction of the proposed system is relatively simple and is mainly comprised of extruded aluminium. The robotic manipulator is part of an education package costing approximately $1600 USD. The calibration and integration of various subsystems is also considered in this paper.

New metrics are proposed that quantify the error introduced by a grasp trial using top-down vision. In previous work [8], it was shown that optimising for the proposed metrics reduced orientational error by 2.7%, reduced translational error by 5.2% and improved grasp rates by up to 4.7%—compared to optimising for grasp success alone. The proposed metrics were demonstrated to objectively evaluate methodology performance and showed more descriptive power than the traditional binary measure of grasp success.

Detailed fabrication schematics, CAD models, accompanying software, training datasets and other related documentation are made available online. A prototype benchmarking system was constructed and used to evaluate a 3-stage grasp synthesis methodology in previous papers [8, 9]. The results of four grasping strategies from 4000 trials are published online to help establish baselines for this benchmark. A dataset of 144,000 annotated grasps is also made available to aid in methodology development.

This paper is organised as follows. Section 2 introduces the various approaches toward standardising manipulation research. Section 3 provides an overview of the proposed benchmarking platform. The experimental protocol and grasp quality scores are described in Section 4. Section 5 details an example application of the benchmarking system and the resultant outcome. Finally, the conclusion of this research is presented in Section 6.

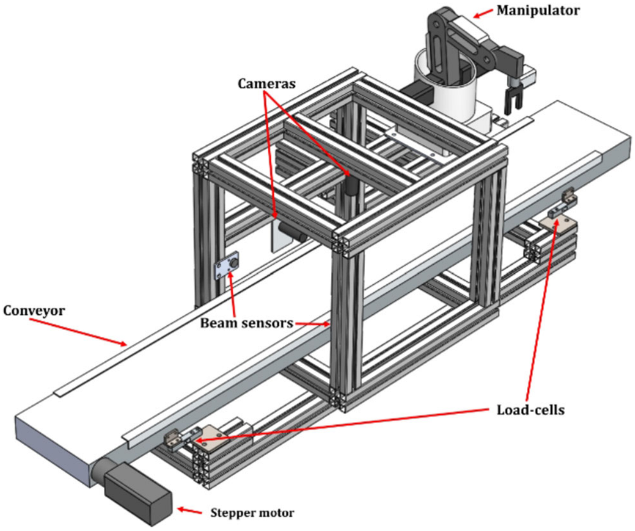

Fig. 1 Annotated isometric view of the proposed benchmarking system with covers removed from the vision enclosure

2 Related Work

In general, data-driven, ML-based grasp synthesis methodologies have become over-represented within robotic manipulation literature. Supervised schemes that leverage large numbers of example grasps are common. To avoid the time-consuming nature of gathering real-world data, many works utilise human-annotate datasets to train CNNs to recognise new grasp instances. Such approaches exploit human expertise and have been shown to be very effective. Jiang, Moseson and Saxena for instance, successfully grasped a range of novel objects 87.9% of the time with a pipeline trained using the Cornell Grasp Detection Dataset [10]. Chu, Xu and Vela achieved a grasp rate of 89.0% when implementing this dataset during training [11]. More complex schemes have also been suggested. Pinto and Gupta for example, employed self-supervision to learn grasp behaviours from 50,000 grasp attempts [12]. Similarly, Levine, Pastor, Krizhevsky and Quillen utilised 14 manipulators over the course of two months to collect over 800,000 grasp attempt samples [13].

In response to the issue of standardisation, a subset of the manipulation community has adopted a fixed set of test objects. Such sets aim to provide a baseline for comparison by trialling an approach on the same set of objects. The YCB object set for example, consists of 75 objects categorised by utility, e.g., food items, kitchen items, tool items, shape items and task items [14, 15]. Leitner at al. proposed the ACRV picking benchmark comprised of 42 widely available objects, in addition to evaluation protocols and suggested arrangements [16, 17]. Despite many efforts, standardised object test sets have not been adopted by the wider community—attributed to methodological differences and hardware constraints. Instead, works often use a *common household* or *common laboratory* object set [18–23]—which varies and is self-defined, but usually well-documented.

Trialling a physical system using a standardised set of objects can be time-consuming and relies on the availability of expensive hardware. Many works have sought to circumvent these requirements by proposing benchmark datasets—enabling rapid assessments of methodologies on real sensor data without any physical equipment. The Cornell Grasp Detection Dataset for instance, has been used extensively for training and testing [11, 24–29]. Some works assess their methodology via dataset performance alone [24, 25, 30–32]. In such cases, dataset accuracy is generally interpreted as a potential grasp rate—without conducting physical trials for validation. Applying this approach to manipulation research relies on the assumption that dataset performance translates into real-world performance. Evidently this is not the case as many works report disparities between real-world grasp rates and dataset accuracy. Sun, Yu, Liu and Gu for instance, noted an 11% disparity between dataset classification accuracy and physical trial outcome [27]. Watson, Hughes and Lida reported a 26% drop in expected performance [33].

Gathering large amounts of real-world data is particularly cumbersome in this field and can be avoided by leveraging simulation. GraspIt! [34–36] is a popular simulation environment commonly used in conjunction with standardised mesh model datasets. Other similar tools include OpenGRASP [37, 38], Gazebo [39–41], openRAVE [42], VisGraB [43] and MoveIt motion planning [44, 45]. Unfortunately, it is not clear whether simulated environments currently resemble the real-world well enough to transfer learned behaviour. Though synthetic data can significantly increase the number of training samples and simulated worlds make research timely, accessible and directly comparable, bridging the so-called *reality gap* has been problematic [46–49]. James et al. for instance, noted their 99% simulated grasp rate drop to 70% when tested in the real-world [50].

Due to the lack of benchmarks and metrics, a common practice has emerged where the success rate of an approach is determined, given a fixed number of trials and objects—where a trial is framed as either successful or unsuccessful, i.e., *pass* or *fail*. This mode of comparison is problematic because what constitutes a successful grasp is not well-established. Several works utilise force sensors when defining a successful grasp outcome [19, 51, 52]. Many works pose task-based definitions, e.g., object transportation to a bin [53–56], complex task descriptions [57, 58], shake protocol [23, 59] and lifting the object above some pre-defined height [10, 12, 29, 60–66]. Though defining a grasp as either a *pass* or *fail* has been pragmatic for comparison, training, and assessment purposes, it is not clear whether this single-faceted criterion is related to grasp quality, handling quality or placement quality. Therefore, this work proposes new metrics in conjunction with a testing protocol and a self-contained benchmarking platform.

3 Benchmarking Platform

Key aspects of design, calibration and other details related to the proposed benchmarking platform are covered in this section. Comprehensive documentation for construction is available at: https://drive.google.com/open?id=1VsEjCl6hrX3FeL9VRF-J9CzVL7JHXO15.

The proposed benchmarking platform illustrated in Fig. 2a is portable and self-contained. The layout consists of a conveyor system, vision enclosure and robotic manipulator, framed by 45 × 45 mm slotted aluminium extrusion. Figure 1 illustrates an annotated diagram of the proposed system.

The 460 × 430 × 420 mm (L × W × H) vision enclosure employs HD webcams, diffuse LED lighting and a photoelectric through-beam sensor. The inside panels were matt white. 2 LED strips were mounted at the top of the enclosure, facing downward. A plastic diffusor was added at the top of the enclosure, below the lighting. Two identical Microsoft Livecam Studio webcams were used for vision—offset by 180°. Specifications related to the components present in the vision enclosure are tabulated in Table 1.

The 1600 × 210 × 60 mm conveyor system was driven by a NEMA 23 stepper motor directly coupled to an aluminium roller, fixed with bearings. The belt was matt white and moved over an aluminium base, with two plastic guard rails to guide objects entering the vision system away from the edges. The conveyor system rested on 4 load-cells used to compute the COG coordinates of an object in vision space. Shielded cabling was used to transmit data from the load-cell amplifiers to the microcontroller. Amplifiers were located directly under each load-cell to minimise the length of wiring, thereby reducing the amount of voltage induced by the AC power system. This resulted in significant noise reduction. The cabling used between amplifiers and load-cells were also shielded. Table 2 provides information regarding the specific componentry used by the conveyor system.

The robotic manipulator used in this benchmark system is part of the Dobot Magician pedagogical package [73]. The system is relatively low-cost and provided ample features. The Dobot Magician was rated to carry a maximum payload of 500 g, while maintain a position repeatability of 0.2 mm up to 320 mm from the base. Several end-effectors were included in this package. A programmable I/O interface was present on the forearm and base of the robot to encourage the development of custom end-effectors and peripheral hardware. A comprehensive specification list is tabulated in Table 3.

Power, communication, and control subsystems were housed in a control box mounted to the apparatus. AC power was converted to appropriate DC voltages. 24 V was used to power the stepper motor driver and LEDs. 5 V was used to power the microcontroller. Details regarding the componentry within the control box are provided in Table 4.

The complete system occupies a cuboid of approximately 1600 × 700 × 470 mm. Dimensions and workspace are illustrated in Fig. 2b. The total cost of a single benchmarking platform proposed in this paper is roughly $2300 USD. The robotic manipulator accommodates gripping locations of up to approximately 28 mm and objects weighting up to 500 g. During development and experimentation, the manipulator performed more than 10,000 grasp actions. Therefore, the conveyor completed over 20,000 translation actions. No breakages were encountered, and no significant wear is visible. No parts were replaced during this period, no screws were tightened, and no cables were frayed.

