Received March 11, 2020, accepted April 7, 2020, date of publication April 13, 2020, date of current version May 1, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2987341

A 3-Stage Machine Learning-Based Novel Object Grasping Methodology

JACQUES JANSE VAN VUUREN, LIQIONG TANG, (Senior Member, IEEE), IBRAHIM AL-BAHADLY, (Senior Member, IEEE), AND KHALID MAHMOOD ARIF, (Senior Member, IEEE)

Department of Mechanical and Electrical Engineering, SF&AT, Massey University, Auckland 0632, New Zealand

Corresponding author: Jacques Janse Van Vuuren (j.jansevanvuuren@massey.ac.nz)

This work was supported in part by the Massey University Manawatū, New Zealand, and in part by the Richard and Mary Earle Technology Trust, the Ken and Elizabeth Powell Bursary, and the Massey University Foundation, New Zealand.

ABSTRACT The automatic grasping of objects previously unseen by a robotic system is a difficult task—of which there is currently no robust solution. The research presented in this article improves upon previous works that employ depth data and learning techniques to generate and select from a pool of hypothesised grasps by focusing on the pruning and selection process. In this work, a vision-based, sampling methodology that generates candidate grasps through a convolutional neural network is proposed. Each candidate grasp is assessed using scores derived from the candidate itself and other related input modalities—such as the centre of gravity of the object. The final selection is determined by a learning algorithm. To overcome human bias, objective measures of grasp performance are established that comprehensively measure the error introduced by the grasp trial itself. The proposed metrics are empirically demonstrated to quantify grasp quality, offer useful criteria for network training and provide better descriptive power than traditional measures of grasp outcome. Experimentation showed that the proposed methodology can generate a meaningful, final grasp within 1.3 seconds. Trials quantitatively demonstrate a small-object-in-isolation performance of 99%. For unknown objects, this equates to a 10% improvement relative to other similar methodologies. Testing also showed that grasp performance was improved by 5% when implementing the proposed metrics—compared to the baseline.

INDEX TERMS Object grasp detection, robotic grasping, part-handling, machine vision, machine learning, robotic learning, AI-based flexible automation system.

I. INTRODUCTION

Autonomous novel object grasping and handling is a wide-ranging, high-impact field with many implications, especially within domestic and industrial application. Some instances where automatic grasping has been studied include an automated checkout robot [1], garbage sorting [2], cloth manipulation [3], bed making [4], dishwasher unloading [5], automated cooking [6], [7], service robotics [8]–[11], general household-related grasping [12]–[14], clutter clearing [15], [16] and stowing, picking and packing for warehouse automation [17]–[20]—which has gained significant traction since the 2017 Amazon Robotics Challenge [21].

The automated manipulation of objects previously unseen by a robotic system is an extremely difficult task, as a good grasp is related to object shape, size, material, weight-distribution, surface properties, friction coefficients and object deformability, and can be severely affected by sensing and actuation accuracy. Moreover, the relationship between these variables and a specific grasping strategy, robotic hardware and a gripper is not always clear.

Research in this field has been active for decades, yielding a colourful range of promising avenues—especially with the recent interest from well-known and well-resourced research institutions, such as Google and the Massachusetts Institute of Technology (MIT). With the attention of such institutions, we have seen unprecedented, large-scale dataset generation frameworks that allow for the training of complex, self-supervised neural networks [22]–[25]. Moreover, machine learning in general has become overwhelmingly

The associate editor coordinating the review of this manuscript and approving it for publication was Ehsan Asadi.

represented within this area, with a great deal of work utilising RGB and/or depth input modalities [12], [26], [27]. Despite the success seen throughout literature, the automated manipulation of novel objects remains challenging and an active topic of research.

Over the past three decades consumers have increasingly been demanding a wider variety of goods in smaller batches, resulting in rapid changes in production technologies [28]–[30]. This trend reveals the importance for flexible, reconfigurable and automated production systems for future markets [31]–[33]—which is not usually considered by object manipulation literature. Although flexible robotic hardware is progressively becoming a popular topic, such as the dual-arm, scalable concept developed by ABB [34], [35], the adaptability of the related grasping methodology is not usually considered to the same degree. Fully manual assembly lines are still common in low-wage countries, particularly for manufacturers of consumer electronics, small appliances, toys, etc. Novel object manipulation methodologies that utilise object identification, accurate grasping location/orientation and a robust handling process play a crucial role in future automation and production lines. This paper presents a grasping methodology based on machine learning that aims to improve on research related to novel object detection, grasping and handling with autonomous robotic systems—further closing the gap between manual and fully automated production.

Automated grasping is typically posed as a search problem—find the location that will best facilitate handling of the object from a potential infinite number of grasps. The goal of our research is to sample some of these candidate locations and select a meaningful subset of grasp hypotheses, which may then be pruned based on quality metrics to select a suitable and reliable grasp for execution. In contrast to many current systems that utilise depth information [18], [20], [27], [36], 3D models of objects [37]–[39] or wrist-mounted sensors [12], [26], [40], our methodology operates on raw, monocular RGB observations of the scene.

This paper presents an approach that utilises machine learning and part-related information to find, grade and select suitable robotic grasping locations in 3D space and is an extension of our previous work [41]. In this paper, a selection-stage is established. The datasets used for training have also been considerably improved. Moreover, the proposed methodology is implemented on a physical robot and trials are conducted for validation. The method consists of three main stages, each coupled with a learning component. First, a classifier is trained to determine whether the object within a region of interest is known or unknown. For unknown objects, a small convolutional neural network (CNN) quickly classifies segments of the object through vision to identify potential grasping locations. A scoring network is then used to rank these locations and decide on the final grasping position. Input features for the scoring network are derived from the assessed grasp itself and other features related to the object or grasp location. Methodology performance is quantified as per literature and the proposed set of metrics. To evaluate the proposed methodology, experimentation and testing focus specifically on 2-fingered, parallel jaw gripping that uses force-closure within the scope of object-agnostic grasping, approached from an industrial perspective.

The performance evaluation of our system through physical trials demonstrates quantitatively that our approach can grasp small objects not seen during training 98.9% of the time—despite relying only on rudimentary sensing, such as an RGB webcam. Compared to the relevant literature, this constitutes an improvement of roughly 10%. An illumination-controlled imaging chamber and conveyor system was constructed for dataset generation and methodology testing. Objects are placed haphazardly at one end of the conveyor and grasped at the other end. The final system generated a grasp within 1.3 seconds, producing on average 83 viable grasps per object. New quantitative metrics that more accurately reflect the quality of a grasp have also been proposed. Trials revealed that the proposed metrics are capable of further improving grasp rates by 2.7% for unknown objects and 5.3% for known objects—compared to highest confidence selection.

This paper is organised as follows. Section 2 introduces current approaches to novel object grasping and describes the associated difficulties therein. Section 3 states the proposed metrics used to define and improve grasping performance. Sections 4 to 7 cover the proposed methodology, experimentation, analysis and future research. Finally, section 8 presents some of the conclusions from this research.

II. GRASPING LITERATURE

A. OBJECT GRASPING CHALLENGES

The robotic grasping of unfamiliar objects has grown to be a well-studied field within manipulation and is approached in many ways. Although a comprehensive overview of grasping is not within the scope of this paper, we refer the reader to the widely cited grasp synthesis survey by Bohg *et al.* [42].

Some grasping approaches have tried reducing the importance of gripper placement by increasing the dexterity of the end-effector itself. Brown *et al.* considered a granular jamming end-effector design, in which a mass is pressed onto an object [43]. By eliminating the air within the mass via vacuum, the shape conforms to the candidate object. Their methodology showed excellent performance for a wide range of objects and significantly reduced hardware and software complexity—but lacked gripping force for round, flat and small objects. Welhenge, Wijesinghe and Rajakaruna proposed a universal 3-fingered gripper designed to emulate human finger motion when grasping objects [44]. Similarly, Huang, Lehman, Mok, Miikkulainen and Sentis made use of an evolutionary search model and the *MekaHand*—a simulated 5-fingered, humanoid gripper [45]. Odhner, Ma and Dollar took inspiration from the 2-fingered strategy humans employ to grasp objects from a table with their underactuated gripper design [46]. They showed good performance for small, thin objects like keys or coins—which is a common issue for many current works. The problem with fixed, design-based approaches is two-fold. Complex gripper designs usually excel at complicated tasks with specific requirements and are difficult to extend to a wider range of objects. Also, grasping locations for such end-effectors can be difficult to represent conceptually. 2-fingered gripping is usually favoured for research within this area [16], [47]–[49]. Our research is also focused on 2-fingered grasping as there is a plethora of representations for this modality. Moreover, 2-fingered gripping has demonstrated state-of-the-art performance for novel objects [12], [22], [23].

Point cloud and model-based approaches that use a 3D model of the candidate object to generate and select appropriate grasping locations are common. Arruda, Wyatt and Kopicki provide one such example, in which a wrist-mounted depth camera is driven to multiple locations around an object [50]. Up to 7 views are collated to optimise the surface reconstruction of the object, which may then be used to assess the quality of contact points around potential grasps. Unfortunately, the reliability of such methodologies declines as the quality of the model declines. Furthermore, obtaining accurate and complete model reconstructions has proven extremely difficult in practice due to depth sensor noise and the response of surface-dependent sensors [51]. To combat this, many works have turned to machine learning. Pas, Gualtieri, Saenko and Platt frame point cloud-based grasp detection as a binary classification problem, wherein partial or occluded views of objects are used to train a 4-layer CNN classifier [36]. Their dataset consists of 1.5 million hand-labelled examples of positive and negative grasps of 55 unique objects. Similarly Mahlet *et al.* use Dex-Net, a grasp quality CNN which predicts grasp robustness directly from reconstructed 3D models [39]. A synthetic dataset containing 6.7 million point clouds and other metrics was used to train their network. Their work was later posed as a cloud-based, grasp planning system specifically for 2-fingered grippers [52]. Both methodologies were physically trialled and achieved grasp rates of 93% or above for novel objects. Fischinger, Weiss and Vincze discretised point cloud data into many small topographical features [53]. A support vector machine (SVM) classifier was trained to recognise the pattern of such features that correspond to potential grasps. Their methodology showed good grasping rates in clutter for 3 varying robotic arms and 4 unique grippers.

Machine learning techniques have also been used to detect local grasp locations directly from sensor data without considering the object in context. This concept was pioneered by Jiang, Moseson and Saxena in their early work which focused on grasp representation [54]. By representing a 2-fingered grasp in terms of a *grasping rectangle*, hand-labelled, supervised learning approaches could be used to generate and evaluate numerous candidate grasps. Their later work [13] showed that their representation, coupled with depth data, could effectively grasp novel objects. Their dataset has since been adopted by others [55]–[61]. A plethora of similar generate-and-test methodologies utilising analogous representations have also since been proposed. Sun, Yu, Liu and Gu, for example, extract a histogram of gradient features from the Cornell Grasp Detection Dataset [55]. These features are used to train their classifier to find candidate grasping rectangles. A second network is then used for final candidate selection.

More sophisticated learning methodologies have also benefitted from the grasping rectangle representation. Pinto and Gupta used an unsupervised learning technique with over 50,000 grasp attempts and 700 robot hours to find appropriate grasps in terms of 2-dimensional RGB rectangle representations [24]. Adversarial learning has also been investigated within this context [25]. Other representations have also been proposed. kPAM from MIT, for example [12], use semantic 3D keypoints to strictly represent an object in terms of task-relevant geometric detail. Their representation proved effective for manipulating objects within the context of the desired task.

The research presented in this article is closely related to that of Lenz *et al.* [13] and Sun *et al.* [55]. They exploited RGB-D data and learning techniques to generate numerous candidate grasps, of which a final grasp was selected. However, neither methodology put enough emphasis on the candidate pruning and selection process. We propose an RGB sampling-based method that generates candidate grasps through a CNN. Each potential grasp is assessed using scores derived from the candidate, as well as other related input modalities, such as the centre of gravity (COG) of the object. The final selection is determined by a learning algorithm.

B. LACK OF STANDARDISATION

A current major issue within this field is the lack of shared benchmarks and performance metrics for comparison. This was specifically noted by Bohg, Morales, Asfour and Kragic in their grasp synthesis survey [42] and more recently by Morrison *et al.* [26]. Some methodologies, for example, simply train a new learning algorithm on popular grasp detection datasets such as the Cornell dataset [56], [57], [59]–[61] and report their classification accuracy as a potential grasp success rate—without physical trials. Although this might provide some basis for comparison for learned model performance, many systems are so drastically different that this single-faceted comparison is irrelevant in the context of how well the system will grasp objects in practice. Moreover, some only use parts of the database for training and comparison. Training success rates are also usually higher than physical trial success rates [24], [36]. Sun, Yu, Liu and Gu, for instance, noted an 11% disparity between dataset classification accuracy and physical trial outcome [55]. Commonly, datasets within this field are hand-annotated [12], [19], [54], [61]–[63]. This is somewhat problematic because training for dataset performance tunes a methodology toward grasps the human creators consider optimal for real-world implementation. Quantifying the degree to which such annotated grasps map to a physical system is extremely difficult—as evidenced by the various works that report similar performances with differing datasets.

Some approaches have tried to avoid human influence altogether by implementing self-supervised learning. The Google-affiliated work of Levine, Pastor, Krizhevsky and Quillen, for example, sees the implementation of an unsupervised convolutional neural network for closed-loop robotic grasping that utilises a single RGB input modality [23]. Over the course of three months, they generated over 800,000 training samples, using anywhere between 6 and 14 robotic manipulators at any one time. Their work was later improved upon by Kalashnikov *et al.* through the introduction of Qt-Opt—a scalable, deep reinforcement learning methodology that utilises a self-supervised framework to learn real-world grasping behaviours [22]. Without human intervention, they showed that—given enough training data—a neural network can learn distinct behaviours that facilitate grasping, e.g., re-grasping strategies for badly grasped objects, object probing, object repositioning for better manipulation and disturbance response [22]. Despite their success, hand-engineered labels still outperform unsupervised algorithms. To overcome human bias, this work proposes the institution of objective measures of grasp performance. By measuring the error introduced by a grasp trial itself, metrics can be established to assess grasp outcome, quantify grasp quality and provide useful criteria for network training.

