2022 IEEE International Conference on
Autonomous Robot Systems and Competitions (ICARSC)
Santa Maria da Feira, Portugal – April 29-30, 2022

2022 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC) | 978-1-6654-8217-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICARSC55462.2022.9784795

Bin Picking Approaches Based on Deep Learning Techniques: A State-of-the-Art Survey

Artur Cordeiro*†, Luís F. Rocha†, Carlos Costa†‡, Pedro Costa†‡, and Manuel F. Silva*†
*ISEP/IPP - School of Engineering, Polytechnic Institute of Porto, Porto, Portugal
†INESC TEC - INESC Technology and Science, Porto, Portugal
‡FEUP - Faculty of Engineering, University of Porto, Porto, Portugal
1160952@isep.ipp.pt,{luis.f.rocha,carlos.m.costa,pedro.g.costa,manuel.s.silva}@inesctec.pt

***Abstract*—Bin picking is a highly researched topic, due to the need for automated procedures in industrial environments. A general bin picking system requires a highly structured process, starting with data acquisition, and ending with pose estimation and grasping. A high number of bin picking problems are being presently solved, through deep learning networks, combined with distinct procedures.**

This study provides a comprehensive review of deep learning approaches, implemented in bin picking problems. Throughout the review are described several approaches and learning methods based on specific domains, such as gripper oriented and object oriented, as well as summarized several methodologies, in order to solve bin picking issues. Furthermore, are introduced current strategies used to simplify particular cases and at last, are presented peculiar means of detecting object poses.

***Index Terms*—Bin picking, Deep learning, Robot grasping, Neural networks.**

I. Introduction

Alongside all the problems that have emerged with the advance in technology, bin picking is a dominant one. It was introduced more than 50 years ago [1], and it is a problem still highly researched nowadays, recently becoming mainstream due to the need for automation in industrial scenarios. Bin picking is a discipline that combines several sub-disciplines, such as scene analysis, object recognition, object localization, grasp planning, motion planning, task execution, and error detection [1].

In the majority of cases, the objects are randomly arranged in the bin (that is why it is also called random bin picking), and for this case, the commercial products implemented in the modern industry can only solve one particular case. There is no single bin picking application, but rather a spectrum of specific situations based on any number of unique constraints [2]. These circumstances can vary with peculiar objects, environments, and processes. That is why the methods are currently growing in the direction of making a robust solution, that can work with almost every product of the industry, besides having speed, efficiency, and accuracy [2].

Several methods were developed to resolve this problem. Beginning in the early 2000s with conventional approaches, consisting of feature based methods and algorithms of segmentation or detection. However, recent methods are predominantly based on deep learning techniques, that are based on deep neural networks, like the ones that are going to be mentioned and compared in this article.

It is worth mentioning, that in 2015 Amazon created a picking challenge, called Amazon Picking challenge, aimed for bin and warehouse picking, as a way of encouraging and assisting the development of automated item picking. This challenge persisted for three years, until 2017, and it was an event where many excellent solutions were proposed.

The primary contributions of this article are as follows:

- It provides a description and a revision of how current bin picking methods based on deep learning approaches are developed;
- It gives a comprehensive and summarized description of bin picking domains;
- It is created an overview table with current remarkable methodologies.

This paper is structured as follows. In the second section is mentioned the strategy and review protocol utilized as a means of writing this article. The third section mentions current bin picking strategies, such as grasp and data acquisition techniques, as well as different hardware choices. The succeeding section describes different deep learning methods to solve bin picking problems, as well as neural networks learning methodologies, and it is created a structured graph of bin picking crucial domains. In the sixth section are mentioned pose estimation algorithms utilized in these approaches, divided into two main topics: model based and model free approaches. In order to complete the article, is written a revision on several approaches, that obtained good results in the area. In the seventh section is briefly discussed the current bin picking procedures. To conclude, the paper presents possible future developments to enhance deep learning methodologies in bin picking problems and is summarised the current state of bin picking approaches.

This work has been supported by the European Regional Development Fund (FEDER) through a grant of the Operational Programme for Competitivity and Internationalization of Portugal 2020 Partnership Agreement (PRODUTECH4S&C, POCI-01-0247-FEDER-046102).