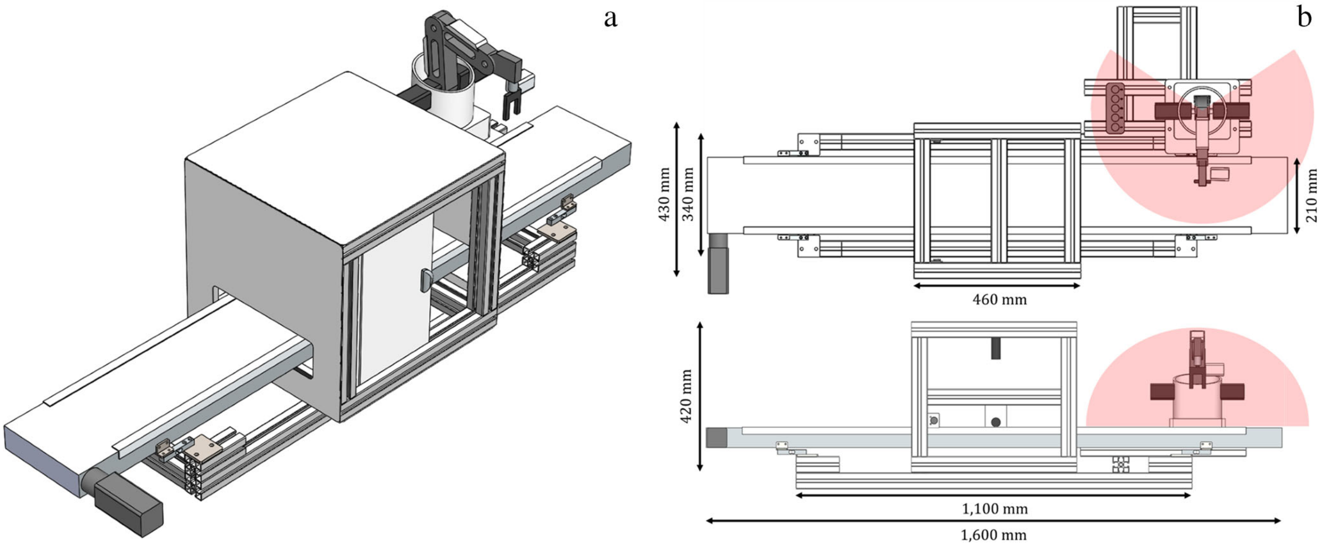

Fig. 2 **a** isometric view of the proposed benchmarking platform. **b** top and side views of the model, with measurements and approximate robot workspace annotated

4 Benchmarking Protocol

4.1 Grasp Quality Scores

Many works throughout literature quantify the performance of their method in terms of grasp success rate. To broaden the scope of method assessment, this work proposes to measure grasp-induced error through vision from a single top-down perspective. First, an image of the object is captured. The object is then picked, lifted, and subsequently placed at the same location. By solely supporting the object with the gripper, weight is taken off the surface to remove any friction which may interfere with the measurement. A final image is then captured of the resultant. By comparing these two images, error introduced by the gripping process itself can be observed.

Two criteria referred to as *similarity metrics* are computed: *OS* and *OE*. The overlap score *OS* is described by the Jaccard similarity index—introduced by Paul Jaccard in 1901 [78]. This index is commonly used for object detection tasks that employ ML [79, 80]. *OS* is computed as the intersection over union (IoU):

$$OS = \frac{|A_{pre} \cap A_{post}|}{|A_{pre}| + |A_{post}| - |A_{pre} \cap A_{post}|} \tag{1}$$

where A_{pre} is the top-down area of the object prior to the grasp. A_{post} is the area of the object after the grasp. Figure 3 illustrates the area of intersection between pre-grasp and post-grasp images used to calculate *OS* in Eq. 1. This criterion measures overlap, strongly responding to translation error—though rotational misalignment is also captured. *OS* ranges from [0, 1]. As the area of overlap tends toward 0%, this score also tends toward 0. Conversely, *OS* will register a value of 1 in the case of a perfect grasp—where no difference between images is measured. The orientation error score *OE* compares the rotational change of the object:

$$OE = 1 - \frac{|(\theta_{post} - \theta_{pre})|}{180} \tag{2}$$

Table 1 Details of componentry utilised in the vision enclosure subsystem

Component	Details
Cameras	Microsoft Livecam Studio [67] • Sensor resolution: 1920 × 1080 • 30 FPS • Manual override on image adjustments
LED strips	RS PRO 136–3579 white LED strips [68] • Colour temperature: 6500 K
Light diffusor	Frosted acrylic sheet
Beam sensor	PA18C [69] • Response time: ≤ 1.0 ms • Operating frequency: 500 Hz • Range: 1 m

Table 2 Details of componentry utilised in the conveyor subsystem

Component	Details
Load-cells	TAL220 Parallel beam load-cell [70] • Capacity: 10 kg • Rated error: ± 0.05%
ADC amplifier	HX711 ADC [71] • Operating frequency (max): 80 Hz • 24-bit, serial output
Stepper motor	NEMA 23 SY57STH76-2804A [72] • Holding torque: 19 kg-cm • 200 steps per revolution

Table 3 Dobot Magician operational specifications. Source: Dobot Magician User Guide V1.5.1 [74]

Component	Details
Robot model	Dobot Magician [73]
Max payload	500 g
Max reach	320 mm
Number axes	4
Repeatability	0.2 mm
End-effectors	3D printing kit Laser engraver Pen holder Vacuum suction cup 2-fingered pneumatic gripper • Range: 0–27.5 mm • Maximum force: 8 N • Payload: 500 g
Software	Dobot Studio Dobot Blocky Dobot programming library
Exensions	10 × I/O interfaces 4 × Controllable 12 V power output UART communication interface
Communications	USB / Wi-Fi / Bluetooth

where θ_{pre} is the major orientation of the object pre-grasp and θ_{post} is the major orientation of the object post-grasp. θ_{pre} and θ_{post} are measured between the horizontal axis of the image and the major axis of the object before and after a grasp attempt, respectively, and range from [0°, 180°]. This metric scores the orientational change of an object introduced by the executed pick-and-place action. It should be noted that it does not respond to translational error—as illustrated in Fig. 4. *OE* ranges from [0, 1]. The relationship between this score and the angular difference between pre- and post-images is linear. As the angular change tends toward 180°, *OE* tends linearly toward 0. If no orientation change is measured, *OE* produces a value of 1.

Table 4 Details of componentry within the control box

Component	Details
AC-DC PSU (24 V)	Mean Well MDR-60-24 [75] • Input: 100–240 VAC, 1.8A • Output: 24 V, 2.5 A
AC-DC PSU (5 V)	Chinfa AMR1–05 [76] • Input: 90–264 VAC • Output: 5 V, 1.5 A
Stepper motor driver	Leadshine M542 [77] • Input: 20–50 VDC • Output: 4.2 A per phase
Microcontroller	Arduino Uno REV3 • Input: 5–20 VDC • Digital I/O: 14 • Analog I/O: 6 • Clock speed: 16 MHz

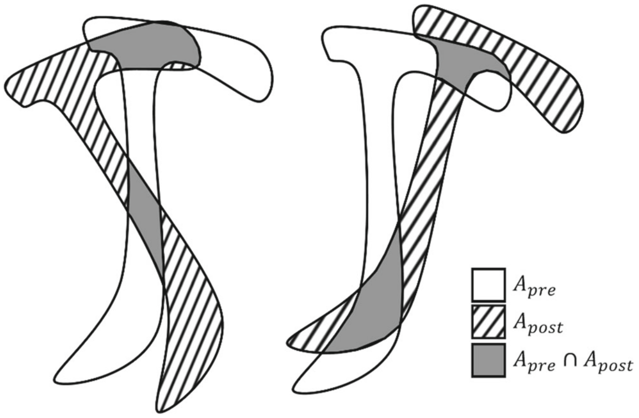

Fig. 3 Illustration of pre-grasp object area ($\boldsymbol{A_{pre}}$), post-grasp object area ($\boldsymbol{A_{post}}$) and the area of intersection between pre-grasp and post-grasp images ($\boldsymbol{A_{pre}} \cap \boldsymbol{A_{post}}$). Two examples of grasp outcome are shown

The proposed metrics may be of interest to other similar works for improvement and as an objective baseline for comparison between methodologies—since they are not dependent on the system or object test pool. *OS* and *OE* measurement is relatively simple, requiring only a single top-down RGB camera and basic DIP processing. Since *OS* and *OE* are continuous, they can be used to train regression models and assess grasp quality. Alternatively, classification models can be trained via class-based thresholds.

4.2 Experimental Protocol

To objectively trial a methodology deployed on the discussed apparatus, single objects are placed haphazardly on the conveyor belt furthest from the robotic manipulator. Detected objects are translated to I_t x-axis centre utilising a beam sensor. Once stationary, objects are analysed by the employed grasp synthesis methodology. After finalising a grasp pose, the object is translated into robot space—where it is subjected to a manipulation procedure. Depending on the outcome, the trial may be labelled as either *pass* or *fail*. Post-manipulation,

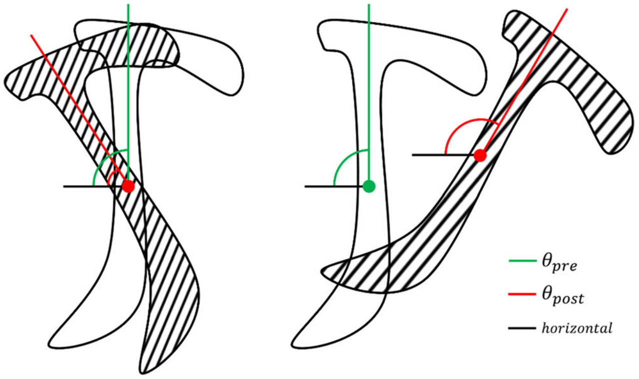

Fig. 4 Illustration of pre-grasp object angle ($\boldsymbol{\theta_{pre}}$) and post-grasp object angle ($\boldsymbol{\theta_{post}}$), relative to the horizontal axis of the image (***horizontal***). Two examples of grasp outcome are shown

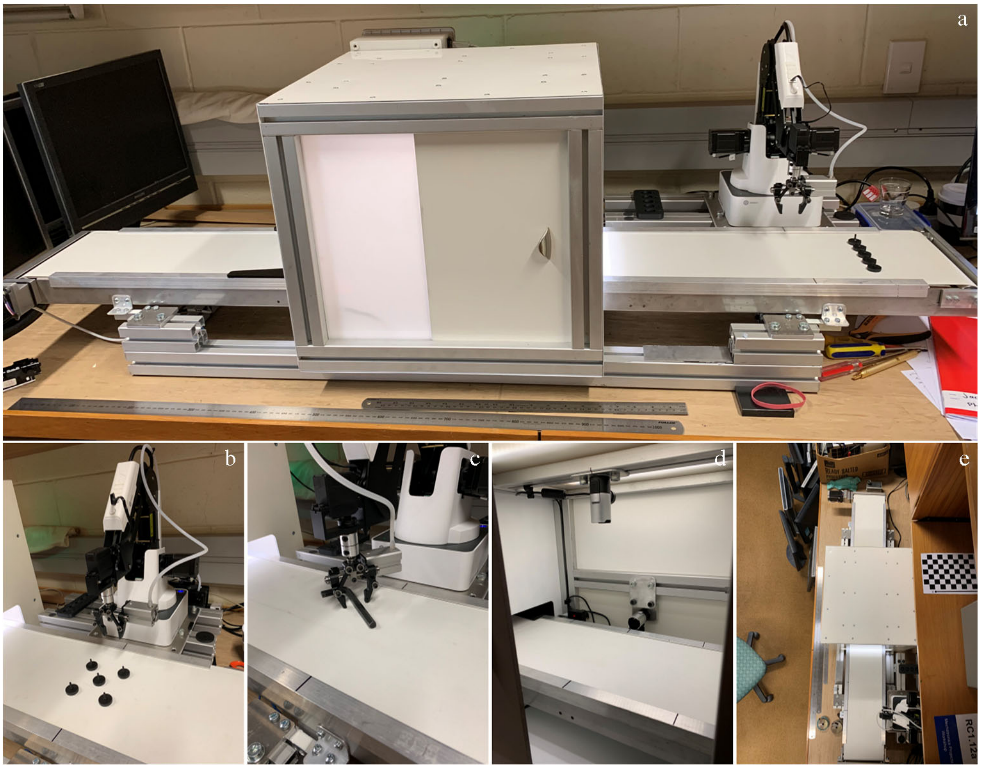

Fig. 5 Various prototype-related images. **a** frontal view of the complete apparatus. **b** Dobot Magician manipulator stacking calibration discs. **c** Dobot Magician picking an object. **d** internal view showing cameras and beam sensor. **e** top down view of the apparatus

the object is translated back into vision space to assess the quality of the performed grasp in terms of *OS* and *OE*. The above described process constitutes one trial.