Due to these broad comparison issues, many works cite their real-world, tested robotic grasp success rates [15], [22], [24], [25]. To help facilitate this, some have proposed standardised object test sets, where the aim is to have a shared pool of objects so that differing methodologies can be tested in the same way. Popular object test sets include the 42-object ARCV picking benchmark [64] and the 72-object YCB object and model set [65]. Despite many efforts, no standard object pools have been adopted by the wider community and generally many works will default to a ‘common household’ or ‘common laboratory’ object test group—which varies and is self-defined, but usually well-documented. Morrison *et al.* [26] suggest that the lack of standardisation within this field is related to the wide range of methodologies, lack of shared object test sets and limitations of physical hardware, e.g., the gripper may limit object shape or size and the robotic arm may limit object weight.

C. DEFINITION OF A SUCCESSFUL GRASP

As stated previously, many works evaluate performance in terms of their physically-trialled grasp success rate—which is the rate at which their system is deemed to successfully grasp objects. To add to the lack of clear baselines for comparison, this metric is also somewhat problematic because what constitutes a successful grasp is defined differently. Some consider a grasp successful if the object in question can be lifted to a pre-defined height without falling [22, 24-26, 54, 66]. Pas and Platt [37] consider a trial successful if the object can be grasped, lifted and transported to a collection box. Pinto and Gupta make use of force sensors [24], whereas others loosely define a successful grasp as lifting an object. To add to the confusion, some simply refer to an executed grasp as ‘successful’ or ‘unsuccessful’, without clearly defining what is meant by these terms. Based on such definitions, the current state-of-the-art novel object grasp rate sits between 85-95% [22], [38], [39], [52], [55]–[57].

Although such loose definitions have demonstrated capable of grasping objects with moderate reliability, it is clear that the binary pass/fail metric is not well-related to the quality of a grasp—in the sense that post-grasp placement is not considered. Although a grasp may be robust enough to facilitate the lifting of an object, the object may be displaced due to the grasp itself, resulting in poor object manipulation or placement—which is key for industry applications. Thus, this study aims to develop better notions of a ‘good grasp’ by quantitatively measuring the quality of a grasp based on how well an object has been picked up, transported and placed. This work is not only concerned with grasp outcome—but also grasp quality—and moves away from describing a grasp trial as successful or unsuccessful, opting rather for continuous scores that provide a larger spectrum to describe grasp outcome.

III. SIMILARITY METRICS

An object grasping study reveals that two factors contribute to how well an object is gripped and handled when using a 2-fingered gripper. The first is gripper alignment, demonstrated in Figure 1. When the two plates of a gripper are not parallel with or geometrically suited to the local gripping area, the object is forced into some gripper-relative position.

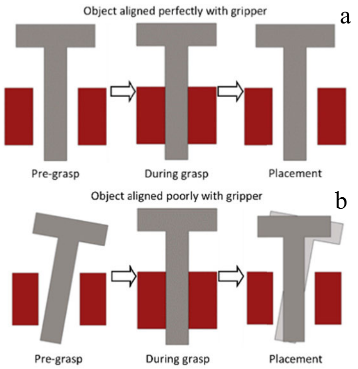
Object aligned perfectly with gripper
a
Pre-grasp
During grasp
Placement
Object aligned poorly with gripper
b
Pre-grasp
During grasp
Placement

FIGURE 1. **2-fingered gripping alignment illustration. (a)—depiction of a grasped object in which the two parallel plates of a gripper are perfectly aligned with the gripping area. (b)—depiction of poor alignment.**

A lack of consideration of the COG of an object can also affect manipulation quality. If sufficient gripping force has not been applied to the object, for example, it may droop, as depicted in Figure 2.

To capture as much of the problem as possible, we propose to measure the error introduced by the grasp in terms of translation and rotation. An overlap score *OS* is described by

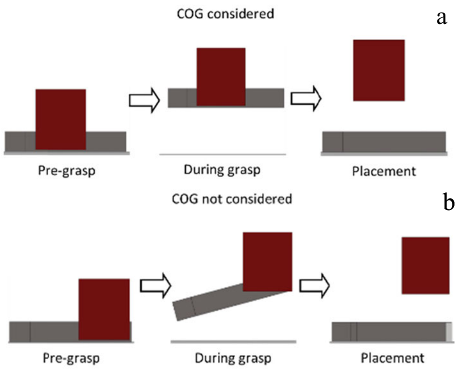

FIGURE 2. 2-fingered gripping COG illustration. (a)–depiction of a grasp in which the COG of the object has been considered. (b)–a grasp in which the COG was not considered.

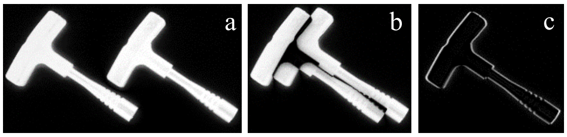

FIGURE 3. (a)–illustration of pre- and post-grasp object locations with no overlap. (b)–depiction of pre- and post- grasp locations with some overlap. (c)–depiction of pre- and post- grasp locations with significant overlap.

the Jaccard similarity index:

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

where A_{pre} is the top-down area of the object prior to the grasp and A_{post} is the top-down area of the object after the grasp. Similarly, an orientation error score OE compares the rotation of the object pre- and post-grasp:

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

where θ_{pre} represents the major orientation of the object pre-grasp. θ_{post} represents the major orientation of the object post-grasp. Both OS and OE range from $0 - 1$. OS returns a value of 0 if the performed grasp translates the object such that pre- and post-grasp objects do not overlap, as shown in Figure 3-a. OS approaches 1 (Figure 3-c) as the applied grasp tends toward a perfect grasp, where no pre- or post-grasp difference in translation or rotation is measured.

OE varies, depending on the amount of rotation introduced by the grasp, resulting in 1 with no change and 0 if the object is rotated by 180°. OE does not respond to translation. Such continuous scores can be used to train regression models or alternatively for classification through threshold techniques.

IV. PROPOSED METHODOLOGY AND SYSTEM DESCRIPTION

The methodology put forward in this paper consists of three main stages, each with their own learning component. Figure 4 illustrates the conceptual stages of the proposed system.

Stage-1 is a learner and classifier. It can identify known and unknown objects and learn new classes. Prior to stage-1, an image of the workspace is captured. Present objects are segmented and cropped using digital image processing techniques. Such images are then classified by the stage-1 learner. If the object within the image is correctly classified, a previously implemented grasp may be fitted to the object and passed to the manipulator for direct execution—completing the task. The methodology proceeds for further processing if the object is not confidently classified by the stage-1 network.

Stage-2 is an orientation and pose generator. Prior to stage-2, strong object orientations within the cropped image are found. The cropped image is then rotated based on these orientations. A small grasping window is iterated across each rotated image and assessed by the stage-2 classifier. Each correctly classified window and its relative location and orientation is recorded.

Stage-3 integrates information and optimises for grasp selection. It gathers information about the object, such as a secondary view of the object, the COG of the object, etc. The stage-3 learning framework scores each previously recorded window and its accompanying object-related information. The highest scoring grasping window may then be converted to robot space for physical implementation. The executed grasp location, orientation and object class may then be saved for future use, depending on the outcome.

A. GRASP REPRESENTATION

Machine vision was employed in this research for object grasp location and orientation detection. Consider the problem of detecting grasps for known and unknown objects given two monocular RGB observations of the scene (Figure 5).

Grasping positions from the perspective of the camera mounted at the top of the enclosure fall within top-view image space:

$$I_t\,[x, y] \in \mathbb{R}^{3\times H\times W} \tag{3}$$

where $[x, y]$ are each $\in \mathbb{R}^{3\times H\times W}$ and contain real pixel values. H and W are the height and width of the captured image respectively. Sensor properties are known *a priori*. To characterise a grasp within I_t, a 5-dimensional, rectangular planar representation is assumed—graphically illustrated in Figure 6.

The rectangular grasping window as an image is referred to as I_{RCW}. The orientation of this representation is defined by its rotation angle θ_t with respect to the $I_t x$ *axis* about centre position x_t, y_t. The relationship between I_{RCW} and the physical gripper placement is direct. The robotic tool is centred about x_t, y_t in real space and the two parallel plates of the gripper close perpendicular to θ_t. Rectangle height h_t relates to the physical width of the gripper and w_t relates to the available distance between the two plates of the gripper when fully extended. The top-view pose

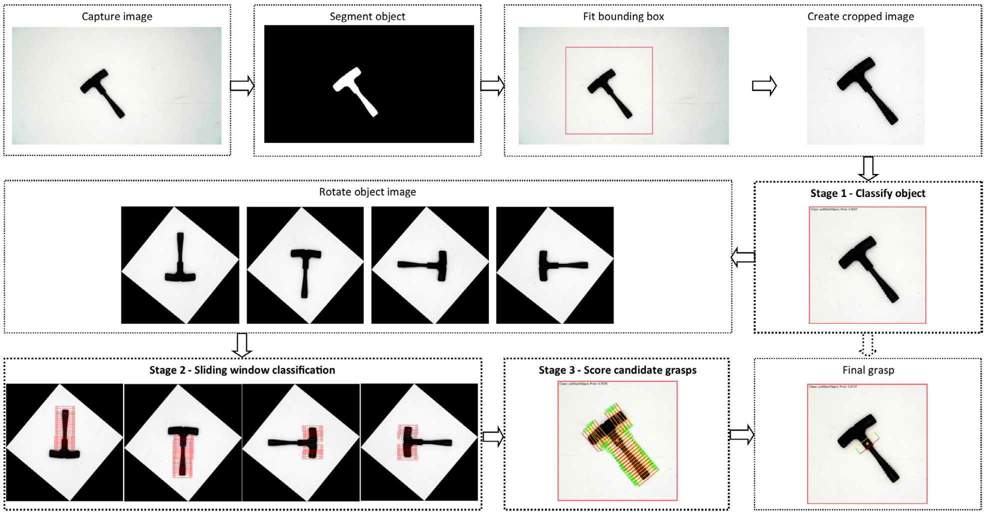
Capture image
Segment object
Fit bounding box
Create cropped image
Rotate object image
Stage 1 - Classify object
Stage 2 - Sliding window classification
Stage 3 - Score candidate grasps
Final grasp

FIGURE 4. Proposed 3-stage grasping methodology top-level process diagram. First, an object is captured and segmented from its background. Second, a bounding box is fitted and classified. If classified correctly, the object is rotated. Rotated images can then be used to generate numerous candidate grasps, of which a final grasp is selected.

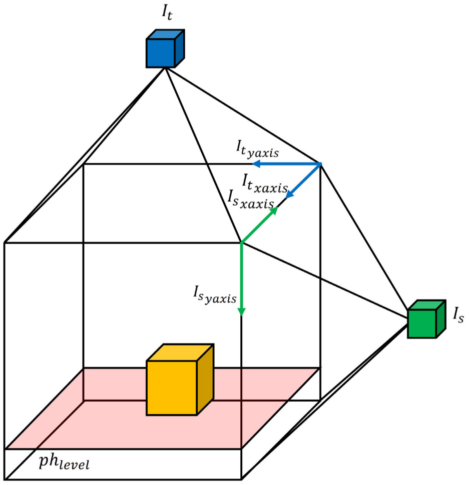

FIGURE 5. Illustration of object imaged from two distinct viewpoints. Camera space and the object lower limit have also been annotated.

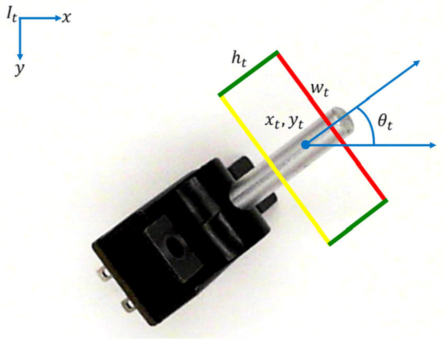

FIGURE 6. Illustration of 5-dimensional top-view grasping rectangle representation. $\mathbf{h_t}$ and $\mathbf{w_t}$ represent the height and width of the representation, respectively. θ_t is the orientation and $\mathbf{x_t}$,$\mathbf{y_t}$, the centre position of the rectangle.

component of this window in image space I_t can therefore be defined as:

$$pose_t = \{x_t, y_t, \theta_t, h_t, w_t\} \in I_t \qquad (4)$$

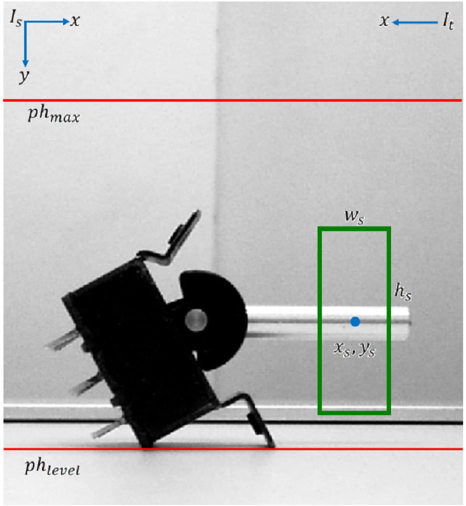

FIGURE 7. 4-dimensional planar representation of a grasping rectangle from the side view. $\mathbf{h_s}$ and $\mathbf{w_s}$ represent the height and width, respectively, and $\mathbf{x_s}$,$\mathbf{y_s}$ is the rectangle centre position.