II. Search strategy and review protocol

This study was developed by applying a systematic type of review. The objective was to comprehensively collect and review information from relevant available studies. Separated searches were carried out to write this article, for the different areas of this study. Bin picking was the primary topic (keyword), followed by deep learning and, lastly, a modal revision of the two main topics, emphasizing the newest approaches, clutter scenarios and, industry applications.

Most articles considered were mainly researched from the IEEXplore, Cornell University Library, Scopus and ResearchGate databases.

III. Current bin picking strategies

In this section are introduced general tasks of bin picking, divided into specific areas. Each sub-section represents an important part of the procedure, namely, robot grasping and data acquisition, where recurrent bin picking approaches based on deep learning works, are presented.

A. Robot Grasping

Grasping is widely explored in the robotics research area, but there are still many problems to be solved due to the variety of environments and the uncertainty of perception and execution [18]. The theoretical research work on grasping still lacks the maturity for some cases of industrial implementation.

Grasping is the act of taking hold of some object by some kind of manipulator, in our case usually the human hand. Robot grasping is the act of grasping by a robotic manipulator, which may have quite different designs [19]. There are mainly four types of robotic grippers: vacuum, pneumatic, hydraulic, and servo-electric grippers. Each can have diverse configurations, such as two fingers, five fingers (human hand), vacuum cup, among others, influencing the motion planning of the robot given its objective. The grasping operation relies on the type of grippers. In conclusion, with different end effectors, different grasping strategies are needed. In general, most works and applications implement vacuum cups or two parallel grippers configurations.

Vacuum grippers are highly reliable in industrial environments. When the gripper comes in full contact with the surface of an item, creates a vacuum area, detailed in [20]. As a result, it allows picking items requiring only one point position, reducing the risk of picking several items simultaneously, though the types of picked items will be limited by the mesh that wraps the object.

Parallel grippers can pinch and pick items, as described in [20]. In contrast to vacuum or suction grippers, the pose of the gripper must be determined while taking into consideration the collision with nearby items and gripper constraints.

B. Data acquisition

Data acquisition is the process of acquiring information about the real or simulated world from different sensors to generate data. Choosing these sensors is a delicate task that will affect the whole robotic bin picking process in general. Currently, the largest number of bin picking solutions are based on Red, Green, Blue - Depth (RGB-D) sensors to acquire data, and occasionally with the assistance of pressure sensors.

Deep Learning sensor data can be classified as RGB data, depth data, skeleton sequences, and methods using a combination of these data modalities (multi-modal) [22].

In the research works analyzed, it is verified that most implementations require RGB images and depth maps, taken by RGB-D sensors. This data can be utilized in distinct ways, at different steps of the bin picking procedure. For example, Wu *et al.* [10] made use of RGB images and depth maps to construct a point cloud, defined by a group of points manifested in the same 3D system, with the objective of detecting different clusters in the same environment. Unlike Wu and his research team, Zhuang *et al.* [15] implemented this data differently, dividing it into two procedures, using only RGB images for the segmentation method, and depth maps to calculate the objects segmented poses with a previously known Computer Aided Design (CAD) model.

IV. Deep Learning methods for bin picking

Vision-based robotic grasping can be categorized along with a set of multiple different criteria. It can be divided into different domains, depicted in Figure 1. Bin picking is divided into two main topics, inspired by the work of Kleeberger et.al [25] and Matthieu Grard [39]: Gripper Oriented and Object Oriented, each of them analysed in the succeeding subsections.

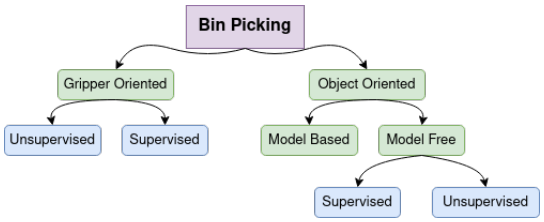

Fig. 1. Bin picking based on deep learning domains [39].

Gripper oriented techniques lack the notion of instance, meaning that the approach doesn't identify the different objects present in the workspace, which in this case is important for handling occlusions in dense piles of instances. It can be subdivided into unsupervised and supervised learning. Object oriented approaches rely either on the notion of pose or on generic instance segmentation techniques. This approach is divided into model based, also refereed in [13] as analytical approaches, and model free as data-driven. Supervised and unsupervised learning depends on the input given to the network. As the name suggests, supervised learning is a method where the model is administered by an external entity, for example, a labeled dataset, in order to train an algorithm to learn a mapping function that maps input features to output variables. On the other hand, unsupervised models try to discover patterns and structure of unlabeled data, in order to

discover information related to the objective, employed, for example, in self-learning networks [35].