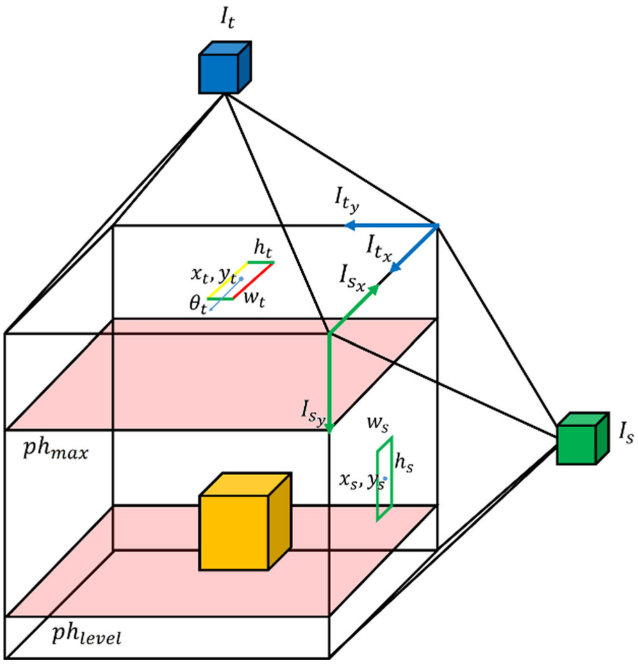

Fig. 6 Diagram illustrating the conceptual basis and relative positions of the top-view grasping rectangle representation and the side-view grasping rectangle representation

In the manipulation procedure, autonomously computed grasp poses are physically attempted by the robotic manipulator. The object is grasped and lifted vertically to a height of 15 cm. The object is then suspended for 10 s, after which it is placed back on the conveyor platform at the initial grasp location. For a grasp trial to be labelled as *pass* or 1, the object must not fall at any stage of the trial. Moreover, objects must be solely supported by the gripper—such that no other part of the object touches any other hardware. Any trial which does not conform to the outlined criteria is labelled as *fail*, or 0. The derivation of a *pass* label definition stems from many works throughout literature. Saxena et al. for instance [65, 66], consider a trial successful if an object can be lifted to a height of 1 ft and held stationary for 30 s. Similarly, Pinto and Gupta [12], Johns, Leutenegger and Davison [63] and Kopicki et al. [64] consider the same standard, but substitute a height of 20 cm. Many other works also consider a grasp attempt as successful if the trialled object is lifted above some predefined height for some time interval [29, 56, 60, 61].

Table 5 Error associated with the conveyor system, found from 50 measurements

Component	Associated error
Conveyor belt translational error (pix)	± 3.4 pixels
Conveyor belt translational error (mm)	± 0.5 mm

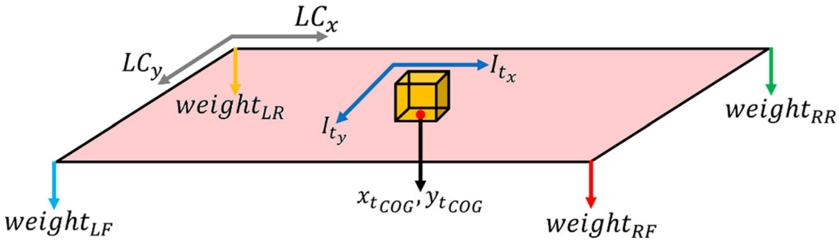

Fig. 7 Graphical illustration of the load-cell arrangement and related coordinate frames

5 Example Application

5.1 Prototype System

An experimental prototype was constructed to physically validate the proposed benchmarking platform. A grasp synthesis methodology presented in previous works [8, 9] was deployed and trialled using the test system. Various images of the prototype system are illustrated in Fig. 5.

The trialled synthesis methodology employed a 3-stage, CNN-based structure to generate numerous candidate grasps—of which a final grasp was selected for implementation. A 5-dimenstional, rectangular representation was used to characterise a grasp pose within top-view camera space I_t:

$$pose_t = \{x_t, y_t, \theta_t, h_t, w_t\} \in I_t \tag{3}$$

where x_t, y_t refer to the gripper's centre position in Cartesian coordinates within I_t. Rotation of the rectangle is defined by angle θ_t about centre point x_t, y_t, with respect to the horizontal axis I_{tx}. Physically, the 2 fingers of the gripper close perpendicular to θ_t. h_t relates to the physical width of the gripper in pixels and w_t refers to the available distance between the two parallel plates of the gripper at maximum extension, also in pixels. Note that h_t and w_t are fixed, as the dimensions of the gripper from perspective I_t do not change. In addition, side-view pose component of a grasp, $pose_s$, is defined as:

$$pose_s = \{x_s, y_s, h_s, w_s\} \in I_s \tag{4}$$

where x_s, y_s refer to the centre position of the ideal contact area of the gripper pads, within side-view vision space I_s. h_s and w_s are the height and width of this contact area in pixels, respectively. Figure 6 graphically depicts the relative grasping poses.

5.2 Calibration, Coordinate Frame Consolidation and Design Considerations

The conveyor belt is actuated in one axis only. The x-axis of the top camera and conveyor coordinate frame act in the same plane—graphically depicted in Fig. 7a. The conveyor coordinate system is expressed in terms of motor steps. One step was found to correspond to approximately 0.096 mm, or 0.64 pixels. The only function of the conveyor is to move objects in and out of coordinate frames. Initially, objects placed on the conveyor belt are translated to an approximate x-axis mid position within top-view image space by utilising the beam sensor. At this stage, objects are discoverable through vision and grasp configurations may be computed by the deployed grasp synthesis methodology. Top-view image space I_t and

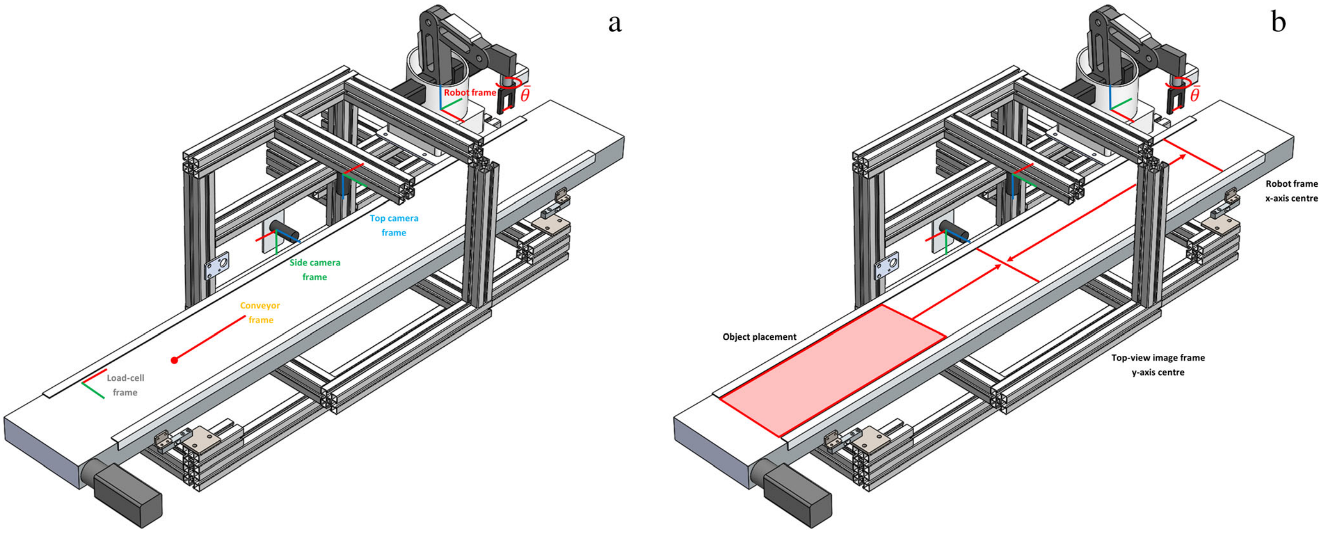

Fig. 8 **a** visualisation of various coordinate frames associated with the proposed benchmark system. **b** illustration of the function of the conveyor system. Object are placed haphazardly at one end of the prototype, where they may be moved into vicinity of the desired coordinate frame by the conveyor. Red, green, and blue coordinate lines express the x, y, and z-axes, respectively

Table 6 Error and calibration tool weights associated with the load-cell-conveyor subsystem

Component	Associated error
Combined load-cell error	± 0.10 g
Calibration tool weight	1079.60 g ± 0.01 g
Calibration disc weight	133.63 g ± 0.01 g

robot workspace $\overline{G}$ may be treated as though they occupy the same space by translating the object a set amount between frames. Effectively, conveyor space is ignored and treated as an intermediary consolidation step prior to an interaction. Figure 8b illustrates the function of the conveyor system. Note that the intermediate conveyor translate step introduces a fixed error, shown in Table 5.

Each load-cell unit was calibrated individually. Note that a load-cell unit consists of an HX711 amplifier and TAL220 load-cell pairing. Calibration was specific to each unit and varied slightly between units. 50 initial measurements were taken with no weight present to tare the unit. A calibration tool was bolted to the load-cell and used for calibration. The calibration tool and bolts weighed 1079.60 g. Measurement accuracy was found to be consistent within ± 0.1 g of the calibration weight for individual load-cell units. A 133.63 g calibration disc was used to measure the combined load-cell error. Errors associated with this subsystem are tabulated in Table 6.