In addition to the top mounted camera, a side camera produces a side-view image space:

$$I_s\,[x, y] \in \mathbb{R}^{3\times H\times W} \qquad (5)$$

Since two identical sensors are used, the height H and width W of I_s are identical to I_t. The planar representation within this space is graphically depicted in Figure 7.

The side-view rectangular window as an image is referred to as I_{SVW}. The side-view pose component of this window within I_s is represented as follows:

$$pose_s = \{x_s, y_s, h_s, w_s\} \in I_s \qquad (6)$$

where x_s, y_s is the window centre position and h_s, w_s are the height and width, respectively. Height h_s and width w_s vary depending on the distance between the object and the side camera, i.e., h_s and w_s increase in size as an object approaches the side camera. Physically, I_{SVW} relates to the side-on maximal contact area of the gripper when closing onto the object. x_s and y_s reflect the centre position of gripping pads and h_s and w_s are related to the height and width of the side gripping area respectively.

Combining both pose representations yields a multi-perspective location array that is referred to as a candidate grasping window:

$$g = \{pose_t, pose_s, S_c\} \tag{7}$$

where S_c is the candidate score array containing 10 scores related to the specific candidate. A comprehensive explanation of this array is provided in Section 4.2. Effectively, $pose_t$ can be used to find the x, y and θ components of an end-effector in robot space and $pose_s$ can be used to find the z component.

B. GRASP POSE GENERATION AND SELECTION

To facilitate the selection of a final grasp, a set of top-view grasp hypotheses is generated:

$$pose_{t_{1,\ldots,m}} = \left\{ \begin{array}{c} x_{t_1}, y_{t_1}, \theta_{t_1}, h_t, w_t \\ \vdots \\ x_{t_n}, y_{t_n}, \theta_{t_n}, h_t, w_t \\ \vdots \\ x_{t_m}, y_{t_m}, \theta_{t_m}, h_t, w_t \end{array} \right\} \tag{8}$$

where n ranges from 1 to m. h_t h_t and w_t are fixed according to the gripper size in pixels. Figure 8 provides a full graphical representation of this notation.

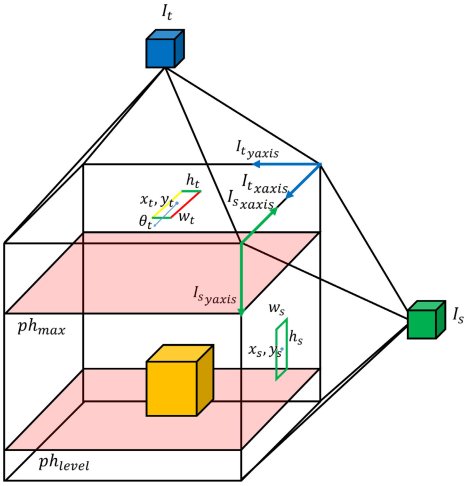

FIGURE 8. **Illustration of top-view and the side-view rectangle representations. Camera space, object lower limit and object maximum height have also been annotated.**

To find θ_{t_n}, I_t and I_s are first converted to grayscale using the Rec. 601 conversion, resulting in a pixel intensity range of $0 - 255$. Currently RGB data is ignored to improve performance, as a strong correlation between the number of input channels and computation time was observed—supported by general machine learning theory [67], [68]. A binary image I_{tTh} of height H and width W—which segments the object from its background—is created by thresholding the intensity values of each pixel within I_t:

$$I_{tTh}[x, y] = \begin{cases} 1, & if\ I_t[x, y] \geq T_t \\ 0, & if\ I_t[x, y]\ < T_t \end{cases} \tag{9}$$

where x ranges from 1 to W and y ranges from 1 to H. T_t is the threshold level applied to I_t. Several morphological filters are applied to I_{tTh} for noise reduction and to improve object clarity. For additional information related to digital image processing, an excellent resource is provided by Gonzalez and Woods [69]. I_{tTh} can be used to find the centroid of the object x_o, y_o by using the centre of mass calculation:

$$x_o = \frac{\sum_{x=1}^{W} \sum_{y=1}^{H} x I_{tTh}[x, y]}{N_{pix}} \tag{10}$$

$$y_o = \frac{\sum_{x=1}^{W} \sum_{y=1}^{H} y I_{tTh}[x, y]}{N_{pix}} \tag{11}$$

where N_{pix} is the number of pixels within I_{tTh} that register a value of 1:

$$N_{pix} = \sum_{x=1}^{W} \sum_{y=1}^{H} I_{tTh}[x, y] \tag{12}$$

N_{pix} is also considered to be the pixel-wise area of an object. A square, fixed-size bounding box I_{BB} of size $H_{I_{BB}} \times H_{I_{BB}}$ is fitted onto I_t at centre x_o, y_o. A graphical representation of this process is illustrated in Figure 3, step 3. $H_{I_{BB}}$ is constrained by the desired input size of the stage-1 classifier and the desired maximum object size. For this application, a value of 790 provided an optimal input size-maximum object trade-off.

I_{BB} is assessed by the stage-1 classifier to determine if the object is likely known. A pre-determined grasp may be fitted to the object directly if the classifier soft-max output $K_{sf_{class}}$ is above some classification acceptance threshold T_{class}. If $K_{sf_{class}} < T_{class}$, a Sobel filter is applied to each pixel within I_{BB} to calculate the horizontal gradient value $gradient_x$ and vertical gradient value $gradient_y$ of each pixel:

$$gradient_x[x, y] = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} * A \tag{13}$$

$$gradient_y[x, y] = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} * A \tag{14}$$

This filter responds to strong edges within an image. The resultant gradient angle $\theta_{gradient}$ is also calculated for each pixel within I_{BB}:

$$\theta_{gradient}[x, y] = arctan\left(\frac{gradient_y}{gradient_x}\right) \tag{15}$$

A number of pixels at specific angle vs. pixel angle histogram is tabulated. Local maxima within this plot (Figure 9) are used to determine the major orientations of an object.

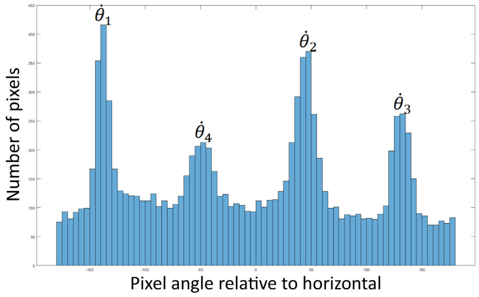

FIGURE 9. Typical number of pixels at unique gradient angle vs. pixel angle histogram. Peaks correspond to the angle of strong edges within an image.

The top four maxima are selected:

$$\theta_{Sobel} = \begin{pmatrix} \dot{\theta}_1 \\ \dot{\theta}_2 \\ \dot{\theta}_3 \\ \dot{\theta}_4 \end{pmatrix} \tag{16}$$

where $\dot{\theta}_1$ denotes the highest number of pixels counted at a specific angle and $\dot{\theta}_4$ denotes the 4th highest. I_{BB} is rotated 4 times about centre x_o, y_o by $\dot{\theta}_{1,\ldots,4}$, creating $I_{BBR_{1,\ldots,4}}$. A rectangular classification window I_{RCW_n}—which might contain an appropriate gripping area—of height h_t and width w_t is iteratively translated across each rotated bounding box image $I_{BBR_{1,\ldots,4}}$. At each stepped location $I_{BBR_{1,\ldots,4}}[i,j]$, I_{RCW_n} with centre $[i,j]$ is assessed by the stage-2 classifier. The soft-max output of this network is denoted by K_{sf_n}. Note that a binary variant of $I_{BBR_{1,\ldots,4}}$ is used to largely avoid object-absent areas, derived from $I_{t_{Th}}$. If K_{sf_n} is higher than some heuristic acceptance threshold T_{sf} then $x_{t_n}, y_{t_n}, \theta_{t_n}$ are known and can be added to the $pose_t$ set. x_{t_n} and y_{t_n} are found by transforming the I_{RCW_n} centre location within rotated space $I_{BBR_{1,\ldots,4}[i,j]}$ to the respective non-rotated space position $I_{BB}[x,y]$, which shares a coordinate frame with I_t. θ_{t_n} is taken as the respective angle $\dot{\theta}_{1,\ldots,4}$ at classification time. For each candidate, 6 related scores ranging from $0-1$ are calculated:

- $S_{ss_n}(symmetryscore)$. This score assesses the symmetry of the grasp within I_{RCW_n}. S_{ss_n} will score 1 if the grasp perfectly symmetric.
- $S_{cs_n}(centrescore)$. This score responds to the horizontal location of the grasp within I_{RCW_n}. If the grasp is perfectly situated around centre centre y_{t_n}, S_{cs_n} will produce a score of 1.
- $S_{lsl_n}(linestrengthleft)$. This score assesses how parallel the left-most gripping area is with respect to the $I_{RCW_n}yaxis$, which is consequently parallel with the left plate of the gripper. S_{lsl_n} responds to the gripping area edge strength, e.g., a corrugated edge from a bolt may score low, whereas a straight edge from a pencil might score high.
- $S_{lsr_n}(linestrengthright)$. Similar to S_{lsl_n}, this score assesses how parallel the right-most gripping area is relative to the $I_{RCW_n}yaxis$, as well as line strength.
- $S_{ptp_n}(proportiontruepixels)$. This score measures the proportion of I_{RCW_n} filled with the grasping area.
- $S_{COG_n}(COGdistancescore)$. This score relates to the distance between the centre of the candidate grasp x_{t_n}, y_{t_n} and the measured COG x_{tCOG}, y_{tCOG} of the assessed object. This score is relative to I_{BB} and will produce a 0 if x_{tCOG}, y_{tCOG} lies outside the bounding box. x_{tCOG}, y_{tCOG} lies outside the bounding box. S_{COG_n} will score 1 if x_{t_n}, y_{t_n} and x_{t_n}, y_{t_n} and x_{tCOG}, y_{tCOG} are identical.

Now that the top-view grasp hypothesis set $pose_t$ has been populated, the corresponding side-view set can be addressed:

$$pose_{s_{1,\ldots,m}} = \left\{ \begin{matrix} x_{s_1}, y_{s_1}, w_{s_1}, h_{s_1} \\ \vdots \\ x_{s_n}, y_{s_n}, w_{s_n}, h_{s_n} \\ \vdots \\ x_{s_m}, y_{s_m}, w_{s_m}, h_{s_m} \end{matrix} \right\} \tag{17}$$

where rectangle width w_{s_n} and height h_{s_n} vary based on side camera-object distance. x_s and y_s represent the side-on rectangular window pose, which relates to the gripper placement within I_s. Side camera-object distance can be measured by the top camera in the $I_t yaxis$. The relationship between the top-view rectangle y-position y_{t_n} and side rectangle width w_{s_n} is described by:

$$w_{s_n} = a y_{t_n}^2 + b y_{t_n} + c \tag{18}$$

where a, b and c are constants found to relate w_{s_n} and y_{t_n} through testing. Since the ratio of the grasping rectangle does not change and is the same as the top-view rectangle, h_{s_n} can be calculated as follows:

$$h_{s_n} = w_{s_n}\left(\frac{h_t}{w_t}\right) \tag{19}$$

Prior to object placement in I_s, a blank grayscale side-view image is captured with no objects present I_{sBlank}. A 2-dimensional Gaussian blur is performed on both I_s and I_{sBlank}, defined by:

$$GB[x,y] = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{20}$$

where σ is the standard deviation of the Gaussian distribution. The blurred side-view image containing an object is denoted by I_{sGB} and the blank variant by $I_{sBlank_{GB}}$. A side-view absolute difference image $I_{s_{diff}}$ is calculated by comparing pixel-wise intensity levels:

$$I_{s_{diff}}[x,y] = \left| I_{sGB}[x,y] - I_{sBlank_{GB}}[x,y] \right| \tag{21}$$

A binary image $I_{s_{Th}}$ of the difference image $I_{s_{diff}}$ of height H and width W is created by thresholding each pixel:

$$I_{s_{Th}}[x,y] = \begin{cases} 1, & if\ I_{s_{diff}}[x,y] \geq T_s \\ 0, & if\ I_{s_{diff}}[x,y] < T_s \end{cases} \tag{22}$$

Some noise is removed from $I_{s_{Th}}$ using morphological filters. Since the $I_t x\ axis$ and the $I_s x\ axis$ share the same plane, the top-view rectangle x-position x_{t_n} is used to calculate the side-view rectangle x-position x_{s_n} by applying a series of known transforms:

$$x_{s_n} = tr_{TC}\left(x_{t_n}\right) \tag{23}$$

A side-view rectangular window I_{SVW_n} of height h_{s_n} and width w_{s_n} is iterated through multiple y-positions along the vertical axis of I_{sTh} at x-position x_{s_n}, through the range $\left(ph_{level} - (h_{s_n}/2)\right) \leq y_{s_{step}} \leq ph_{max}$. ph_{level} is the height of the platform (i.e., the lowest point of an object, shown in Figure 8) described by the polynomial function:

$$ph_{level} = ay_{t_n}^3 + by_{t_n}^2 + cy_{t_n} + d \tag{24}$$

where a, b, c and d are constants found to relate ph_{level} and y_{t_n} through experimentation. ph_{max} is the maximum y-position of a grasp physically limited by the cover around the enclosure, which limits object height (Figure 8). At each iteration, 3 scores ranging from $0 - 1$ are calculated:

- $S_{svol_n}(side - view\ over\ lapscore)$. This metric scores the proportion of I_{SVW_n} occupied by the candidate grasping area. Higher degrees of gripper overlap will produce larger scores. If the entire gripper overlaps with the side gripping area, this score will produce a 1. Similarly, if there is no gripper overlap from this perspective, a 0 will be scored.
- $S_{hs_n}(height\ score)$. This score assesses the vertical symmetry of the grasping area within I_{SVW_n}.
- $S_{ws_n}(width\ score)$. This score assesses the horizontal symmetry of the grasping area within I_{SVW_n}.