A. Gripper oriented methods

Gripper oriented methods consist of detecting grasp opportunities concerning the robot end effector physics. In early approaches were employed unsupervised heuristic methods with RGB and depth images to detect the best locations for parallel grippers, or the best locally planar areas for vacuum cups. Miller *et al.* [37] introduced one of the first and most known heuristic procedures. In this case, Miller used heuristic rules in order to generate and evaluate grasps for three fingered hands, modeling an object as a set of shape primitives, such as spheres, cylinders, cones, and boxes.

In the later stages, for more complex gripper oriented methods, ranking heuristics based grasp candidates were boosted by deep convolutional networks (DCN). Some of these methods based on DCN are going to be briefly described to have a better understanding of their implementation in bin picking environments.

The majority of the methods with the best results and with more depth were developed since 2017. One of the most popular ones is Dexterity Network (Dex-Net) 2.0 [38] by Mahler *et al.* It resulted in a Grasp Quality Convolutional Neural Network (GQ-CNN) model that rapidly predicts the grasp success probability from depth images. Grasps are specified with a planar position, angle, and depth of a gripper relative to an RGB-D sensor [38]. The pipeline of this network can be identified in Figure 2.

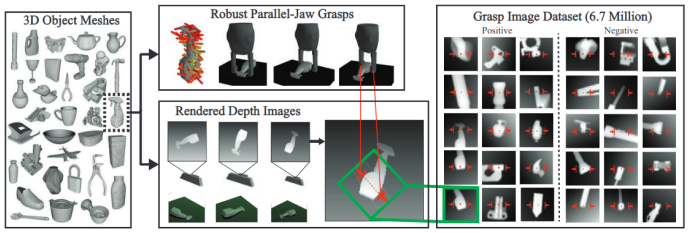

Fig. 2. Dex-Net 2.0 pipeline for training dataset generation [38].

B. Object oriented methods

Object-oriented bin-picking aims to locate affordable instances independently of the gripper model. From such a perspective, the disparity between object and gripper oriented approaches are that object oriented is strongly related to the perception of occlusions, while gripper oriented approaches are more related to friction forces and gripper torques [39].

Object oriented approaches are divided into two categories: Analytic or Model based methods and Data-driven or Model Free methods [25]. Analytic methods assume an explicit model of the target, while Data driven methods are mostly driven by image segmentation techniques, like image segmentation neural networks [39].

1) Analytic approaches (Model based): Also called geometric, analyze the shape of a target object to identify a suitable grasp pose. Frequently are based on certain information about the models, such as points of contact, Coulomb friction, and rigid body modeling. These characteristics are needed so it can be formulated a constrained optimization problem based on geometric, kinematic, and dynamic formulations [45]. In conclusion, this solution is a model based method that, as the name adduces, depends on the model, like a CAD model, given to solve the considered task. Many methods that apply this approach are based on traditional point cloud processing but can also be based on deep learning techniques. For instance, works such as LineMOD [40] and the one described by Lee *et al.* [41] implement neural networks based on this approach.

Model-based robotic grasping can be considered as a three-stage process where object poses are first estimated, then a grasp pose is determined, and, finally a collision-free kinematic feasible path is planned towards the object to pick it [25].

When utilizing object-specific knowledge, approaches typically require an object-specific configuration, which means it requires a high amount of manual tuning and a high computational power until the system reaches a good performance / accuracy ratio [25]. Consequently, this approach has the advantage of considering the geometric values of the objects but, in return, requires a high amount of time and computation power to solve the problem.

In the modern days, the majority of analytic methods incorporate trained deep convolutional networks for representation learning. In 2018, Junesuk Lee *et al.* proposed a method to estimate the 3D pose of an object using a RGB-D camera. The method consists of two modules: object detection by deep learning, and pose estimation by Iterative closest point(ICP) algorithm [41]. It is also proposed a depth based filtering method to improve the precision of object detection by preprocessing input data, and was applied a plane fitting method in the second module to increase the accuracy of the estimated pose (Figure 3).