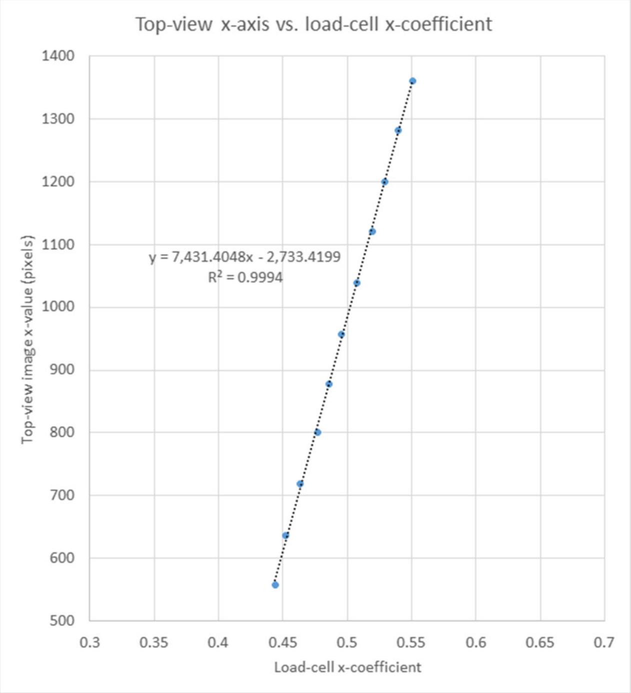

Fig. 9 Relationship between top-view image x-axis and load-cell x-coefficient

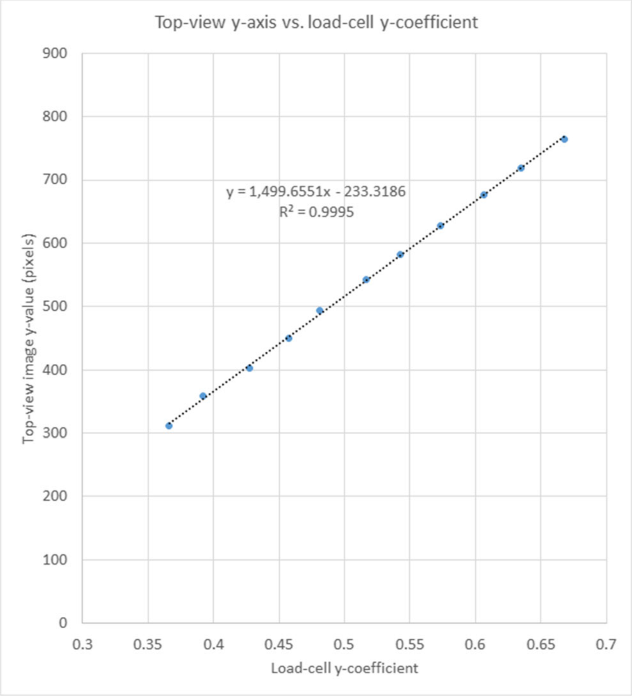

Fig. 10 Relationship between top-view image y-axis and load-cell y-coefficient

Because an object is supported by a fixed number of load-cells—and the total weight is known—the ratio of weight distribution between axes can be calculated. This ratio may be used in conjunction with known distances between load-cells to compute a relative COG position within vision. The load-cell coordinate frame *LC* is quantified by 4 sensitive components—the conveyor platform rests on these components. Load-cell space *LC* and the top-view camera frame I_t share two axes—Fig. 8a and Fig. 8.

To formalise load-cell coordinate space in terms of vision x- and y-axes, the output of each component is measured. Ratio coefficients are used to estimate the COG position of an object $[x_{t_{COG}}, y_{t_{COG}}]$ in image space I_t directly:

$$x_{t_{COG}} = acoeffx{-}b \tag{5}$$

$$y_{t_{COG}} = ccoeffy{-}d \tag{6}$$

where *a*, *b*, *c* and *d* are found through experimentation. Ratio coefficients are defined by:

$$coeffx = \frac{weight_{RF} + weight_{RR}}{weight_{total}} \tag{7}$$

$$coeffy = \frac{weight_{RF} + weight_{LF}}{weight_{total}} \tag{8}$$

where $weight_{total}$ is computed as the sum of all weight measurement inputs. A 133.63 g calibration disc was used to

Table 7 Error associated with the COG position estimate, taken from 20 measurements using the 133.63 g calibration disc

Component	Associated error
COG position error @ 133.63 g (pixels)	± 4.1 pixels
COG position error @ 133.63 g (mm)	± 0.6 mm

measure the response of *coeffx* and *coeffy* as the disc was iterate across $I_t[x, y]$. Disc centre location was found through vision. The resulting relationships between $I_t[x]$ and *coeffx* and $I_t[y]$ and *coeffy* are graphed in Figs. 9 and 10, respectively.

The error associated with the COG measurement subsystem is shown in Table 7.

Due to inaccuracies in construction, the conveyor platform surface was not perfectly perpendicular with the robotic manipulator z-plane. This offset is illustrated in Fig. 10.

To compensate for this error, bed level was modelled in terms of robot z-axis values. This was achieved by sampling end-effector z-axis positions of the conveyor surface along the x-axis with y-axis coordinates set to 0. A second-degree polynomial relationship was found to relate the true conveyor bed height with the robot x-axis—illustrated in Figs. 11 and 12. This model was used to mitigate the slight error between the robot z-plane and conveyor level surface, improving manipulation accuracy.

5 calibration discs with rubber bottoms were used to tune the vision system, find threshold values, and consolidate coordinate systems—shown in Fig. 5a. Discs were matt black with 20 mm diameters. An approximate pixel to mm conversion was found by calculating the disc area in mm:

$$A_{discmm} = \frac{\pi D^2}{4} \tag{9}$$

where A_{discmm} is the calculated area of a calibration disc from a top-down perspective in squared millimetres. A conversion may then be established through relationship:

$$mm\ per\ pixel = \sqrt{\frac{A_{discmm}}{A_{discpix}}} \tag{10}$$

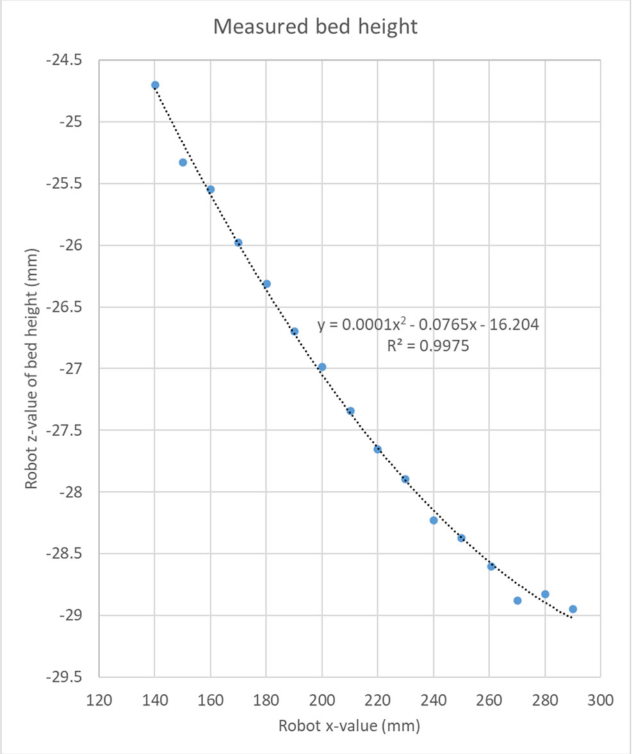

Fig. 12 Measured conveyor level height in terms of robot z-value along the robot x-axis. This relationship was used to mitigate misalignment errors due to inaccuracies in construction

where $A_{discpix}$ is the area of the disc measured by the vision system, in pixels. This calculation provides an approximate conversion from mm to pixel. Note that this conversion is only an estimate and assumes a linear relationship between pixel and mm. Due to minor lens distortion, this value may vary slightly between x-axis, y-axis, and distance from the camera. 5 calibration discs were used to consolidate the $\overline{G}[\overline{x}, \overline{y}]$ components of the robot coordinate frame with the top-view camera frame $I_t[x, y]$. The calibration pattern is shown in Fig. 13.

Disc centre positions were found through vision. Figs. 14 and 15 graphically illustrate the relationship between $I_t[x]$ and $\overline{G}[\overline{y}]$ and $I_t[y]$ and $\overline{G}[\overline{x}]$, respectively.

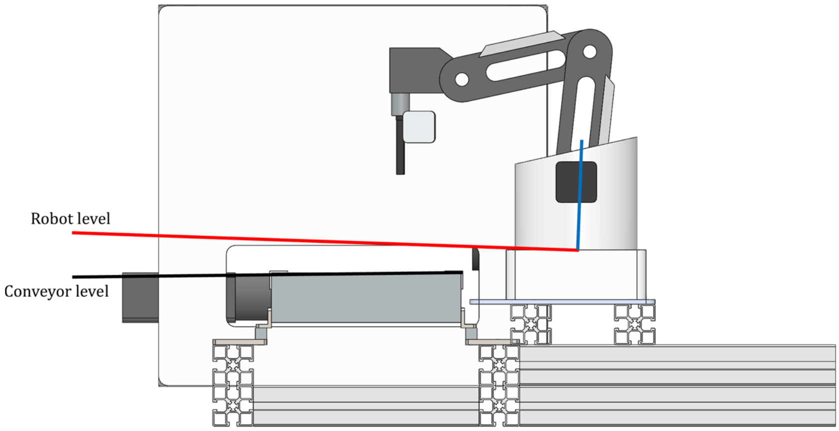

Fig. 11 Illustration of misalignment due to inaccuracies in construction. Note this offset has been exaggerated for clarity

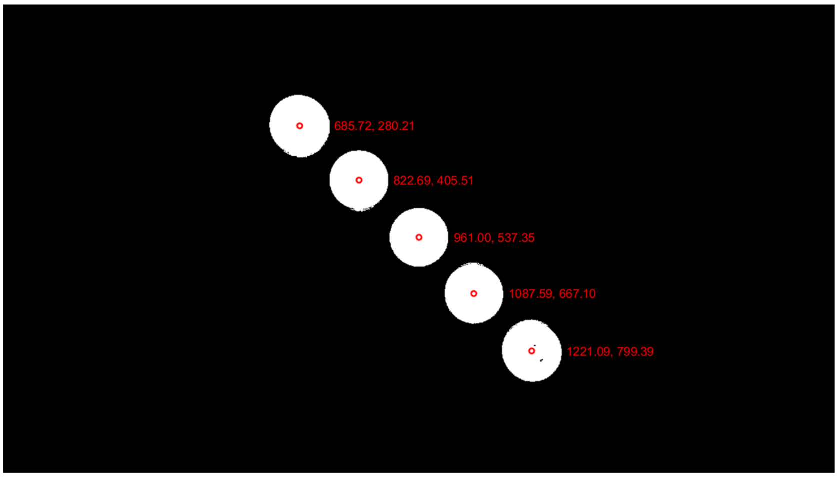

Fig. 13 Calibration pattern used to consolidate robot and vision coordinate frames. Vision coordinates are annotated

Linear models were fit to the resulting data. These relationships are used to estimate robot coordinates $\overline{G}[\overline{x}\ ,\overline{y}]$, given top-view camera coordinates $I_t[x, y]$. Therefore, grasps generated in vision space by a grasp synthesis methodology may be converted to robot space for implementation via the following transform:

$$\overline{x} = a\, I_t[y] + b \tag{11}$$

$$\overline{y} = cI_t[x] - d \tag{12}$$

$$\overline{\theta} = \theta_t + e \tag{13}$$

where a, b, c, d and e are found through experimentation. Robot end-effector angle $\overline{\theta}$ and top-view grasping rectangle angle θ_t are related by a fixed offset. No significant rotational offset was found between coordinate frames. As such, simple linear relationships could be applied accurately, as opposed to computationally expensive rotation matrices that account for rotational offsets between coordinate systems. The error associated with the transformation between vision and robot frames is shown in Table 8.