The y_{s_n} value is taken from the iterated side-view rectangular window I_{SVW_n} as the respective $y_{s_{step}}$ with the highest combination of S_{svol_n}, S_{hs_n} and S_{ws_n}. At $pose_{t_n}$ and $pose_{s_n}$, the corresponding candidate score values S_{c_n} are collected in an array:

$$S_{c_{1,\ldots,m}} = \left\{ \begin{array}{c} K_{sf_n}, S_{ss_n}, S_{cs_{n'}} S_{lsl_n}, S_{lsr_n}, \\ S_{ptp_n}, S_{coG_n}, S_{svol_n}, S_{hs_n}, S_{ws_n} \end{array} \right\} \tag{25}$$

Collating $pose_t, pose_s$ and S_c gives the candidate grasp matrix:

$$g_{1,\ldots,m} = \left\{pose_{t1,\ldots,m}, pose_{s1,\ldots,m}, S_{c1,\ldots,m}\right\} \tag{26}$$

Two output scores OS_{pred_n} and OE_{pred_n} are predicted for each candidate grasp n by the stage-3 framework with the 10 scores from S_{c_n} as input features. This gives the selection-stage grasp set:

$$\dot{g}_{1,\ldots,m} = \left\{ \begin{array}{l} pose_{t1,\ldots,m}, pose_{s_{1,\ldots,m}}, \\ OS_{pred_{1,\ldots,m}}, OE_{pred_{1,\ldots,m}} \end{array} \right\} \tag{27}$$

A final grasp is selected for execution based on the candidate n with the highest combined OS_{pred_n} and OE_{pred_n} scores:

$$G = max\left(\dot{g}_{1,\ldots,m}\left\{OS_{pred_{1,\ldots,m}}, OE_{pred_{1,\ldots,m}}\right\}\right) \tag{28}$$

Finally, the selected grasp G is transformed from camera coordinate frames to the robot coordinate frame:

$$\bar{G}\left\{\bar{x}, \bar{y}, \bar{z}, \bar{\theta}\right\} = t_{CR}\left(G\right) \tag{29}$$

where t_{CR} is a known camera-robot space transformation which extracts robot Cartesian coordinates $\bar{x}, \bar{y}$ and robot end-effector angle $\bar{\theta}$ from $pose_{t_n}$ and robot z-value $\bar{z}$ from $pose_{s_n}$. $\bar{\theta}$ is perpendicular to the image space grasp angle θ_t. Since a 2-fingered grasp is symmetric around ±90°, $\bar{\theta}$ falls within the range $0 - 90°$.

C. GENERATING TRAINING DATA

Deep learning brings robustness at the expense of amassing large quantities of data. Noise within this data can affect the robustness of the learned algorithm.

To mitigate some of this noise we avoid human interaction where possible. Training samples are taken from real sensor data. 100 objects were chosen as the object test pool. Samples are generated exclusively from the 15 objects labelled as 'known' (Figure 10). The 'unknown' subset contains 45 unique objects that vary in size, shape and complexity (Figure 11). The remaining 40 objects have been labelled as 'etc'. This subset is shown in Figure 12 and contains items that may be similar to items within the unknown subset. The unknown and etc subsets contain novel objects, used exclusively for testing purposes. Examples of the object test pool include a USB drive, side cutters, hex key, bolt, screwdriver, wrench, pneumatic flow regulator, pneumatic T-junction, small cross wrench and an eyelet screw. To help diversity the object test pool, objects have been chosen that belong to 3 categories: general household items, tool items and component items. The complete set is shown in Figures 10, 11 and 12. Care has been taken to curate an object test set with minimal redundancy, i.e., selecting objects that sufficiently differ in shape, size, mass-distribution and grasp difficulty.

The stage-1 dataset contains images of entire objects within the known subset. Example images are shown in Figure 13. Samples are generated automatically via the same process as obtaining I_{BB}. This dataset contains 80,000 samples, split into 70% training/validation data and 30% test data. The dataset contains 5,000 images for each of the 15 known object classes. An additional 5,000 images with no objects present were added for the 16th, blank class. This class was added for two reasons. First, there may not be an object present within the classification window. Second, object-absent data was found to significantly reduce inter-class confusion—as evidenced by the generally increased precision and recall rates for classifiers trained in this way.

Samples within the stage-2 dataset are directly related to the top-down gripper placement window I_{RCW} from perspective I_t (Figure 6). Samples were manually labelled within rotated space $I_{BBR_{1,\ldots,4}}$ of size I_{RCW}. To facilitate binary classification, this dataset contains one positive class and one negative class—which relates to the traditional pass/fail metric from literature. Examples are provided in Figure 14. Each class was defined subjectively according to what the user considered a positive or negative class, with the intention

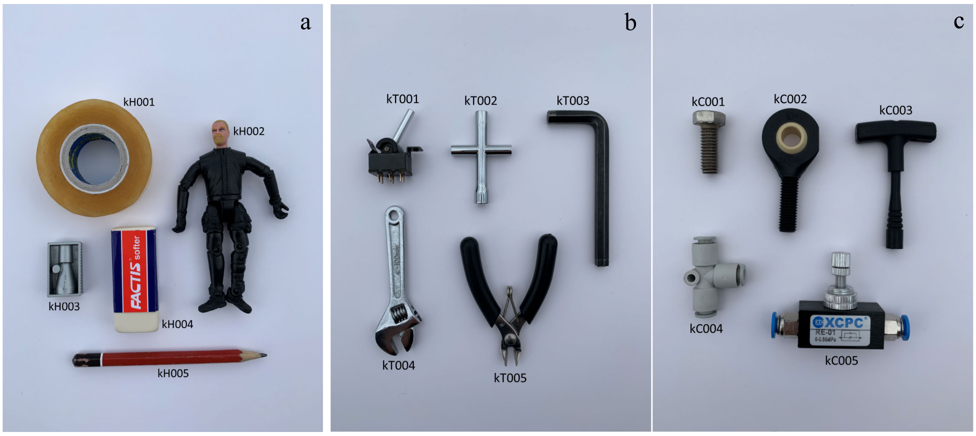

FIGURE 10. 15-object subset denoted as known. An object is considered known if it was included in any of the training datasets. (a)–known household items. (b)–known tool items. (c)–known component items.

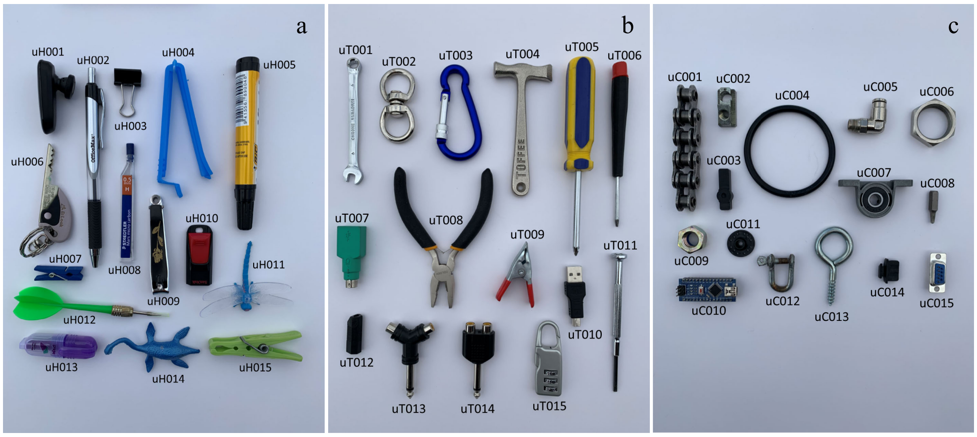

FIGURE 11. 45-object subset denoted as unknown. An object is considered unknown if it was not included in any of the training datasets. This subset is explicitly used for testing purposes. (a)–unknown household items. (b)–unknown tool items. (c)–unknown component items.

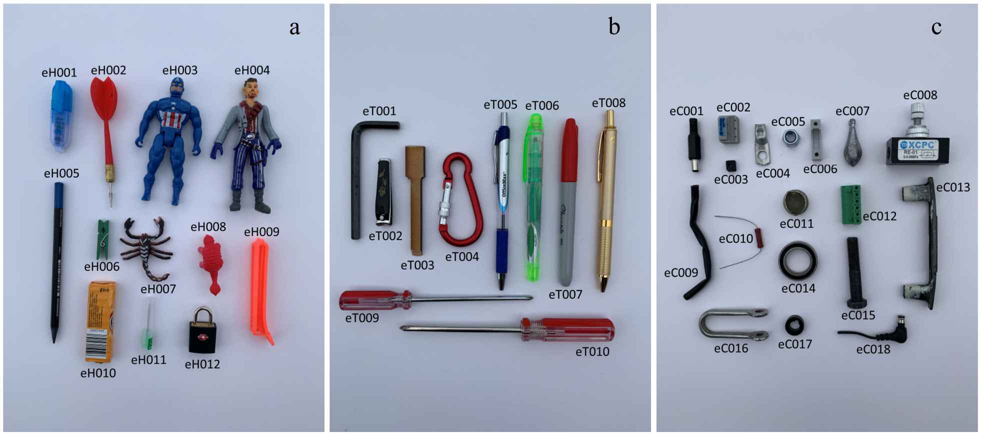

FIGURE 12. 40-object subset denoted as etc. Objects within this subset are unknown but contain object variants and objects similar in shape and size to the unknown subset. This subset is explicitly used for testing purposes. (a)–etc household items. (b)–etc tool items. (c)–etc component items.

of reflecting a location that may facilitate a successful grasp (Figure 14-b) or an unsuccessful grasp (Figure 14-a) in practice. This dataset contains 141,000 samples, of which 70% were used for training/validation and 30% for testing. 47,000 samples were dedicated to the positive class and 94,000 samples to the negative. The cost of a false

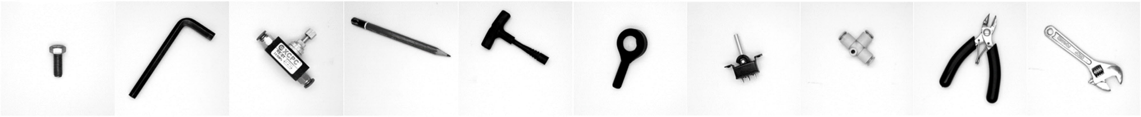

FIGURE 13. Stage-1 training data examples. Samples are generated by haphazardly placing objects within the known subset on the apparatus conveyor belt. Detected objects are moved to the vision system, which automatically locates and segments the object from its background.

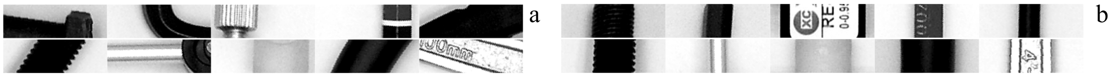

FIGURE 14. Stage-2 training data examples. (a)—negative class. (b)—positive class. Samples are generated by manual annotation according to what the user considers a successful grasp (b), or an unsuccessful grasp (a).

positive is high for the grasp detection component of the system. Therefore, we biased our dataset toward the negative class, which has been shown to reduce the number of false positives [67].

Two individual learning algorithms were trained for the stage-3 component of this methodology: one to predict the overlap score OS_{pred} and another to predict the orientation error score OE_{pred}. Both networks used the same dataset. This dataset was generated automatically by physically testing and recording the outcome of attempted grasps in terms of input features S_c and output features OS and OE for the known object subset. Note that stage-2 must be somewhat functional to generate this set. This dataset contained 2,000 samples, split into 80% training/validation data and 20% testing data for classification. The grasp attempted by the robot $\bar{G}$ was taken randomly from the candidate grasp matrix g. Generating samples for this dataset is extremely slow.

D. LEARNING TO GRASP

The stage-1 CNN framework consisted of 4 convolutions, followed by one fully connected layer. Network architecture was found through experimentation, with the goal of optimising both computational performance and dataset accuracy. In total, over 200 networks were trained prior to settling on a base architecture for this application. Each convolution was followed by a max-pool layer. Network input size was 790 × 790, as per the dimensions of I_{BB}. These dimensions were found to appropriately limit object size as per the constraints of the robotic manipulator while providing the network with sufficient information, without compromising performance. We used the Stochastic gradient descent with momentum (SGDM) optimiser. The network was trained for 15 epochs, with a mini-batch size of 30 and initial learning rate of 5×10^{-4}. 21 similar networks were also trained. The implemented network correctly classified 99.3% of its respective test set. It took 241 minutes to train using the hardware and software detailed in Table 1. The loss curve for this network is illustrated in Figure 17.

The stage-2 classifier followed a very similar CNN design as stage-1, but input size was 164 × 52. This size was set according to gripper size, corresponding to I_{RCW}. SGDM was used and the network was trained for 60 epochs with a mini-batch size of 200 and initial learning rate of 1×10^{-4}. 44 slight variants of this network were also trained. Several networks achieved roughly the same classification perfor-

TABLE 1. Implementation details.