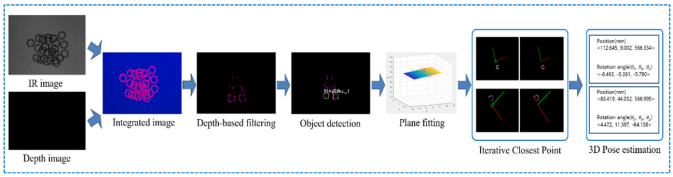

Fig. 3. Junesuk Lee *et al.* proposed approach [41].

Generally, analytical approaches face two limitations in bin picking: they require explicit object models that are not always available, and the notion of pose is only defined for rigid objects, because these can maintain the same shape when moved, making it possible to estimate the final pose [39].

2) Data-driven approaches (Model Free): Also called empirical approaches are based on machine learning [25], and gained popularity in recent years. These methods sample grasping candidates for unknown objects and rank them according to a certain metric [45]. Unlike analytic methods, they

do not need a model, such as a CAD model or a previously scanned model; the only requirements are: labels collected by humans or labelling processes, physical trial and error, heuristic methods, or a process based on a demonstration by humans or a developed robot, as described in sub-subsection V-1; therefore, it is considered a model-free method. Usually these approaches directly propose grasp candidates, and typically aim for a generalization to novel objects [25].

Data-driven approaches can be interesting in bin picking and robot grasping challenges, like the Amazon Picking Challenge (APC). In 2017 the Massachusetts Institute of Technology (MIT) Princeton Team presented an approach to solve warehouse pick and place problems, concerning cluttered environments and self-occlusions. They developed a self supervised, data-driven system based on RGB-D data. The framework was divided into two major parts. First, the views of the scene were segmented by a fully convolution neural network, and secondly, a pre-scanned 3D object model was fit in the scene to then get a six-dimension (6D) object pose [46].

With the same context, and very similar to the case just presented, Tuan-Tang Le *et al.* [6] and Peichen Wu *et al.* [10], developed methods to solve bin picking [6] and predicting grasps [10], based on deep learning networks. Typically, these approaches (data-driven) consist of first identifying the several objects in the scene, secondly creating a region or a segmentation proposal, followed by a 3D scene representation, for example by a point cloud, and lastly estimating the final object pose. Two of these cases are represented in Figure 4, where it is possible to notice the high versatility and accuracy of data-driven approaches.

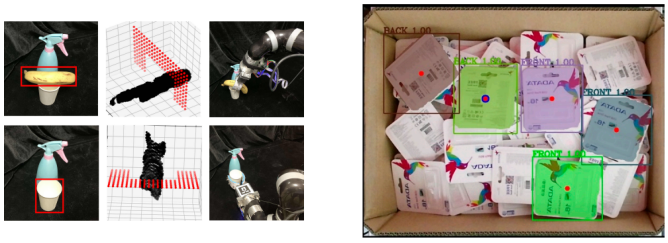

Fig. 4. Data-driven approaches [6] [10].

V. Learning Methodologies

Bin picking based on Deep Learning techniques is a common topic explored nowadays, due to the advances in neural networks. These networks require previous training before their application on a bin picking problem, otherwise, they are going to be completely "lost" with the intended objective. There are several learning methods established for distinct networks and these are differentiated based on the time required, their objective, and accuracy.

The three typical learning methods are briefly described next, associating some works that implemented them. Jeannette Bohg *et al.* wrote a depth review of each method presented, in [13]. Despite all of them being regularly used in several projects, labelling is mainly utilized due to the practicality and the results it provides.

1) Demonstration process: One of the most naive methods to use from the viewpoint of the consumer, and hardest to implement from a programmer's point of view, due to uncertainty in friction, push mechanics, and the variety of objects encountered, is the demonstration process [26]. As the name indicates, this process is based on a demonstration by a human supervisor or an already developed robot for grasping objects [37]. Based on this concept, Laskey *et al.* [26] trained a Deep Neural Network, to grasp a desired object in a cluttered situation, such as an Amazon warehouse. This network, combined with Online Learning from demonstration (LfD) algorithms, such as Dataset Aggregation (DAgger) and Svm-based reduction in Human InterVention (SHIV), empower the robot to learn control policies for such tasks, where the input is a camera image and system dynamics. The idea explored by Laskey *et al.* and his research team is visualized in Figure 5 [26].