Photoelectric diffuse-reflective sensors situated at the entrance to the vision enclosure were used to translate objects to I_t x-axis centre. As the assessed object is stepped toward the

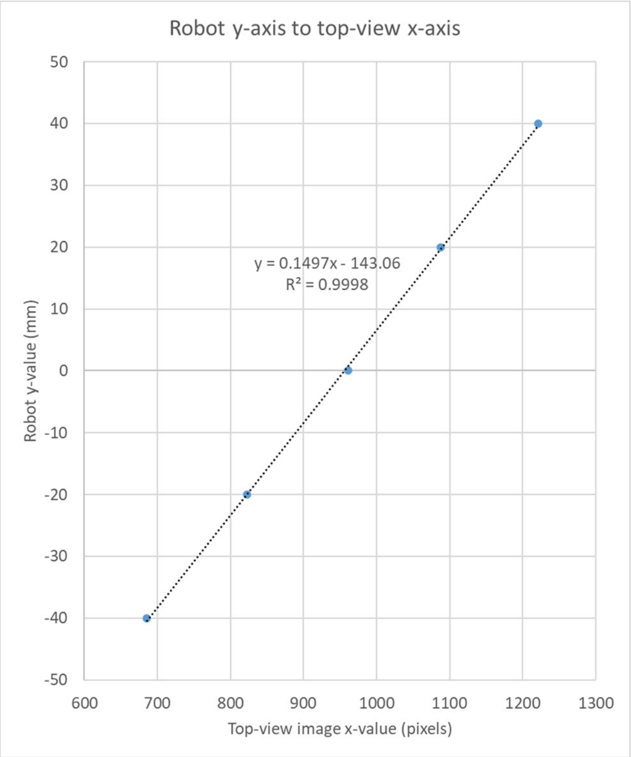

Fig. 14 Relationship between robot y-axis and top-view vision x-axis

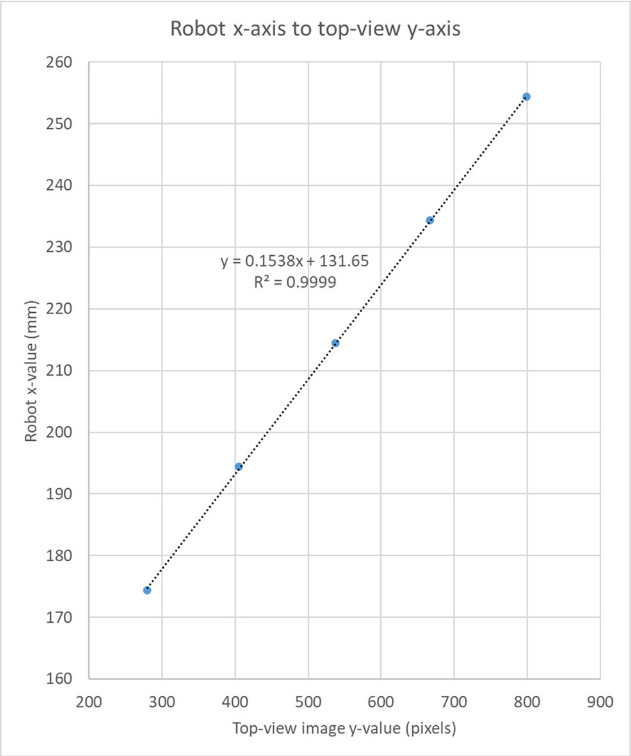

Fig. 15 Relationship between robot x-axis and top-view vision y-axis

Table 8 Error associated with the vision-robot transform, found from 50 measurements

Component	Associated error
Robot-vision transformation error (pix)	± 1.70 pixels
Robot-vision transformation error (mm)	± 0.24 mm

vision system, the beam is interrupted. This initial object-beam incidence step location is denoted as N_i. As the object continues past the beam sensor, the beam can resume uninterrupted. This change in signal is transmitted to the microcontroller, which notes the associated step position. The step location at which the beam resumes is denoted as N_s. By recording the step count at N_i and N_s, the length of the object may be approximated in steps as $N_s - N_i$. The distance from the end of the beam sensor to I_t x-axis centre N_{mid} is known a priori. Therefore, the steps required to translate the object to I_t centre is computed as:

$$N_{step} = N_{mid} - \frac{N_s - N_i}{2} \tag{14}$$

I_t centre is denoted as $I_t\left[\frac{W}{2},\frac{H}{2}\right]$, where W and H are the width and height of image space, respectively. Note that the conveyor only translates objects within the I_t x-axis. Object length approximated in the above manner does not necessarily relate to a true dimension of the object. This length simply refers to the number of beam steps interrupted by the object. Consequently, object orientation does not affect the accuracy of centring objects near $I_t\left[\frac{W}{2}\right]$. This process is graphically represented in Fig. 16.

Due to noise present in the vision system, conveyor, and robotic manipulator, OS and OE assessments were prone to minor errors in measurement. Table 9 lists the amount of error associated with each respective component.

5.3 Data Collection

The grasping pipeline employed to evaluate the proposed benchmarking platform used a CNN to detect grasps through vision. The dataset used to train this network was composed of 144,000 hand-annotated samples from 15 unique objects. 47,000 samples related to a positive class—defined to reflect grasping areas that likely facilitate successful grasps when implemented. The remaining samples pertain to areas that may result in unsuccessful grasps when implemented. 70% of this dataset was used for training/validation and 30% was used for testing. The resulting network correctly classified 99.5% of the training set and 99.6% of the validation set. Time to classify was 0.31 ms using an NVIDIA Quadro K2200 GPU.

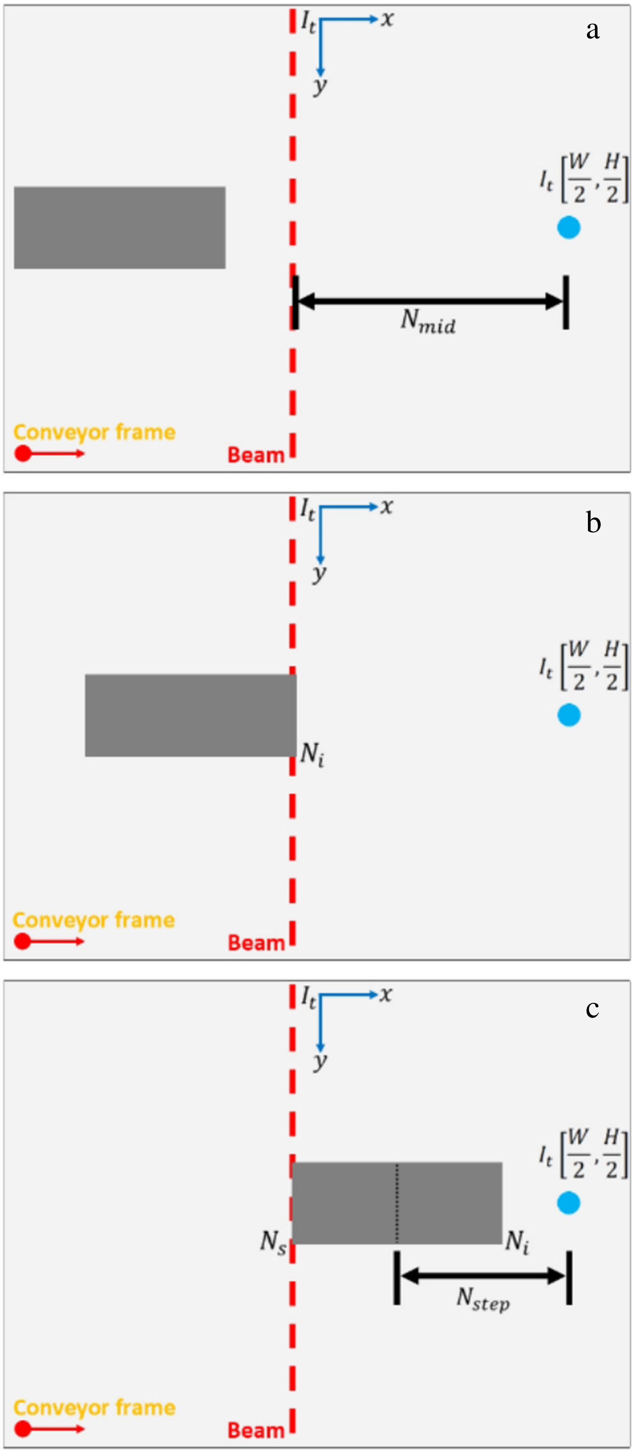

Fig. 16 Illustration of the beam sensor process used to measure object length. **a** depiction of object heading toward the vision frame. **b** depiction of initial object-beam incidence. **c** depiction of beam resuming as object moves past sensor. By recording the step locations of change in beam signal, object length may be estimated

The selection components of the employed pipeline were trained using real data from actual grasps—recorded using the prototype benchmark system. 2000 grasp trials were conducted by randomly selecting a sample generated by the grasp

Table 9 Error associated with ***OS*** and ***OE***, found from 20 measurements per type of error for a total of 120 samples

	Vision	Vision + conveyor	Vision + conveyor + robot
OS	± 0.0012	± 0.0204	± 0.0545
OE	± 0.0001	± 0.0007	± 0.0033

detection CNN. Note that many candidate grasp poses were generated by the detection network, but only one candidate was attempted per object instance. A trial consisted of attempting one grasp pose per object and recording outcome in terms of *pass/fail* label, *OS* and *OE*—as described in Section 4.B. This dataset was used to train various regression models to predict the outcome of a grasp, given various grasp-related scores. The predicted value was used to select a grasp for implementation from the pool of candidates. Refer to our most recent publication for more information regarding the training methodology, datasets, network architectures, and details related to the object test pool used in this study [8].