Computer	Computer specs: NIVIDIA Quadro K2200 [70]: 4GB GDDR5 Compute capability: 3.0 Intel core i7 4790 16 GB ram Windows 10 Enterprise
Software	Matlab R2018a: Parallel computing toolbox Neural network toolbox
Robot	Dobot Magician [71]: 4 axes Maximum payload of 500g Repeatability of 0.2 mm 2-fingered gripper properties: Range: 0 – 27.5 mm Maximum force: 8N
Camera	Microsoft Livecam Studio [72]: 1920 x 1080 sensor resolution CMOS sensor 75° field of view Auto focus: 0.1 to ≥ 10m 30 FPS Automatic image adjustment with manual override
Enclosure	Diffuse LED lighting Stepper-controlled, matt-white conveyer belt 460L x 430W x 420H
Load cells	TAL220 Parallel beam load cell [73]: Rated to 10kg Load cell amplifier: HX711 ADC [74]

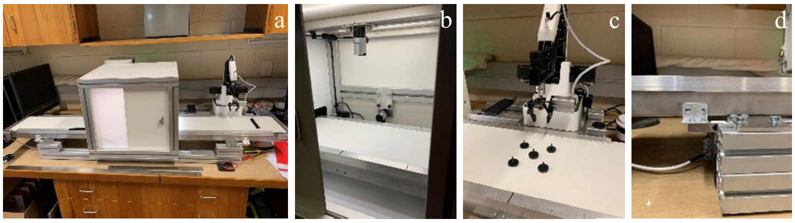

FIGURE 15. Various test rig related images. (a)—front view of test rig. (b)—view from inside the enclosure, showing two perpendicular webcams. (c)—Dobot Magician manipulator used for testing. (d)—image of one of the 10kg load cells used to find the COG of an object.

mance, but the smallest network was implemented. It was found to correctly classify 98.9% of the stage-2 test set. This network took 145 minutes to train. The loss curve is shown in Figure 18. The stage-2 classifier is rotation variant. As shown by Sun*et al.* [55], better performance can be achieved by rotating the image prior to classification—as opposed to training a rotation invariant algorithm. Alternatively, Pinto and Gupta [24] classify graspable patches, then additionally classify for angle.

Initially we attempted to cast the stage-3 component as a classifier through thresholds. 5 classes were defined for each network by grouping the measured output features OS and OE by range $0-0.2, 0.2-0.4, 0.4-0.6, 0.6-0.8, 0.8-1.0$. Unfortunately, this approach performed poorly. 30 various classifiers were trained in this way, with the top performing network achieving a mere 63% classification accuracy.

Framing stage-3 as a regression problem yielded better results. The stage-3 OS_{pred} network utilised a regression SVM. 5-fold cross validation was used. The best network predicted OS with input S_c with a root mean square error (RMSE) of 0.07. Simple linear regression was used for the stage-3 OE_{pred} network, resulting in an RMSE of 0.14. Other regression algorithms were also trained, including regression trees, Gaussian process regression models and other ensemble learners.

V. EVALUATION

A. HARDWARE SET-UP

Methodology evaluation and testing was conducted using the apparatus shown in Figure 15. The test rig consists of a conveyor belt, robotic system and imaging enclosure. Two Microsoft Livecam Studio webcams were installed to capture images of the object within the enclosure and diffuse LED lighting was used to maintain a consistent imaging environment. The conveyor belt rested on 4 load cells, rated to 10kg each. Hardware and software details are provided in Table 1.

Measured system properties:

- Conveyor belt repeatability: ±3.4 *pixels*($\sim$ 0.5 *mm*)
- Combined load cell error: ±0.1 *g*
- Robot-vision transformation error: ±1.7 *pixels* ($\sim$ 0.24 *mm*)
- Calibration disc weight: $133g \pm 0.01$ *g*
- COG position error: ±4.1 *pixels* ($\sim$ 0.6 *mm*) @200 *g*

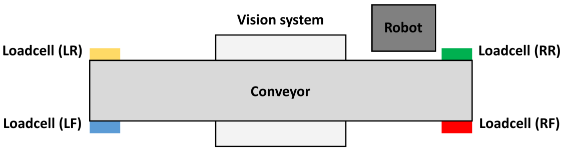

FIGURE 16. Load cell configuration and layout of test rig.

The COG of an object x_{tCOG}, y_{tCOG} within top-view image space I_t is related to the output of each load cell. The interaction between these planes is illustrated in Figure 19.

To facilitate the transform between load cell space and image space, *coeffx* and *coeffy* are calculated:

$$coeffx = \frac{weight_{RF} + weight_{RR}}{weight_{total}} \tag{30}$$

$$coeffy = \frac{weight_{RF} + weight_{LF}}{weight_{total}} \tag{31}$$

$$\begin{aligned} weight_{total} &= weight_{LR} + weight_{RR} \\ &\quad + weight_{LF} + weight_{RF} \end{aligned} \tag{32}$$

where $weight_{LR}$, $weight_{RR}$, $weight_{LF}$ and $weight_{RF}$ are load cell measurements corresponding to the locations in Figures 16 and 19. *coeffx* and *coeffy* effectively describe the ratio of weight distribution in two perpendicular axes. These coefficients can be used to devise linear relationships that relate x_{tCOG}, y_{tCOG} and *coeffx*, *coeffy*:

$$x_{tCOG} = Acoeffx + B \tag{33}$$

$$y_{tCOG} = Ccoeffy + D \tag{34}$$

where A, B, C and D are constants found through experimentation. For a more in-depth description of this method, please refer to Patel and Topiwala [75].

B. EXPERIMENTAL PROTOCOL

To test the proposed methods and methodologies, single objects were placed haphazardly at the end of the conveyor belt furthest from the robotic manipulator. Detected objects were automatically moved to the centre of the vision system along I_tx *axis*. The length of an object was approximated using a beam sensor and known distance travelled by the conveyor belt. The object length approximation accuracy was measured as $\pm 1mm$ and could be used to move an object to a desired I_t location along the *xaxis* only. Once imaged,

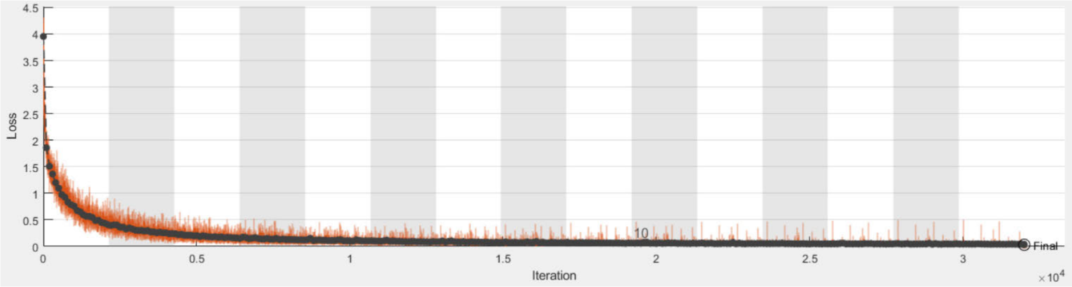

FIGURE 17. Loss curve for the stage-1 network used to classify object images. This network was trained for 15 epochs and reached a validation accuracy of 99.3%.

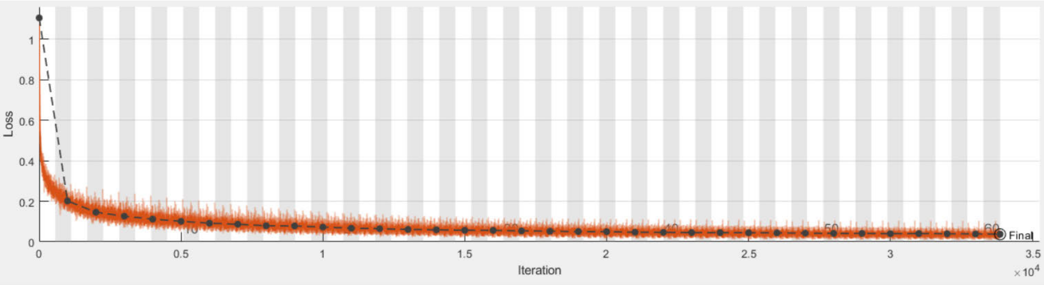

FIGURE 18. Loss curve for the stage-2 network used to classify grasping rectangles. This network was trained for 60 epochs and reached a validation accuracy of 98.9%.

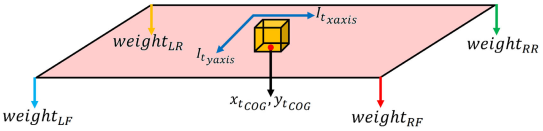

FIGURE 19. Graphical depiction of load cell arrangement. $x_{t_{COG}}$, $y_{t_{COG}}$ can be calculated in I_t image space by measuring load cell outputs.

the object is moved by a set amount into the robot frame. A grasp is then attempted. For comparison we take our working, literature-based definition of a successful attempt to be a physically performed grasp in which an object is lifted vertically to a height of 15 cm and then held stationary for at least 10 seconds without falling. The object must be solely supported by the gripper, so that no other part of the object touches any other hardware. After being suspended for 10 seconds, the object is placed on the conveyor belt at the grasp attempt location and moved by the set amount back toward the centre of our vision system, where the resultant OS and OE scores are measured. This experimental evaluation aims to quantitatively assess the effectiveness of the methodology by measuring the error introduced by the grasping process itself in terms of OS and OE. The outcome was also recorded in terms of the pass/fail metric. The above process constitutes one trial.

The system was physically trialled 3,000 times on the entire object test pool consisting of the 15 known objects used to generate training datasets and the remaining 85 objects not seen during training. Each object was trialled 10 times based on 3 different selection criteria. 1,000 trials were conducted using the stage-2 network, in which a grasp was selected at random from grasp matrix g. 1,000 trials were also conducted by selecting the grasp with the highest K_{sf} value produced by the stage-2 network. The remaining 1,000 trials were conducted using the complete system, in which stage-3 automatically chose and attempted a grasp from g using input features S_c. Results are shown in Figure 20. 3,000 trials corresponds to approximately 100 robot-hours.

VI. RESULTS AND DISCUSSION

As stated in Section 2.2, it is difficult to determine the relative performance of this research due to the lack of clear benchmarks. However, the comparison of our method to other open-loop methodologies that grasp a single object in isolation shows improvement. An object is considered isolated if it does not touch another object in the field of view. To aid the comparison, common 'household' and 'laboratory' items that are similar in shape and size to objects used by other methodologies reported in literature and within the capability of our robot were selected.

A. QUANTITATIVE ASSESSMENT

The grasp success rate of the proposed methodology over 3,000 trials is illustrated in the form of a histogram in Figure 20-a. Grasping via random selection from g successfully lifted objects within the object test pool to a height of 15 cm for 10 seconds, 94.0% of the time. This could be increased to 96.4% by selecting the grasp within g with the largest K_{sf} value. This was further improved to 99.0% by

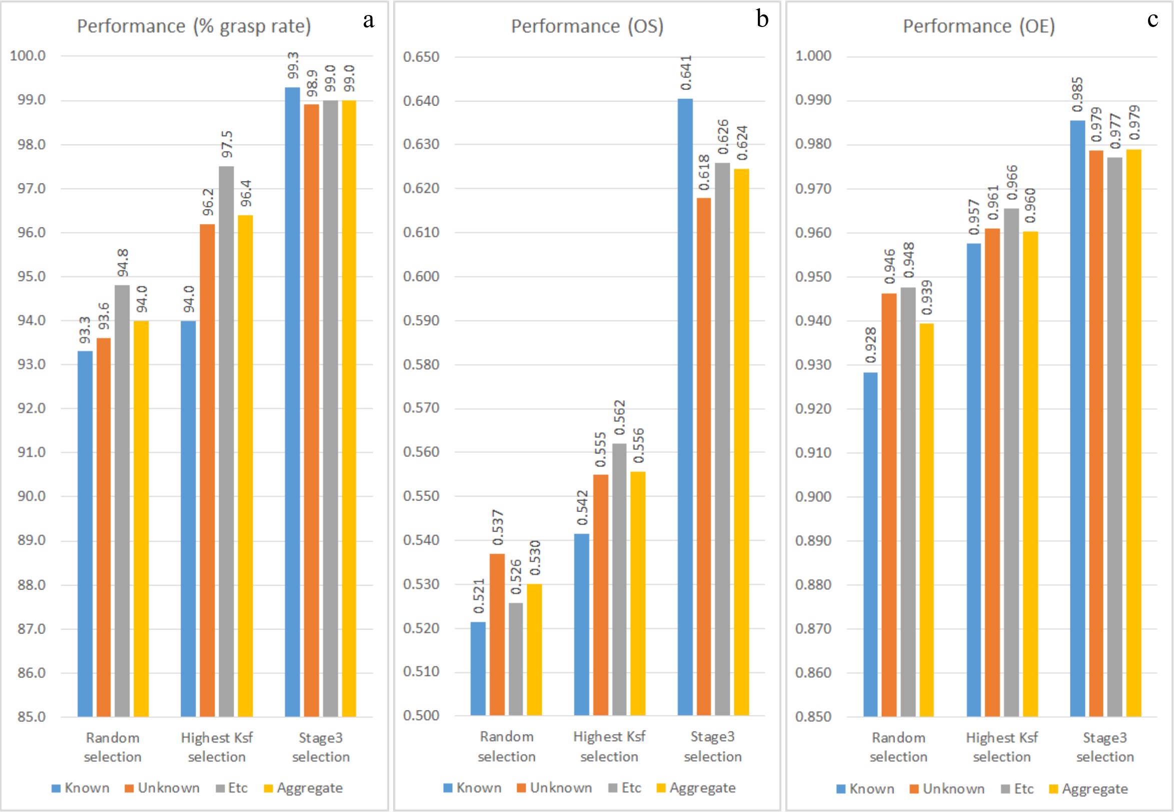

FIGURE 20. Measured results from 3,000 grasp attempts by selection criteria in terms of the pass/fail metric (a), the OS metric (b) and the OE metric (c).

selecting the grasp from g with the largest combined OS_{pred} and OE_{pred} scores, predicted by the two stage-3 networks. Note that the reported grasp rates are defined in terms of the literature-based definition of a successful grasp, as defined in Section 5.2.

The grasp rate via random selection for the known object subset was 93.3%, increasing slightly for the unknown and etc subsets, tested to grasp objects at a rate of 93.6% and 94.8%, respectively. Highest K_{sf} selection increased grasp rates for the known subset to 94%. This grasp strategy performed notably better for objects within the unknown and etc subsets, with grasp rates of 96.2% and 97.5%, respectively. Note the gradual increase in grasp rate for the known, unknown and etc subsets for random and highest K_{sf} selection (Figure 20-a). This increase is consistent across selection criteria and may be related to the difficulty of each respective object subset.