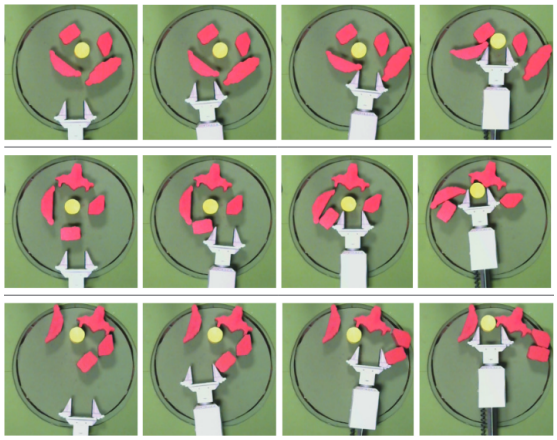

Fig. 5. Laskey *et al.* demonstration process [26].

Apart from implementing the demonstration method, Laskey article's was the first study implementing hierarchical supervisors for LfD. They deduce from physical experiments that, with 160 expert demonstrations, the probability of a successful grasp can increase from 55% to 90% [26].

2) Reinforcement learning (Trial and error): Trial and error is an uncommon method based on neural networks, due to perhaps the low efficiency and high quantity of hours needed to train. It was developed to decrease the human labor produced by the labelling of the extremely lengthiest dataset. This technique usually prones the learner to over-fit, on behalf of the dataset length.

In 2015 Pinto *et al.* [27] increased the training data to 40 times more than the previous work, leading to a dataset of 50,000 data points collected over 700 hours of robot grasping attempts. To achieve this goal a Convolutional Neural Network (CNN) was used, with a pre-trained layer from ImageNet Alexnet. The action of learning by trial and error is explained in Figure 6. The results provided by this approach had a better result than linear Support Vector Machines (SVM) [28] and k Nearest Neighbours (KNN) algorithms [27].

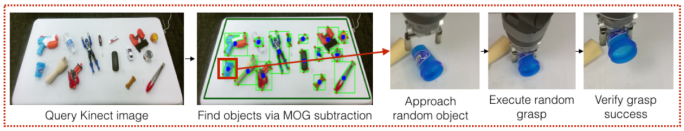

Fig. 6. Overview of how random grasp actions are sampled and executed, so it can self learn from these samples [27].

In conclusion, trial and error is an easy to use method, since it only needs to apply the prior work of other heuristics methods or hand feeding grasp possibilities. Although depending on the workspace and the whole scene, it can take a considerably high amount of time.

3) Labelling: Labelling learning is a method based on successful or unsuccessful labels collected by humans or physical trials [45].

Generating the labels can be done with different techniques. The most common and simple one is human labelling. This method can be implemented by utilizing third-party software, such as LabelMe [49], RectLabel [48], VGG Image ANootator [50], among others, or can be manually implemented by indicating the characteristics of the label.

To generate grasp candidates with high accuracy, are commonly used known datasets, such as Cornell Dataset [33], Dex-Net Dataset [29], Google Grasp Dataset [30], and Jacquard Dataset [31]. For instance, Kumra *et al.* [32] adopted Jacquard [31] and Cornell Datasets [33] to tackle the problem of generating and performing antipodal robotic grasps for unknown objects. Xiang *et al.* [9] estimated a 6D object pose in cluttered scenes based on a large scale video dataset called Yale-CMU-Berkeley (YCB) [34], also developed in the same project. Despite the employment of a model to define the final pose, the key idea was to develop a CNN capable of detecting and performing segmentation of the visualized objects. Shin *et al.* [5] took advantage of the YCB dataset and developed a deep learning-based object recognition and robot manipulator for unknown objects, estimating the direction of the object, the center point of the segment and the edge points, by drawing straight lines from the center.

As an alternative to generate and label a dataset produced by real data, or adopting developed datasets, which is high time-consuming and is sensitive to human errors or mistakes, like the methods referred above, it is possible to create and label synthetic data, such as the Dex-Net 1.0 dataset. This dataset of 3D object models is based on a sampling-based planning algorithm to explore point clouds for robust grasp planning, where the grasp prediction is observed in Figure 7. It contains 6.7 million synthetic point clouds, grasps, and analytic grasp metrics generated from thousands of 3D models. Each grasp includes an estimate of the probability for force closure under uncertainty in object and gripper pose and friction [36].