5.4 Testing and Results

To evaluate the proposed benchmarking system and associated grasp quality scores, four separate grasping approaches were trialled. Table 10 details the various criteria employed during experimentation.

Each approach was trialled 10 times per 100 unique objects, for a total of 1000 physical grasp attempts per criterion. 4000 trials were conducted in total. 15 objects were used to generate training data, while the remaining 85 objects were used for testing purposes only. The object pool consisted of objects such a screwdriver, a pen, several nuts and bolts, a VGA connector, several small toys, several pneumatic components, a hex key, a USB pen drive, and many small components [8]. This pool was split into three series, labelled known, unknown, and miscellaneous. In addition, objects were subcategories according to utility: common household objects, tools, and components. The known set consisted of the 15 objects seen during training. The unknown set contained 45 objects and the miscellaneous set contained the remaining 40.

The results of the 4000 trials are illustrated in Fig. 17. Randomly attempting a grasp classified by the detection network resulted in an aggregate grasp rate of 94.0%. Grasping based on the softmax output of this network improved grasp rates to 96.4%. Implementing a grasp based on a network trained to optimise for grasp success, resulted in a grasp rate of 96.3%. Finally, selecting for the proposed metrics, i.e., highest combined OS_{pred}, OE_{pred}, improved grasp rates to 99.0%. From the trials, it was clear that the quality of a grasp in terms of resultant *OS* and *OE* varied by grasping approach. The average *OS* and *OE* scores from the 1000 trials by the random approach were 0.53 and 0.94, respectively. Highest K_{sf} improved *OS* and *OE* performance to 0.56 and 0.96, respectively. $pass_{pred}$ improved the resultant *OS* to 0.57 but yielded a lower average *OE* score of 0.95. Grasping based on the proposed similarity scores by predicting OS_{pred} and OE_{pred} resulted in an average *OS* score of 0.63 and *OE* score of 0.98. Fig. 17b illustrates the resultant *OS* scores by selection criteria and object series. For the resultant *OE*, refer to Fig. 17c.

Analysis of the 4000 trials revealed a potential relationship between the physically measured similarity scores *OS*, *OE* and the *pass/fail* classification of grasp outcome. The average *OS* and *OE* scores for a *pass* outcome were 0.59 and 0.97, respectively. The average *OS* and *OE* scores for a *fail* outcome were 0.10 and 0.69, respectively. Refer to Figs. 18 and 19 for more details.

Some potential system limitations were observed during testing. The orientation score *OE* is computed by comparing the pre-grasp angle θ_{pre}, with the post-grasp angle θ_{post}. The derivation of these angles is based on the major axis of an ellipse fitted to the object. Circular objects do not have a major orientation. Therefore, *OE* cannot be computed for such objects. In practice, the sensitivity of the vision system captured and responded to irregularities within seemingly circular objects. Although this issue was not found relevant in this study, it may be problematic in some other cases.

Table 10 Description of the four grasping approaches employed during trials

Approach	Description
Random	The grasp to be implemented *G* is randomly selected from candidate grasp matrix *g*—generated by grasp detection network.
Highest $K_{I_{RGW}}$	The highest $K_{I_{RGW}}$ scoring sample from candidate grasp matrix *g* is selected for implementation. $K_{I_{RGW}}$ denotes the softmax function output of the grasp detection network for the positive class.
Highest $pass_{pred}$	Implemented grasps *G* are selected from grasp matrix *g* by a network trained to predict the probability of a *pass* label based on the 10 input scores. The candidate with the highest probability of resulting in a *pass* label is selected for implementation.
Highest combined OS_{pred}, OE_{pred}	A grasp *G* is selected from matrix *g* by predicting the resultant *OS* and *OE* scores given 10 inputs. The candidate with the highest combined predicted scores, OS_{pred} and OE_{pred}, is selected for implementation.

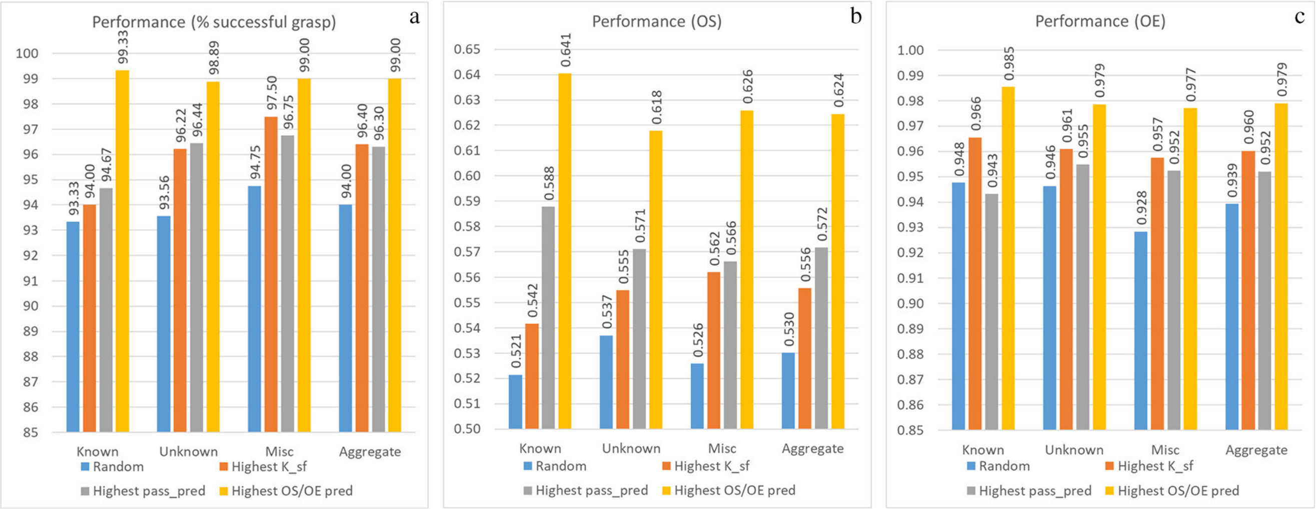

Fig. 17 Performance of various grasp approaches by object series and category from 4000 trials. **a** performance in terms of grasp rate. **b** performance in terms of the proposed overlap score. **c** performance in terms of the proposed orientational score

6 Conclusion

A platform was proposed in this paper to benchmark vision-based grasp synthesis methodologies. In addition, two grasp quality scores were explored to compliment the traditional grasp rate metric used by many works to assess their methodology. A prototype system was constructed. 4000 trials were conducted to evaluate four separate grasp approaches. It was shown that the benchmarking platform was robust to many trials and capable of serving as a platform for the deployment of various grasp approaches. The results revealed that the proposed scores provided useful insights related to the tested methodology. Moreover, the metrics comprehensively quantified the performance of a methodology and provided a fair baseline for comparison between differing approaches.

Exhaustive design details, source code, datasets, trained networks, results, and reference material is available online at the link given in Section 3.

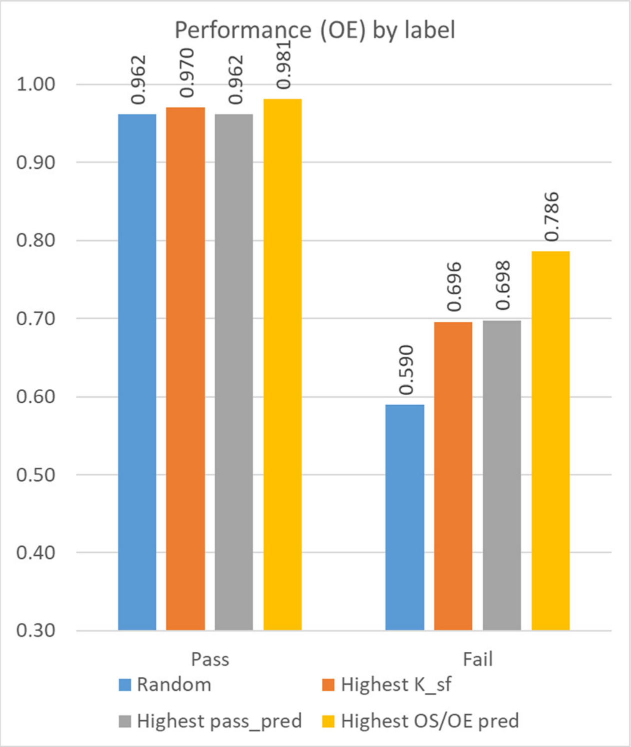

Fig. 18 Performance in terms of the proposed overlap score by *pass/fail* label

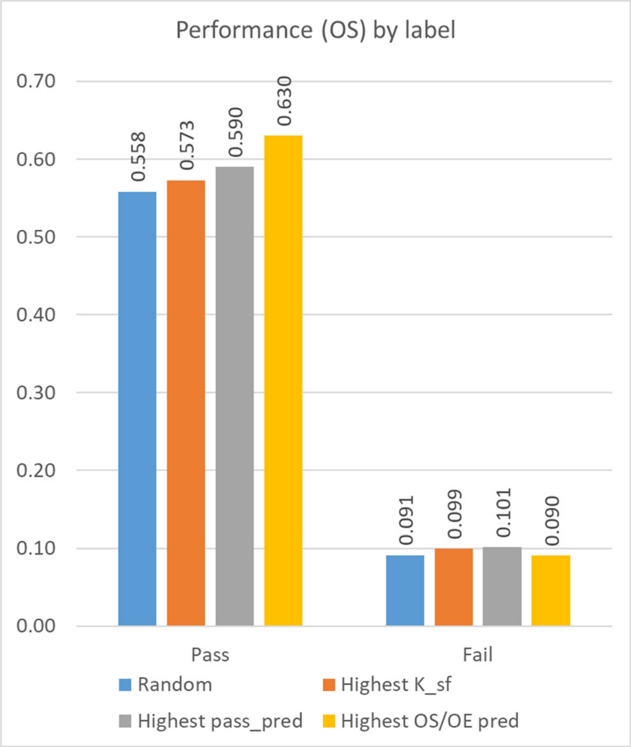

Fig. 19 Performance in terms of the proposed orientational score by *pass/fail* label

Acknowledgements This research was funded and supported by the Richard and Mary Earle Technology Trust, the Ken and Elizabeth Powell Bursary, and the Massey University Foundation.

Data Availability / Code Availability Related resources such as source code, datasets, networks, CAD drawings, designs, and microcontroller code are available at: https://drive.google.com/open?id=1VsEjCl6hrX3FeL9VRF-J9CzVL7JHXO15.

Author's Contributions Jacques Janse van Vuuren (40%). Liqiong Tang (30%). Ibrahim Al-Bahadly (15%). Khalid Mahmood Arif (15%). Please note that all authors contributed equally to the research work of which this paper is related to. However, not all authors contributed equally to the production of this paper.