Selection via stage-3 produced the least amount of variability across each object subset, grasping objects within the known, unknown and etc subsets successfully 99.3%, 98.9% and 99% of the time, respectively.

A comparison of the performance of the proposed methodology vs. the performance of others is shown in Table 2. The grasping success rate for known and unknown objects is improved by approximately 2 and 10%, respectively, compared to the best performance of other similar approaches.

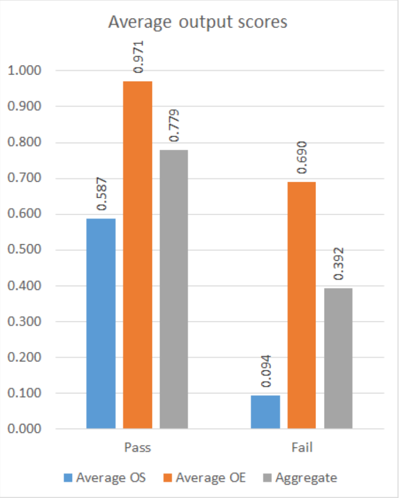

FIGURE 21. Measured output scores averaged over 3,000 grasp attempts.

Relative to the highest K_{sf} aggregate baseline, grasping performance increased by roughly 3% when selecting grasps via highest combined predicted OS and OE scores. Although this improvement can be quantified by defining an attempted grasp as either successful or unsuccessful, grasp quality seems to be more accurately reflected by our metrics. This was particularly evident for the key object (uH006).

TABLE 2. Cited grasping success rates in terms of the pass/fail metric of various 2-fingered gripper-based works, where objects were physically trialled in isolation. Number of trials per object and testing conditions varied. Most of the referenced works make use of some form of deep learning–please refer to the 'Approach' column.

Authors	Approach	Grasping success rate (%)		
		Known objects	Unknown objects	Aggregate
Sun et. al. [55]	RGB-D supervised learning	86.0	86.0	86.0
Johns et al. [66]	RGB-D supervised learning	-	80.3	-
Jiang et. al. [54]	RGB-D supervised learning	-	87.9	-
Pinto et. al. [24]	RGB unsupervised learning	73.0	66.0	69.5
Pinto et. al. [25]	RGB adversarial learning	-	82.0	-
Kopicki et. al. [38]	RGB-D point cloud learning	88.0	86.0	87.0
Arruda et. al. [50]	RGB-D active vision	-	80.4	-
Saxena et. al. [5]	RGB logistic regression	90.0	87.8	88.9
Chu et al. [60]	RGB-D supervised learning	-	**89.0**	-
Viereck et. al. [40]	RGB-D supervised learning	**97.5**	-	-
Mahler et al. [39]	RGB-D supervised learning	93.0	80.0	86.5
ten Pas et. al. [37]	RGB-D supervised learning	-	88.0	-
Proposed 3-stage methodology	Random candidate selected	93.3	93.6	94.0
	Highest K_{sf}	94.0	96.2	96.4
	Selected by stage-3 framework	**99.3**	**98.9**	**99.0**

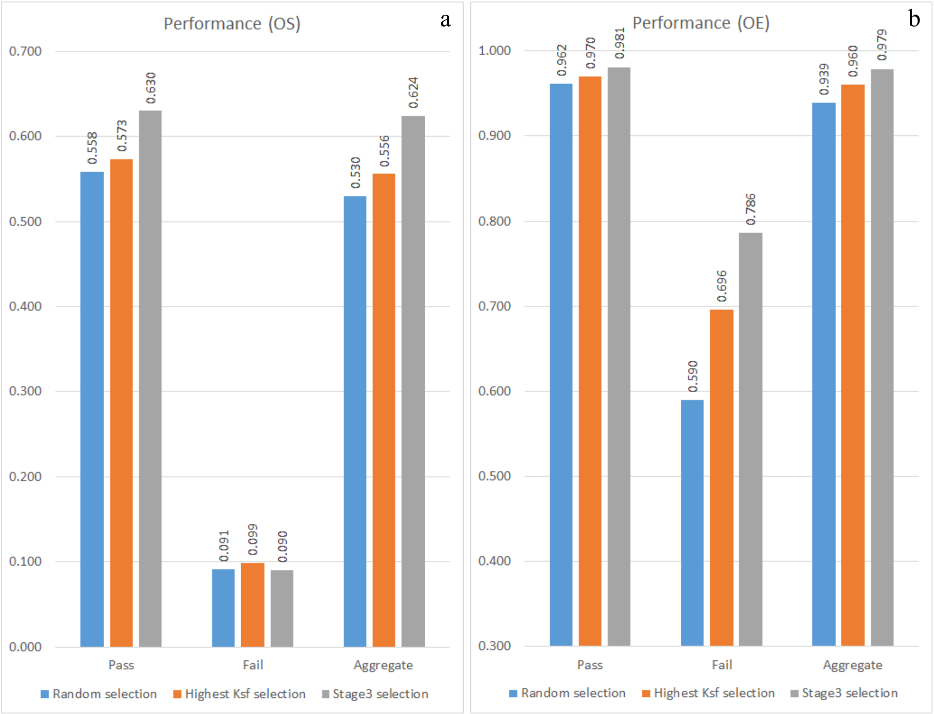

FIGURE 22. (a)–average **OS** response from samples that produce a pass or fail outcome, measured from 3,000 grasp attempts. (b)–average **OE** response from samples that produce a pass or fail outcome, measured from 3,000 grasp attempts.

During trials, this object was frequently grasped by the chain interlinked to the main body of the object (Figure 25-d), generally resulting in a successful grasp, but very poor *OS* and *OE* scores. Moreover, the uneven mass distribution of the wrench object (kT004) tended to produce lower *OS* and *OE* scores. Top-scoring K_{sf} grasps tended toward the handle end of this object, generally far from the COG—resulting in significant droop and low-quality grasps. Although this was qualitatively noticeable, the pass/fail metric did not quantify this error, as all trials for this object were considered

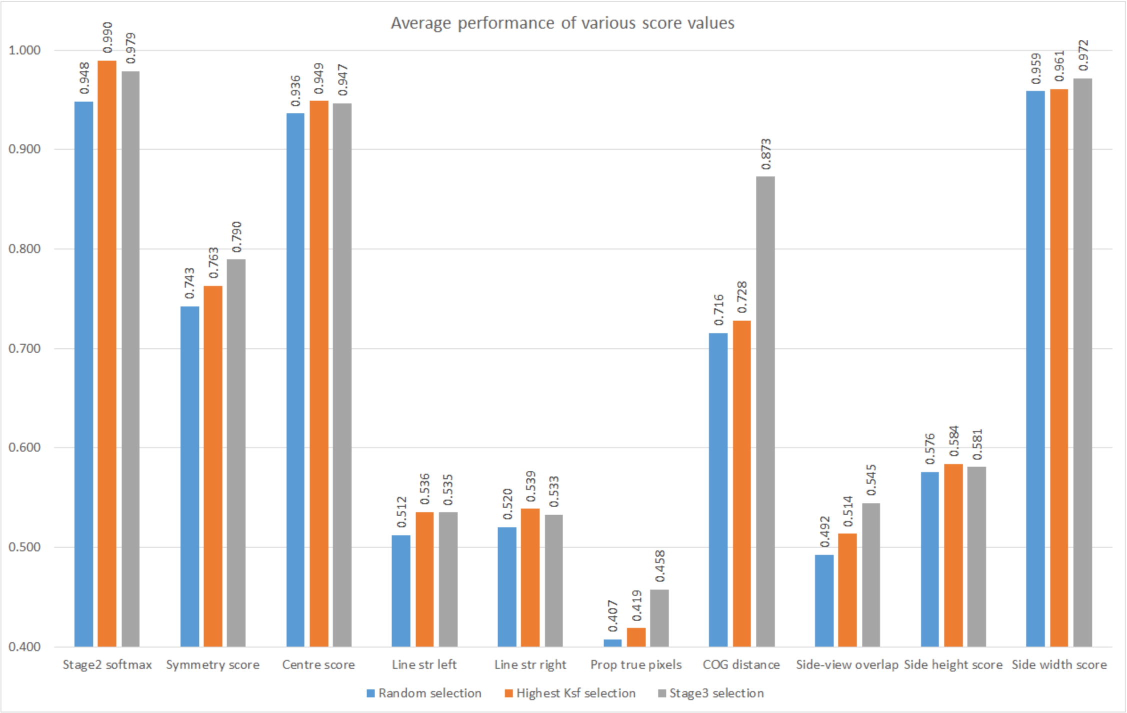

FIGURE 23. Performance of various score values from $\mathbf{S_C}$ by selection criteria, measured from 3,000 grasp attempts.

successful for all 3 modes of selection. This error was significantly mitigated by the stage-3 selection process which tended toward grasps close to the COG. Generally, grasping away from the COG of an object tended to amplify the introduced rotational and translational error. The average response of OS and OE across 3,000 grasp attempts is shown in Figure 20-b and Figure 20-c, respectively. Note OS and OE fluctuate with grasp success rate.

Figure 22 illustrates the response of OS and OE based on grasp outcome. Resultant OE scores from Figure 22-b show that very little orientation error was introduced when a grasp attempt was successful.

Compared to random selection, highest K_{sf} selection tended to increase the outcome of both OS and OE for successful grasps. This was further improved by stage-3 selection. Notably, OE scores for failed grasps differed based on the selection criteria (Figure 22-b). Even though a grasp attempt may have failed, the introduced orientational error was lower when selecting via stage-3. Generally, stage-3 selection produced the highest quality grasps in terms of both OS and OE for grasps considered successful. OS scored very low for failed grasps, although this score was somewhat consistent across selection criteria.

The average OS and OE scores for a failed grasp were 0.094 and 0.690, respectively (Figure 21). The average OS and OE scores for a successful grasp were 0.587 and 0.971, respectively. This suggests a relationship between the literature definition of a successful/unsuccessful grasp and the proposed metrics.

In addition to recording the resultant OS and OE scores from 3,000 physical grasps, the related grasping window scores S_c were also recorded—illustrated in Figure 23. These scores serve as inputs to the stage-3 selection networks. Note that the COG distance score was significantly higher for grasps selected via stage-3. This is consistent with the qualitative assessment in Section 6.2, as grasps tended toward the COG of an object for this mode of selection. Moreover, selection via the stage-3 framework tended toward grasps that were more symmetric, scored higher in terms of proportion of graspable area from the top camera perspective and tended to select grasps with increased gripper overlap from the side camera perspective.

TABLE 3. Run-time per component.

Component	Time to compute
Object segmentation	$4.7 \times 10^{-2}\ s$
Stage-1 network	$3.7 \times 10^{-3}\ s$
Sobel operation	$7.9 \times 10^{-2}\ s$
Stage-2 network	$3.8 \times 10^{-4}\ s$
Stage-3 network (OS_{pred})	$7.9 \times 10^{-5}\ s$
Stage-3 network (OE_{pred})	$4.9 \times 10^{-5}\ s$
Overall system	$1.28\ s$

The computation time for each major component at trial time is shown in Table 3. The methodology produced, on average, 82.7 candidate grasps per object of $K_{sf_n} \geq T_{sf}$ at $T_{sf} = 0.95$, with $I_{BBR_{1,\dots,4}}[i,j]$ step magnitudes of $i = 5$ and $j = 5$ and $y_{s_{step}=5}$. These step magnitudes resulted in the system operating at approximately 0.8 Hz. The search resolution $I_{BBR_{1,\ldots,4}}[i,j]$ and $y_{s_{step}}$ were the biggest contributors to computation time. Lowering this resolution significantly increased performance but reduced the number of candidate grasps generated. Alternatively, limiting the number of Sobel rotations θ_{Sobel} considered also increased performance at the cost of reducing g count. The object segmentation and Sobel operation components were accelerated using GPU-array computation. Computation time varied depending on object. Hardware acquisition times, e.g., camera frame capture and load cell measurement, were the largest contributor to total system operation time.

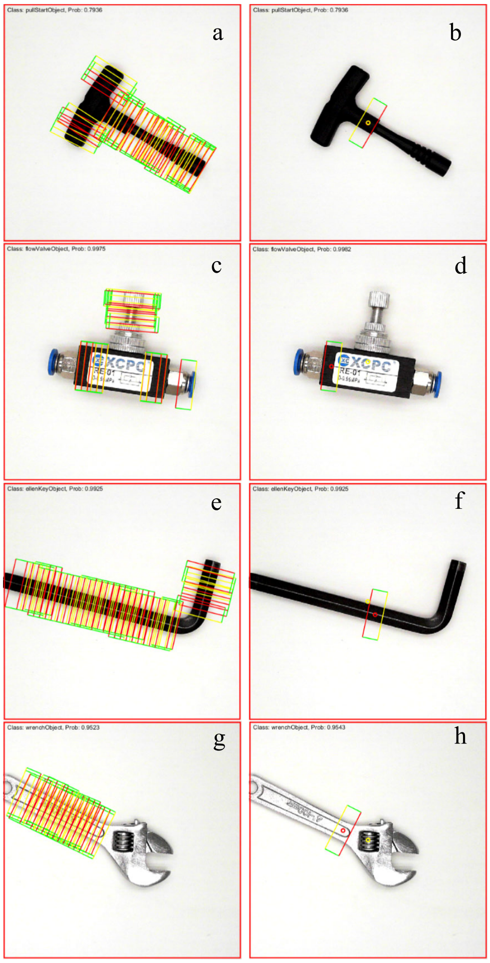
Class: pullStartObject, Prob: 0.7936
a
Class: pullStartObject, Prob: 0.7936
b
Class: flowValveObject, Prob: 0.9975
c
Class: flowValveObject, Prob: 0.9982
d
Class: allenKeyObject, Prob: 0.9925
e
Class: allenKeyObject, Prob: 0.9925
f
Class: wrenchObject, Prob: 0.9523
g
Class: wrenchObject, Prob: 0.9543
h

FIGURE 24. Candidate grasp pool generated through classification by the stage-2 network (left) and the corresponding grasp selected by the stage-3 framework (right).

B. QUALITATIVE ASSESSMENT

It was apparent from the first few trials that stage-3 tended toward grasps close to the COG of objects. Some examples are shown in Figure 24. These sorts of grasps do not generally result in droop—which affects peri-grasp object manipulation quality and post-grasp placement quality. Placement quality is measured by OS and OE.

Selected grasps tended to have larger gripping surfaces and clear, parallel gripping areas. 'Good' grasps tended to score OS and OE values higher than 0.98 and 0.70, respectively, whereas 'bad' grasps tended to produce lower scores. A grasp was considered 'bad' if there was noticeable droop and/or object displacement caused by the gripper. A 'bad' grasp did not necessarily result in a dropped object and was therefore not reflected by the literature definition of a successful/unsuccessful grasp.

Generally, false positives produced by stage-2 were not selected by stage-3. The right-most grasping rectangle in Figure 24-c for the flow valve object (kC005), for example, is likely to produce droop or fail the grasp altogether.

C. COMMON REASONS FOR FAILING TO GRASP AN OBJECT

Some common reasons for failure were observed. The flow-valve object (kC005), for example, was prone to grasps extremely close to other perpendicular surfaces not visible in I_{RCW}. Due to the uncertainty associated with the system, sometimes the gripper would collide with such surfaces, resulting in failure or decreased grasp quality.

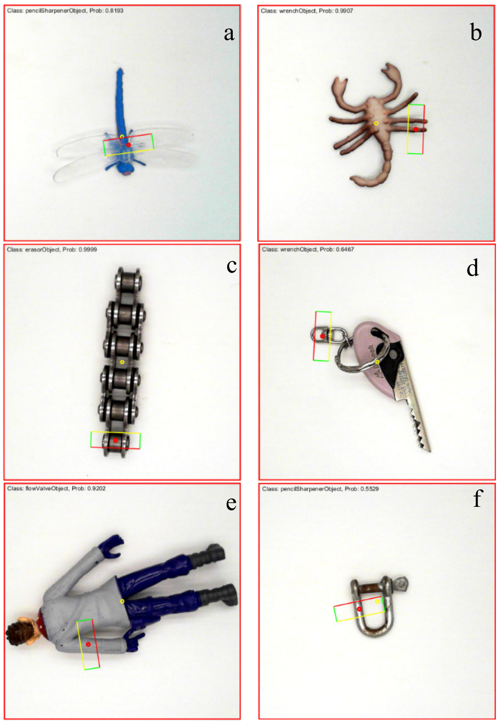
Class: pencilSharpenerObject, Prob: 0.8193
a
Class: wrenchObject, Prob: 0.9907
b
Class: erasorObject, Prob: 0.9999
c
Class: wrenchObject, Prob: 0.6467
d
Class: flowValveObject, Prob: 0.9202
e
Class: pencilSharpenerObject, Prob: 0.5529
f

FIGURE 25. Candidate grasps selected by various selection criteria that will likely result in an unsuccessful grasp or score poorly in terms of OS and OE.

Since I_{RCW} does not capture the physical dimensions of the gripper outside of this window, some grasps were prone to collision in certain situations. Generally, this resulted in lower OS scores, but did not cause the grasp attempt to fail. This was particularly relevant for the small U lock object (uC012)—Figure 25-d.

Deformable objects rarely caused grasp attempts to fail, but usually reduced OS and OE. Examples of deformable objects include the black figurine object (kH002), the brown scorpion toy (eH007), the red Euoplocephalus toy (eH008) and the blue Plesiosaur toy (uH014). Moreover, the brown scorpion toy sometimes suffered from both deformation and gripper edge collision. An example of this situation is illustrated in Figure 25-b.

Many grasps selected via random selection or highest K_{sf} selection failed or performed poorly in terms of OS and OE due to a low COG distance score S_{COG}. The distance between an attempted grasp and the COG of an object becomes increasingly important for heavier objects such as the black bolt (eC015), the hex key objects (eT001, kT003), the screwdriver objects (uT005, eT009, eT010), the wrench object (kT004), the small side cutters (kT005), the motorcycle chain (uC001), the small pliers (uT008) and the toffee hammer (uT004). Droop is the main factor that contributed to poor grasps in this instance. An example of this is shown in Figure 25-c. Generally, this was not an issue for stage-3 selection.

The largest contributor to failed grasp attempts was the selection of incorrectly classified grasps. False positives often caused major collisions (Figure 25-e) or translated the object significantly without fully closing the gripper onto the object (Figure 25-a). Usually, false positives performed poorly in terms of the related grasping window scores S_c and were thus avoided by the stage-3 framework altogether.

An accurate grasp relies on an accurate translation between the vision and robot coordinate frames—which are separated by the conveyor belt. Conveyor translation may introduce a small, yet significant amount of error for very light and rigid objects. This error contributed to a very small percentage of failed grasps.

VII. LIMITATIONS AND FUTURE WORK

Currently, the proposed methodology is only applicable to systems with small 2-fingered grippers. However, since only the soft-max output of the stage-2 network K_{sf} is used, custom networks that correspond to the new gripper size and design can be substituted. Scale-invariant learning approaches should ideally be used for this stage.

The conducted trials were only performed on objects in isolation due to the limitations imposed on the system by the employed COG measurement technique. Removing COG as an input feature can still produce good grasping performance—shown in Figure 20 by the pre-stage-3 performance of the proposed methodology. The COG feature is regarded as one of many potential inputs that provide the proposed methodology with information to improve grasp prediction. It is intended that many varying sensors or known object properties can serve as inputs if they are available, such as temperature, material properties, friction coefficients, gripper feedback, weight, etc. Future work aims to extend the methodology to cluttered environments by approximating the COG of an object through vision. Note that this method assumes uniform mass distribution. In this work, object centroid coordinates x_o, y_o represent a vision-estimated COG. Therefore, true COG measurements may be avoided by simply substituting x_{tCOG}, y_{tCOG} for x_o, y_o, respectively.

The output from stage-1 is not currently used. The aim is to build a database of successful grasps, where unique objects are linked to their grasp outcome. By identifying an object which has previously been well-grasped, stage-2 and stage-3 can be avoided by fitting the known grasp to the known object directly.

Improving the quality and quantity of our stage-3 dataset is another direction for future work. The current dataset is small and noisy.

Single-pass networks have gained significant popularity recently due to their increased computational performance [26], [56], [57], [76], [77]. It is possible to merge the proposed networks into a large, single-pass network which predicts a final grasp G directly from an input image I_t. Of course, a more powerful GPU is required to facilitate this.

Relevant source code and the datasets used in this work will be made available at: https://drive.google.com/open?id =1VsEjCl6hrX3FeL9VRF-J9CzVL7JHXO15.

VIII. CONCLUSION

In this paper we presented a novel 3-stage, learning-based object grasping approach that predicts a well-suited, 2-fingered grasping location for previously unseen objects. The outcome showed an increase in performance of roughly 10% for unknown objects, compared to other similar works. The proposed similarity metrics demonstrated applicability for learning-based algorithm training. 3,000 physical trials revealed a 3% improvement in grasp rate when employing the proposed metrics compared to the more traditional pass/fail grasping definitions. To the best of our knowledge, our methodology is the first to use a COG factor as an input feature for learning.

Our work aims toward a robust methodology to not only grasp objects well, but to grasp objects so that they may be handled well. We believe this is key for industry applications and needed to further close the gap between manual and fully automated assembly. In future work, we would like to improve the fit of the stage-3 networks, tune our methodology to a wider range of objects and extend our methodology to cluttered environments.

REFERENCES

[1] E. Klingbeil, D. Rao, B. Carpenter, V. Ganapathi, A. Y. Ng, and O. Khatib, "Grasping with application to an autonomous checkout robot," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 2837–2844, doi: 10.1109/ICRA.2011.5980287.

[2] C. Zhihong, Z. Hebin, W. Yanbo, L. Binyan, and L. Yu, "A vision-based robotic grasping system using deep learning for garbage sorting," in *Proc. 36th Chin. Control Conf. (CCC)*, Jul. 2017, pp. 11223–11226, doi: 10.23919/ChiCC.2017.8029147.

[3] A. Ramisa, G. Alenya, F. Moreno-Noguer, and C. Torras, "Using depth and appearance features for informed robot grasping of highly wrinkled clothes," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 1703–1708.

[4] D. Seita, N. Jamali, M. Laskey, A. Kumar Tanwani, R. Berenstein, P. Baskaran, S. Iba, J. Canny, and K. Goldberg, "Deep transfer learning of pick points on fabric for robot bed-making," 2018, *arXiv:1809.09810*. [Online]. Available: http://arxiv.org/abs/1809.09810

[5] A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision," *Int. J. Robot. Res.*, vol. 27, no. 2, pp. 157–173, Feb. 2008.

[6] M. Beetz, U. Klank, I. Kresse, A. Maldonado, L. Mosenlechner, D. Pangercic, T. Ruhr, and M. Tenorth, "Robotic roommates making pancakes," in *Proc. 11th IEEE-RAS Int. Conf. Humanoid Robots*, Oct. 2011, pp. 529–536.

[7] *The Moley Robotic Kitchen–Mission & Goals*, Moley Robot., London, U.K., 2015.

[8] C. Militaru, A.-D. Mezei, and L. Tamas, "Object handling in cluttered indoor environment with a mobile manipulator," in *Proc. IEEE Int. Conf. Autom., Qual. Test., Robot. (AQTR)*, May 2016, pp. 1–6, doi: 10.1109/AQTR.2016.7501382.

[9] K. Wu, R. Ranasinghe, and G. Dissanayake, "Active recognition and pose estimation of household objects in clutter," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 4230–4237.

[10] W. Miyazaki and J. Miura, "Object placement estimation with occlusions and planning of robotic handling strategies," in *Proc. IEEE Int. Conf. Adv. Intell. Mechatronics (AIM)*, Jul. 2017, pp. 602–607, doi: 10.1109/AIM.2017.8014083.

[11] S. Levine, N. Wagener, and P. Abbeel, "Learning contact-rich manipulation skills with guided policy search," 2015, *arXiv:1501.05611*. [Online]. Available: http://arxiv.org/abs/1501.05611

[12] L. Manuelli, W. Gao, P. Florence, and R. Tedrake, "KPAM: KeyPoint affordances for category-level robotic manipulation," 2019, *arXiv:1903.06684*. [Online]. Available: http://arxiv.org/abs/1903.06684

[13] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *Int. J. Robot. Res.*, vol. 34, nos. 4–5, pp. 705–724, Apr. 2015.

[14] 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*. [Online]. Available: http://arxiv.org/abs/1809.10790

[15] M. Gualtieri, A. T. Pas, K. Saenko, and R. Platt, "High precision grasp pose detection in dense clutter," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 598–605, doi: 10.1109/IROS.2016.7759114.

[16] M. Gualtieri, A. T. Pas, and R. Platt, "Pick and place without geometric object models," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 7433–7440, doi: 10.1109/ICRA.2018.8460553.

[17] A. Zeng, K.-T. Yu, S. Song, D. Suo, E. Walker, A. Rodriguez, and J. Xiao, "Multi-view self-supervised deep learning for 6D pose estimation in the amazon picking challenge," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 1383–1386, doi: 10.1109/ICRA.2017.7989165.

[18] M. Schwarz, C. Lenz, G. M. Garcia, S. Koo, A. S. Periyasamy, M. Schreiber, and S. Behnke, "Fast object learning and dual-arm coordination for cluttered stowing, picking, and packing," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 3347–3354, doi: 10.1109/ICRA.2018.8461195.

[19] A. Zeng *et al.*, "Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 1–8, doi: 10.1109/ICRA.2018.8461044.

[20] D. Morrison *et al.*, "Cartman: The low-cost Cartesian manipulator that won the Amazon robotics challenge," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 7757–7764.

[21] Amazon Robotics LLC. *2017 Amazon Robotics Challenge Official Rules*. Accessed: Jun. 24, 2019. [Online]. Available: https://www.amazonrobotics.com/site/binaries/content/assets/amazonrobotics/arc/2017-amazon-robotics-challenge-rules-v3.pdf

[22] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, and S. Levine, "QT-opt: Scalable deep reinforcement learning for vision-based robotic manipulation," 2018, *arXiv:1806.10293*. [Online]. Available: http://arxiv.org/abs/1806.10293

[23] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *Int. J. Robot. Res.*, vol. 37, nos. 4–5, pp. 421–436, Apr. 2018.

[24] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50K tries and 700 robot hours," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2016, pp. 3406–3413, doi: 10.1109/ICRA.2016.7487517.

[25] L. Pinto, J. Davidson, and A. Gupta, "Supervision via competition: Robot adversaries for learning tasks," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 1601–1608.

[26] D. Morrison, P. Corke, and J. Leitner, "Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach," 2018, *arXiv:1804.05172*. [Online]. Available: http://arxiv.org/abs/1804.05172

[27] A. Zeng, S. Song, S. Welker, J. Lee, A. Rodriguez, and T. Funkhouser, "Learning synergies between pushing and grasping with self-supervised deep reinforcement learning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 4238–4245.

[28] N. Keddis, G. Kainz, C. Buckl, and A. Knoll, "Towards adaptable manufacturing systems," in *Proc. IEEE Int. Conf. Ind. Technol. (ICIT)*, Feb. 2013, pp. 1410–1415, doi: 10.1109/ICIT.2013.6505878.