Funding This research was funded and supported by the Richard and Mary Earle Technology Trust, the Ken and Elizabeth Powell Bursary, and the Massey University Foundation.

Declarations

Conflicts of Interests/Competing Interests This research is supported by Massey University, New Zealand. The authors declare that they have no known conflicts of interest or competing interests.

References

1. Xiang, Y., Mottaghi, R., and Savarese, S. Beyond PASCAL: A benchmark for 3D object detection in the wild, In IEEE Winter Conf Appl Comput Vis, 24–26, pp. 75–82 (2014), doi: https://doi.org/10.1109/WACV.2014.6836101
2. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010)
3. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Adv Neural Inf Processing Syst. **25**, 1097–1105 (2012)
4. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision, pp. 740–755. Springer, Cham (2014)
5. Robotics, I.E.E.E., Magazine, A.: Special issue on replicable and measurable Robotics research. IEEE Robot Autom Mag. **21**(3), 153–153 (2014). https://doi.org/10.1109/MRA.2014.2346083
6. Bonsignorio, F., Del Pobil, A.P.: Toward replicable and measurable robotics research [from the guest editors]. IEEE Robot Autom Mag. **22**(3), 32–35 (2015)
7. Mahler, J., Platt, R., Rodriguez, A., Ciocarlie, M., Dollar, A., Detry, R., Roa, M.A., Yanco, H., Norton, A., Falco, J., Wyk, K.., Messina, E., Leitner, J.'.J.'., Morrison, D., Mason, M., Brock, O., Odhner, L., Kurenkov, A., Matl, M., Goldberg, K.: Guest editorial open discussion of robot grasping benchmarks, protocols, and metrics. IEEE Trans. Autom. Sci. Eng. **15**(4), 1440–1442 (2018)
8. Vuuren, J.J.V., Tang, L., Al-Bahadly, I., Arif, K.M.: A 3-stage machine learning-based novel object grasping methodology. IEEE Access. **8**, 74216–74236 (2020). https://doi.org/10.1109/ACCESS.2020.2987341
9. Vuuren, J. J. V., Tang, L., Al-Bahadly, I., and Arif, K. Towards the autonomous robotic gripping and handling of novel objects, In 2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA), 19–21 , pp. 1006–1010, (2019) doi: https://doi.org/10.1109/ICIEA.2019.8833691
10. Jiang, Y., Moseson, S., Saxena, A.: Efficient grasping from RGBD images: Learning using a new rectangle representation. In: 2011 IEEE International conference on robotics and automation, pp. 3304–3311. IEEE (2011)
11. Chu, F., Xu, R., Vela, P.A.: Real-world multiobject, multigrasp detection. IEEE Robot Autom Lett. **3**(4), 3355–3362 (2018). https://doi.org/10.1109/LRA.2018.2852777
12. L. Pinto and A. Gupta Supersizing self-supervision: Learning to grasp from 50K tries and 700 robot hours, In 2016 IEEE International Conference on Robotics and Automation (ICRA), 16–21, pp. 3406–3413 (2016), doi: https://doi.org/10.1109/ICRA.2016.7487517
13. Levine, S., Pastor, P., Krizhevsky, A., Ibarz, J., Quillen, D.: Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. Int J Robot Res. **37**(4–5), 421–436 (2018)
14. Calli, B., Singh, A., Walsman, A., Srinivasa, S., Abbeel, P., and Dollar, A. M. The YCB object and Model set: Towards common benchmarks for manipulation research, In 2015 International Conference on Advanced Robotics (ICAR), 27–31, pp. 510–517 (2015), doi: https://doi.org/10.1109/ICAR.2015.7251504
15. Calli, B., Walsman, A., Singh, A., Srinivasa, S., Abbeel, P., Dollar, A.M.: Benchmarking in manipulation research: The YCB object and model set and benchmarking protocols. arXiv. (2015) preprint arXiv:1502.03143
16. Leitner, J., Tow, A.W., Sünderhauf, N., Dean, J.E., Durham, J.W., Cooper, M., Eich, M., Lehnert, C., Mangels, R., McCool, C., Kujala, P.T.: The ACRV picking benchmark: A robotic shelf picking benchmark to foster reproducible research. In: 2017 IEEE International Conference on Robotics and Automation (ICRA), pp. 4705–4712. IEEE (2017)
17. Leitner, J.. The ARCV Picking Benchmark (APB). http://juxi.net/acrv-picking-benchmark/. Accessed 9 July 2019
18. Sainul, I.A., Deb, S., Deb, A.K.: A novel object slicing based grasp planner for 3D object grasping using underactuated robot gripper. In: IECON 2019-45th Annual Conference of the IEEE Industrial Electronics Society, vol. 1, pp. 585–590. IEEE (2019)
19. Rothling, F., Haschke, R., Steil, J.J., Ritter, H.: Platform portable anthropomorphic grasping with the bielefeld 20-DOF shadow and 9-DOF TUM hand. In: 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2951–2956. IEEE (2007)
20. Brown, E., Rodenberg, N., Amend, J., Mozeika, A., Steltz, E., Zakin, M.R., Lipson, H., Jaeger, H.M.: Universal robotic gripper based on the jamming of granular material. Proc. Natl. Acad. Sci. **107**(44), 18809–18814 (2010)
21. Varley, J., Weisz, J., Weiss, J., and Allen, P. Generating multi-fingered robotic grasps via deep learning, In 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 28 Sept.-2 Oct. 2015 2015, pp. 4415–4420, doi: https://doi.org/10.1109/IROS.2015.7354004
22. Rubert, C., León, B., Morales, A., Sancho-Bru, J.: Characterisation of grasp quality metrics. J. Intell. Robot. Syst. **89**(3–4), 319–342 (2018)
23. Goins, A.K., Carpenter, R., Wong, W.K., Balasubramanian, R.: Evaluating the efficacy of grasp metrics for utilization in a gaussian process-based grasp predictor. In: 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 3353–3360. IEEE (2014)
24. Kumra, S., Kanan, C.: Robotic grasp detection using deep convolutional neural networks. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 769–776. IEEE (2017)
25. Caldera, S., Rassau, A., and Chai, D. Robotic Grasp Pose Detection Using Deep Learning, In 2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV), 18–21, pp.

1966-1972 (2018), doi: https://doi.org/10.1109/ICARCV.2018.8581091

26. Wang, Z., Li, Z., Wang, B., Liu, H.: Robot grasp detection using multimodal deep convolutional neural networks. Adv. Mech. Eng. **8**(9), 1687814016668077 (2016)
27. Sun, C., Yu, Y., Liu, H., and Gu, J. Robotic grasp detection using extreme learning machine, In 2015 IEEE International Conference on Robotics and Biomimetics (ROBIO), 6–9, pp. 1115–1120 (2015), doi: https://doi.org/10.1109/ROBIO.2015.7418921
28. J. Mahler *et al.*, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," *arXiv preprint arXiv:1703.09312,* (2017)
29. Morrison, D., Corke, P., Leitner, J.: Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach. arXiv. (2018) preprint arXiv:1804.05172
30. Cheng, H. and Meng, M. Q. A Grasp Pose Detection Scheme with an End-to-End CNN Regression Approach, In 2018 IEEE International Conference on Robotics and Biomimetics (ROBIO), 12–15, pp. 544–549 (2018), doi: https://doi.org/10.1109/ROBIO.2018.8665219
31. Zeng, A. *et al.* Robotic Pick-and-Place of Novel Objects in Clutter with Multi-Affordance Grasping and Cross-Domain Image Matching, In 2018 IEEE International Conference on Robotics and Automation (ICRA), 21–25, pp. 1–8 (2018), doi: https://doi.org/10.1109/ICRA.2018.8461044
32. Zeng, A. *et al.* Multi-view self-supervised deep learning for 6D pose estimation in the Amazon Picking Challenge, In 2017 IEEE International Conference on Robotics and Automation (ICRA), 29 May-3 June 2017, pp. 1386–1383 (2017), doi: https://doi.org/10.1109/ICRA.2017.7989165
33. Watson, J., Hughes, J., Iida, F.: Real-world, real-time robotic grasping with convolutional neural networks. In: Annual Conference Towards Autonomous Robotic Systems, pp. 617–626. Springer, Cham (2017)
34. Miller, A.T., Allen, P.K.: Graspit! A versatile simulator for robotic grasping. IEEE Robot Autom Mag. **11**(4), 110–122 (2004)
35. Miller, A., Allen, P., Santos, V., Valero-Cuevas, F.: From robotic hands to human hands: a visualization and simulation engine for grasping research. An International Journal, Industrial Robot (2005)
36. Miller, A. and Allen, P.. GraspIt! https://graspit-simulator.github.io/. Accessed 2 Mar 2020
37. Ulbrich, S., Kappler, D., Asfour, T., Vahrenkamp, N., Bierbaum, A., Przybylski, M., Dillmann, R.: The OpenGRASP benchmarking suite: An environment for the comparative analysis of grasping and dexterous manipulation. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1761–1767. IEEE (2011)
38. León, B., Ulbrich, S., Diankov, R., Puche, G., Przybylski, M., Morales, A., Asfour, T., Moisio, S., Bohg, J., Kuffner, J., Dillmann, R.: Opengrasp: a toolkit for robot grasping simulation. In: International Conference on Simulation, Modeling, and Programming for Autonomous Robots, pp. 109–120. Springer, Berlin, Heidelberg (2010)
39. Qian, W., Xia, Z., Xiong, J., Gan, Y., Guo, Y., Weng, S., Deng, H., Hu, Y., Zhang, J.: Manipulation task simulation using ROS and Gazebo. In: 2014 IEEE International Conference on Robotics and Biomimetics (ROBIO 2014), pp. 2594–2598, IEEE (2014)
40. Chen, J., Deng, H., Chai, W., Xiong, J., Xia, Z.: Manipulation task simulation of a soft pneumatic gripper using ros and gazebo. In: 2018 IEEE International Conference on Real-time Computing and Robotics (RCAR), pp. 378–383. IEEE (2018)
41. Katyal, K.D., Staley, E.W., Johannes, M.S., Wang, I.J., Reiter, A., Burlina, P.: In-hand robotic manipulation via deep reinforcement learning. In: Proceedings of the Workshop on Deep Learning for Action and Interaction, in Conjunction with Annual Conference on Neural Information Processing Systems, vol. 9. Barcelona, Spain (2016)
42. Diankov, R., Kuffner, J.: Openrave: A planning architecture for autonomous robotics. Robotics Institute, Pittsburgh, PA. Tech. Rep. (2008) CMU-RI-TR-08-34, 79
43. Kootstra, G., Popović, M., Jørgensen, J.A., Kragic, D., Petersen, H.G., Krüger, N.: VisGraB: a benchmark for vision-based grasping. Paladyn, J. Behav. Robot. **3**(2), 54–62 (2012)
44. Chitta, S., Sucan, I., Cousins, S.: Moveit![ros topics]. IEEE Robot Autom Mag. **19**(1), 18–19 (2012)
45. Görner, M., Haschke, R., Ritter, H., Zhang, J.: Moveit! task constructor for task-level motion planning. In: 2019 International Conference on Robotics and Automation (ICRA), pp. 190–196. IEEE (2019)
46. Mouret, J.B., Chatzilygeroudis, K.: 20 years of reality gap: a few thoughts about simulators in evolutionary robotics. In: Proceedings of the Genetic and Evolutionary Computation Conference Companion, pp. 1121–1124 (2017)
47. Koos, S., Mouret, J.-B., and Doncieux, S. Crossing the reality gap in evolutionary robotics by promoting transferable controllers, In Proceedings of the 12th annual conference on Genetic and evolutionary computation, pp. 119–126 (2010)
48. Martinez-Gonzalez, P., Oprea, S., Garcia-Garcia, A., Jover-Alvarez, A., Orts-Escolano, S., Garcia-Rodriguez, J.: Unrealrox: an extremely photorealistic virtual reality environment for robotics simulations and synthetic data generation. Virtual Reality. 1–18 (2019)
49. Collins, J., Howard, D., and Leitner, J. Quantifying the reality gap in robotic manipulation tasks, In 2019 International Conference on Robotics and Automation (ICRA), : IEEE, pp. 6706–6712 (2019)
50. James, S. *et al.* Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 12627–12637 (2019)
51. Wang, L., DelPreto, J., Bhattacharyya, S., Weisz, J., and Allen, P. K. A highly-underactuated robotic hand with force and joint angle sensors, In 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems,: IEEE, pp. 1380–1385 (2011)
52. Dollar, A.M., Howe, R.D.: Joint coupling design of underactuated hands for unstructured environments. Int J Robot Res. **30**(9), 1157–1169 (2011)
53. Klingbeil, E., Rao, D., Carpenter, B., Ganapathi, V., Ng, A. Y., and Khatib, O. Grasping with application to an autonomous checkout robot, In 2011 IEEE International Conference on Robotics and Automation, 9–13 May 2011, pp. 2837–2844 (2011) doi: https://doi.org/10.1109/ICRA.2011.5980287
54. Pas, A. T. and Platt, R. Using geometry to detect grasps in 3d point clouds, arXiv preprint arXiv:1501.03100*,* (2015)
55. Zelenak, A., Brabec, C., Thompson, J., Hashem, J., Fernandez, B., Pryor, M.: Intelligent grasping with the robotic opposable thumb. Appl. Artif. Intell. **28**(8), 737–750 (2014)
56. Kalashnikov, D. *et al.* Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation, arXiv preprint arXiv:1806.10293, (2018)
57. Manuelli, L., Gao, W., Florence, P., and Tedrake, R. kPAM: KeyPoint Affordances for Category-Level Robotic Manipulation, arXiv preprint arXiv:1903.06684*,* (2019)
58. Levine, S., Wagener, N., and Abbeel, P. Learning contact-rich manipulation skills with guided policy search, arXiv preprint arXiv:1501.05611, (2015)
59. Pinto, L., Davidson, J., and Gupta, A. Supervision Via Competition: Robot Adversaries for Learning Tasks, In 2017 IEEE International Conference on Robotics and Automation (ICRA): IEEE, Pp. 1601-1608 (2017)
60. Dollar, A. M. and Howe, R. D., Simple, robust autonomous grasping in unstructured environments, In Proceedings 2007 IEEE