[29] A. R. Mileham, S. J. Culley, G. W. Owen, and R. I. McIntosh, "Rapid changeover–a pre-requisite for responsive manufacture," in *Proc. IEE Workshop Responsiveness Manuf.*, Feb. 1998, pp. 1–12, doi: 10.1049/ic:19980103.

[30] R. Patel, M. Hedelind, and P. Lozan-Villegas, "Enabling robots in small-part assembly lines: The 'ROSETTA approach'—An industrial perspective," in *Proc. 7th German Conf. Robot. (ROBOTIK)*, May 2012, pp. 1–5.

[31] R. McLean, A. J. Walker, and G. Bright, "An artificial neural network driven decision-making system for manufacturing disturbance mitigation in reconfigurable systems," in *Proc. 13th IEEE Int. Conf. Control Autom. (ICCA)*, Jul. 2017, pp. 695–700, doi: 10.1109/ICCA.2017.8003144.

[32] D. Sanderson, J. C. Chaplin, L. De Silva, P. Holmes, and S. Ratchev, "Smart manufacturing and reconfigurable technologies: Towards an integrated environment for evolvable assembly systems," in *Proc. IEEE 1st Int. Workshops Found. Appl. Self* Syst. (FAS*W)*, Sep. 2016, pp. 263–264, doi: 10.1109/FAS-W.2016.61.

[33] R. Frei, L. Ribeiro, J. Barata, and D. Semere, "Evolvable assembly systems: Towards user friendly manufacturing," in *Proc. IEEE Int. Symp. Assem. Manuf.*, Jul. 2007, pp. 288–293, doi: 10.1109/ISAM.2007.4288487.

[34] S. Kock, T. Vittor, B. Matthias, H. Jerregard, M. Kallman, I. Lundberg, R. Mellander, and M. Hedelind, "Robot concept for scalable, flexible assembly automation: A technology study on a harmless dual-armed robot," in *Proc. IEEE Int. Symp. Assem. Manuf. (ISAM)*, May 2011, pp. 1–5, doi: 10.1109/ISAM.2011.5942358.

[35] M. Hedelind and S. Kock, "Requirements on flexible robot systems for small parts assembly: A case study," in *Proc. IEEE Int. Symp. Assem. Manuf. (ISAM)*, May 2011, pp. 1–7, doi: 10.1109/ISAM.2011.5942356.

[36] A. T. Pas, M. Gualtieri, K. Saenko, and R. Platt, "Grasp pose detection in point clouds," *Int. J. Robot. Res.*, vol. 36, nos. 13–14, pp. 1455–1473, Dec. 2017.

[37] A. T. Pas and R. Platt, "Using geometry to detect grasps in 3D point clouds," 2015, *arXiv:1501.03100*. [Online]. Available: http://arxiv.org/abs/1501.03100

[38] M. Kopicki, R. Detry, F. Schmidt, C. Borst, R. Stolkin, and J. L. Wyatt, "Learning dexterous grasps that generalise to novel objects by combining hand and contact models," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2014, pp. 5358–5365, doi: 10.1109/ICRA.2014.6907647.

[39] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," 2017, *arXiv:1703.09312*. [Online]. Available: http://arxiv.org/abs/1703.09312

[40] U. Viereck, A. ten Pas, K. Saenko, and R. Platt, "Learning a visuomotor controller for real world robotic grasping using simulated depth images," 2017, *arXiv:1706.04652*. [Online]. Available: http://arxiv.org/abs/1706.04652

[41] J. J. van Vuuren, L. Tang, I. Al-Bahadly, and K. Arif, "Towards the autonomous robotic gripping and handling of novel objects," in *Proc. 14th IEEE Conf. Ind. Electron. Appl. (ICIEA)*, Jun. 2019, pp. 1006–1010, doi: 10.1109/ICIEA.2019.8833691.

[42] J. Bohg, A. Morales, T. Asfour, and D. Kragic, "Data-driven grasp synthesis—A survey," *IEEE Trans. Robot.*, vol. 30, no. 2, pp. 289–309, Apr. 2014, doi: 10.1109/TRO.2013.2289018.

[43] E. Brown, N. Rodenberg, J. Amend, A. Mozeika, E. Steltz, M. R. Zakin, H. Lipson, and H. M. Jaeger, ''Universal robotic gripper based on the jamming of granular material,'' *Proc. Nat. Acad. Sci. USA*, vol. 107, no. 44, pp. 18809–18814, Nov. 2010.

[44] A. M. Welhenge, R. D. Wijesinghe, and R. M. T. P. Rajakaruna, ''Robotic gripper design to handle an arbitrarily shaped object by emulating human finger motion,'' in *Proc. 8th Int. Conf. Ubi-Media Comput. (UMEDIA)*, Aug. 2015, pp. 330–334, doi: 10.1109/UMEDIA.2015.7297480.

[45] P.-C. Huang, J. Lehman, A. K. Mok, R. Miikkulainen, and L. Sentis, ''Grasping novel objects with a dexterous robotic hand through neuroevolution,'' in *Proc. IEEE Symp. Comput. Intell. Control Autom. (CICA)*, Dec. 2014, pp. 1–8, doi: 10.1109/CICA.2014.7013242.

[46] L. U. Odhner, R. R. Ma, and A. M. Dollar, ''Open-loop precision grasping with underactuated hands inspired by a human manipulation strategy,'' *IEEE Trans. Autom. Sci. Eng.*, vol. 10, no. 3, pp. 625–633, Jul. 2013, doi: 10.1109/TASE.2013.2240298.

[47] S. Levine, C. Finn, T. Darrell, and P. Abbeel, ''End-to-end training of deep visuomotor policies,'' *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1334–1373, 2016.

[48] A. Collet, M. Martinez, and S. S. Srinivasa, ''The MOPED framework: Object recognition and pose estimation for manipulation,'' *Int. J. Robot. Res.*, vol. 30, no. 10, pp. 1284–1306, Sep. 2011.

[49] Z. Wang, Z. Li, B. Wang, and H. Liu, ''Robot grasp detection using multimodal deep convolutional neural networks,'' *Adv. Mech. Eng.*, vol. 8, no. 9, Sep. 2016, Art. no. 168781401666807.

[50] E. Arruda, J. Wyatt, and M. Kopicki, ''Active vision for dexterous grasping of novel objects,'' in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 2881–2888, doi: 10.1109/IROS.2016.7759446.

[51] M. Zollhöfer, ''Commodity RGB-D sensors: Data acquisition,'' 2019, *arXiv:1902.06835*. [Online]. Available: http://arxiv.org/abs/1902.06835

[52] P. Li, B. DeRose, J. Mahler, J. A. Ojea, A. K. Tanwani, and K. Goldberg, ''Dex-net as a service (DNaaS): A cloud-based robust robot grasp planning system,'' in *Proc. IEEE 14th Int. Conf. Autom. Sci. Eng. (CASE)*, Aug. 2018, pp. 1–8.

[53] D. Fischinger, A. Weiss, and M. Vincze, ''Learning grasps with topographic features,'' *Int. J. Robot. Res.*, vol. 34, no. 9, pp. 1167–1194, Aug. 2015.

[54] Y. Jiang, S. Moseson, and A. Saxena, ''Efficient grasping from RGBD images: Learning using a new rectangle representation,'' in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 3304–3311.

[55] C. Sun, Y. Yu, H. Liu, and J. Gu, ''Robotic grasp detection using extreme learning machine,'' in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Dec. 2015, pp. 1115–1120, doi: 10.1109/ROBIO.2015.7418921.

[56] S. Kumra and C. Kanan, ''Robotic grasp detection using deep convolutional neural networks,'' 2016, *arXiv:1611.08036*. [Online]. Available: http://arxiv.org/abs/1611.08036

[57] J. Redmon and A. Angelova, ''Real-time grasp detection using convolutional neural networks,'' in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 1316–1322.

[58] J. Kim, K. Iwamoto, J. J. Kuffner, Y. Ota, and N. S. Pollard, ''Physically based grasp quality evaluation under pose uncertainty,'' *IEEE Trans. Robot.*, vol. 29, no. 6, pp. 1424–1439, Dec. 2013, doi: 10.1109/TRO.2013.2273846.

[59] S. Caldera, A. Rassau, and D. Chai, ''Robotic grasp pose detection using deep learning,'' in *Proc. 15th Int. Conf. Control, Autom., Robot. Vis. (ICARCV)*, Nov. 2018, pp. 1966–1972, doi: 10.1109/ICARCV.2018.8581091.

[60] F.-J. Chu, R. Xu, and P. A. Vela, ''Real-world multiobject, multigrasp detection,'' *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 3355–3362, Oct. 2018, doi: 10.1109/LRA.2018.2852777.

[61] H. Cheng and M. Q.-H. Meng, ''A grasp pose detection scheme with an End-to-End CNN regression approach,'' in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Dec. 2018, pp. 544–549, doi: 10.1109/ROBIO.2018.8665219.

[62] C. Rubert, D. Kappler, A. Morales, S. Schaal, and J. Bohg, ''On the relevance of grasp metrics for predicting grasp success,'' in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 265–272, doi: 10.1109/IROS.2017.8202167.

[63] A. K. Goins, R. Carpenter, W.-K. Wong, and R. Balasubramanian, ''Evaluating the efficacy of grasp metrics for utilization in a Gaussian process-based grasp predictor,'' in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2014, pp. 3353–3360.

[64] J. Leitner. *The ARCV Picking Benchmark (APB)*. Accessed: Jul. 9, 2019. [Online]. Available: http://juxi.net/acrv-picking-benchmark/

[65] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, ''The YCB object and model set: Towards common benchmarks for manipulation research,'' in *Proc. Int. Conf. Adv. Robot. (ICAR)*, Jul. 2015, pp. 510–517, doi: 10.1109/ICAR.2015.7251504.

[66] E. Johns, S. Leutenegger, and A. J. Davison, ''Deep learning a grasp function for grasping under gripper pose uncertainty,'' in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 4461–4468, doi: 10.1109/IROS.2016.7759657.

[67] S. Raschka and V. Mirjalili, *Python Machine Learning*, 2nd ed. Birmingham, U.K.: Packt, 2017.

[68] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[69] R. E. Woods and R. C. Gonzalez, *Digital Image Processing*, 4th ed. New York, NY, USA: Pearson, 2018.

[70] NVIDIA Corporation. *NVIDIA QUADRO K2200 Datasheet*. Accessed: Jul. 16, 2019. [Online]. Available: https://www.nvidia.com/content/dam/en-zz/Solutions/design-visualization/documents/75509_DS_NV_Quadro_K2200_US_NV_HR.pdf

[71] Dobot.cc. *Dobot Magician Specifications*. Accessed: Jul. 16, 2019. [Online]. Available: https://www.dobot.cc/dobot-magician/specification.html

[72] Microsoft. *Microsoft LifeCam Studio Technical Data Sheet*. Accessed: Jul. 16, 2019. [Online]. Available: http://download.microsoft.com/download/0/9/5/0952776D-7A26-40E1-80C4-76D73FC729DF/TDS_LifeCamStudio.pdf

[73] HT Sensor Technology Co. Ltd. *TAL220 Parallel Beam Load Cell*. Accessed: Jul. 16, 2019. [Online]. Available: https://cdn.sparkfun.com/datasheets/Sensors/ForceFlex/TAL220M4M5Update.pdf

[74] Avia Semiconductor. *HX711. 24-Bit Analog-to-Digital Converter (ADC) for Weigh Scales*. Accessed: Jul. 16, 2019. [Online]. Available: https://cdn.sparkfun.com/datasheets/Sensors/ForceFlex/hx711_english.pdf

[75] S. Patel, J. Topiwala, and CGPIT-Bardoli, ''Measuring a centre of gravity of an object using 4 load transducer method,'' *Int. J. Eng. Res.*, vol. V6, no. 1, pp. 210–214, 2017.

[76] A. Milioto, P. Lottes, and C. Stachniss, ''Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in CNNs,'' 2017, *arXiv:1709.06764*. [Online]. Available: http://arxiv.org/abs/1709.06764

[77] J. Redmon and A. Farhadi, ''YOLOv3: An incremental improvement,'' 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

JACQUES JANSE VAN VUUREN received the M.Eng. degree in mechatronic engineering from Massey University Manawatū, New Zealand, in 2017, under the supervision of Dr. L. Tang. His research focused on the development of a new and 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 (Senior Member, IEEE) received the B.Sc. degree in engineering from Southwest Petroleum University, China, in 1982, and the Ph.D. degree from the University of Liverpool, U.K., in 1992. She has research experience in China, U.K., Singapore, and Japan. She has been with Massey University Manawatū, 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 more than ten 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 Ph.D./master's 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. Her research interests include robotics, mechatronics, intelligent control, sensing, and industrial automation.

IBRAHIM AL-BAHADLY (Senior Member, IEEE) received the B.Sc.Eng. degree from the Baghdad University of Technology, in 1987, and the M.Sc. and Ph.D. degrees 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, Newcastle University, Newcastle upon Tyne, U.K. Since 1996, he has been with Massey University Manawatū, where he is currently an Associate Professor in electrical and electronic engineering. His research interests include power electronic applications, variable speed machines and drives, renewable energy systems, instrumentation, robotics, and automation.

KHALID MAHMOOD ARIF (Senior Member, IEEE) received the B.E. degree (Hons.) in mechanical engineering from the University of Engineering and Technology Lahore, in 2000, the master's degree in engineering from The University of Tokyo, in 2004, and the Ph.D. degree in mechanical engineering from Purdue University, West Lafayette, IN, USA, in 2011. He was an Assistant Professor with the Department of Mechatronics and Control Engineering, University of Engineering and Technology Lahore before joining Massey in 2012. He is currently a Senior Lecturer in mechatronics and robotics with Department of Mechanical and Electrical Engineering, SF&AT, Massey University, Auckland, New Zealand. His research interests include sensors, the IoT, robotics, and additive manufacturing.