International Conference on Robotics and Automation, : IEEE, pp. 4693–4700 (2007)

61. Dollar, A.M., Howe, R.D.: The highly adaptive SDM hand: design and performance evaluation. Int J Robot Res. **29**(5), 585–597 (2010)
62. Morrison, D. *et al.* Cartman: The low-cost cartesian manipulator that won the amazon robotics challenge," In 2018 IEEE International Conference on Robotics and Automation (ICRA) : IEEE, pp. 7757–7764 (2018)
63. Johns, E., Leutenegger, S., and Davison, A. J. Deep learning a grasp function for grasping under gripper pose uncertainty, In 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 9–14 Oct. 2016, pp. 4461–4468 (2016), doi: https://doi.org/10.1109/IROS.2016.7759657
64. Kopicki, M., Detry, R., Schmidt, F., Borst, C., Stolkin, R., and Wyatt, J. L. Learning dexterous grasps that generalise to novel objects by combining hand and contact models, In 2014 IEEE International Conference on Robotics and Automation (ICRA), May 31 2014–June 7 2014, pp. 5358–5365 (2014), doi: https://doi.org/10.1109/ICRA.2014.6907647
65. Saxena, A., Driemeyer, J., Ng, A.Y.: Robotic grasping of novel objects using vision. Int J Robot Res. **27**(2), 157–173 (2008)
66. Saxena, A., Driemeyer, J., Kearns, J., and Ng, A. Y. Robotic grasping of novel objects, In Advances in neural information processing systems, pp. 1209–1216 (2007)
67. Microsoft, Microsoft LifeCam Studio Technical Data Sheet. [Online]. Available: http://download.microsoft.com/download/0/9/5/0952776D-7A26-40E1-80C4-76D73FC729DF/TDS_LifeCamStudio.pdf. . Accessed 28 July 2020
68. RS Components Ltd, Datasheet: RS Pro 2.5m White LED Strip 136–3579, n.d. [Online]. Available: https://docs.rs-online.com/256e/0900766b815e69e7.pdf. Accessed 28 July 2020
69. Carlo Gavazzi Holding, Datasheet: Photoelectrics Diffuse-reflective Type PA18C.D..., DC, 2017. [Online]. Available: https://docs.rs-online.com/906c/0900766b8170a64f.pdf. Accessed 05 Aug 2020
70. HT Sensor Technology Co. Ltd. TAL220 Parallel Beam Load Cell. https://cdn.sparkfun.com/datasheets/Sensors/ForceFlex/TAL220M4M5Update.pdf. Accessed 16 July 2019
71. Avia Semiconductor. HX711. 24-Bit Analog-to-Digital Converter (ADC) for Weigh Scales. https://cdn.sparkfun.com/datasheets/Sensors/ForceFlex/hx711_english.pdf. Accessed 16 July 2019
72. Changzhou Songyang Machinery & electronics new technic institute, High torque hybrid stepping motor specifications, 2012. [Online]. Available: https://www.pololu.com/file/0J629/SY57STH76-2804A.pdf. Accessed 06 Aug 2020
73. Dobot.cc. DOBOT Magician: Lightweight Intelligent Training Robotic Arm - An all-in-one STEAM Education Platform. https://www.dobot.cc/dobot-magician/product-overview.html. Accessed 17 Apr 2020
74. Shenzhen Yuejiang Technology Co Ltd., Dobot Magician User Guide, 2018, issue V1.5.1. [Online]. Available: https://download.dobot.cc/product-manual/dobot-magician/pdf/V1.5.1/en/Dobot-Magician-User-Guide-V1.5.1.pdf. Accessed 01 Sept 2020
75. Mean Well Enterprises, 60W Single Output Industrial DIN Rail Power Supply MDR-60 series, n.d. [Online]. Available: https://www.meanwell-web.com/content/files/pdfs/productPdfs/MW/Mdr-60/MDR-60-spec.pdf. Accessed 01 Sept 2020
76. L. Chinfa Electronics IND. CO., AMR1 SERIES, 2014. [Online]. Available: https://docs.rs-online.com/b0cf/0900766b8144d4b2.pdf. Accessed 11 Sept 2020
77. Leadshine Technology Co. Ltd., M542 Economical Microstepping Driver, n.d. [Online]. Available: http://www.leadshine.com/UploadFile/Down/M542d.pdf. Accessed 11 Sept 2020
78. Jaccard, P., Distribution de la Flore Alpine dans le Bassin des Dranses et dans quelques régions voisines. 1901, pp. 241–72
79. Szegedy, C., Toshev, A., Erhan, D.: Deep neural networks for object detection. Adv. Neural Inf. Proces. Syst. 2553–2561 (2013)
80. Sobti, A., Arora, C., and Balakrishnan, M. Object detection in real-time systems: Going beyond precision, In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV): IEEE, pp. 1020–1028 (2018)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Jacques Janse van Vuuren received his M.Eng. degree in Mechatronic Engineering from Massey University, New Zealand in 2017 under the supervision of Dr. Liqiong Tang. His research focused on the development of a new, industrial modular robotic system capable of reconfiguration into commonly used robot types. His work also produced a patent. His research interests include robotics and automation, machine learning, artificial intelligence, digital image processing and autonomous object handling.

Liqiong Tang received a BSc (Eng) in 1982 from Southwest Petroleum University in China and a PhD in 1992 from the University of Liverpool in UK. She is a senior member of IEEE and has research experience in China, UK, Singapore, and Japan. Her research interests include robotics, mechatronics, intelligent control, sensing, and industrial automation. She has been with Massey University since 1996. She was one of the pioneers who established the Mechatronics and Robotics teaching and research at Massey University and the program leader for over 10 years. She has been involved in a number of research projects from industry, healthcare and defence force and continuously obtained research funding from industry and major funding bodies such as National Science Challenge, Callaghan Innovation, and Schulenburg to support her PhD/Master research students. She has served as an assessor for major science and technology funding bodies and board member/reviewer for international journals in robotics, mechatronics, and control.

Ibrahim Al-Bahadly received a B.Sc. (Eng.) degree from Baghdad University of Technology in 1987, followed by M.Sc. and Ph.D. from Nottingham University, in 1990 and 1994, respectively, all in electrical and electronic engineering. From 1994 to 1996, he was a research associate with the electric drives and machines group at the University of Newcastle upon Tyne, UK. Since 1996, he has been with Massey University, where he is currently Associate Professor in electrical and electronic engineering. His research interests include power electronic applications, variable speed machines and drives, renewable energy system, instrumentation, robotics and automation. Ibrahim is a senior member of the Institute of Electrical and Electronics Engineers (IEEE).

Khalid Mahmood Arif is a senior lecturer in Mechatronics and Robotics at Massey University Auckland. He holds a PhD in Mechanical Engineering from Purdue University West Lafayette, Indiana (2011), a Master of Engineering from The University of Tokyo (2004), and a BE (Hons) in Mechanical Engineering from the University of Engineering and Technology Lahore (2000). Prior to joining Massey, he was an Assistant Professor in the Department of Mechatronics and Control Engineering at the University of Engineering and Technology Lahore. His research interests include sensors, IoT, robotics, and additive manufacturing.