VI. Pose Estimation

To pick an object from the bin, it is essential to estimate the position and orientation (pose) of the object or to generate a grasp solution, recently applied by several deep learning

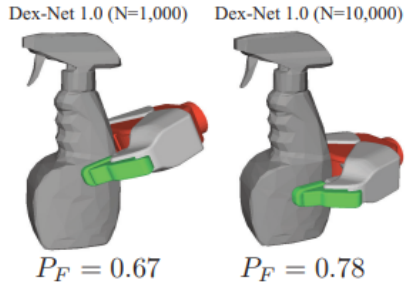

Fig. 7. Grasp predictions: 100 prior objects on the spray bottle (left), and 10,000 prior objects (right) [36].

works. Pose estimation has numerous challenges. One example is dealing with a shiny object, where the challenge comes from the fact that the object appearance largely changes with its pose and illumination. Therefore, conventional 3D-2D correspondence search usually fails due to the inconsistency of feature descriptors. For textureless objects, features matching is not appropriated, due to the absence of texture features [43]. These adversities affect the complexity of pose estimation algorithms, making them hard to be implemented.

In contrast, there are approaches that do not directly implement pose estimation, such as the work by Quanquan Shao *et al.* [18], illustrated in Figure 8. This method is based on self-supervised learning. Basically, a special framework of CNN is implemented, that combines Resnet with U-net, to predict the picking region without recognition and pose estimation. Essentially, the robot learns to grasp cylindrical objects in a cluttered bin from the results of the previous grasp, that can be successful or failures.

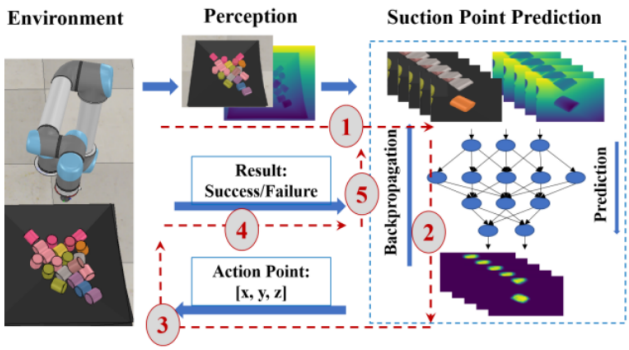

Fig. 8. Structure of self-learning robotic picking system [18].

Next are going to be briefly presented some algorithms and techniques to estimate the pose of an object in a bin picking environment. As referred, this method can differentiate on account of the approach chosen, from model based or model free approaches.

A. Model based

Model based approaches are hard to be implemented but can provide a high accuracy ratio. To estimate a pose of a visualized object based on a known object model, algorithms that match the object with the pre-known model are implemented.

Feature based methods are commonly used in model based scenarios. The objective is to find features (edges, corners, lines) and match them between images. Nevertheless, these methods are very dependent on the type of object mesh, as previously described, considering that different meshes can change the way a system "visualizes" an object [43].

The most common technique nowadays for pose refinement with point clouds is ICP, initially introduced in [44]. It is an algorithm employed to minimize the difference between two point clouds with a low transformation difference. In other words, it aims to find the transformation between point clouds by minimizing errors. Essentially, ICP is not used to find a 6D pose but to refine a pose already known. In terms of multi-modal methods, one of the most common ones is LineMOD introduced by Hinterstoisser *et al.* [40].

A popular work that implements these methods is PoseCNN [9], by Xian *et al.*. This work provides 6D poses of random objects, using only color images as input. PoseCNN estimates the 3D translation of an object by localizing its center in the image and predicting the distance from the camera. The framework of the developed convolutional neural networks is divided into three different stages: feature extraction, embedding, and classification/regression.

Zhuang *et al.* [15] proposed a pose estimation framework based on a semantic Point Pair Feature (PPF) method. This method is divided into, first, an instance segmentation process, operated by Mask-RCNN, and a matching process with PPF described in [42], that estimates the final pose, by matching the extracted point clouds scene with a model.

B. Model Free

Model independent approaches can be based on different methods, such as deep learning methods (as the ones referred in the gripper oriented section), feature based methods, and heuristic methods.

The usual deep learning based methods, normally require a 2D image and a depth map. This information can be employed as an input for a neural network, although several approaches only acquire and apply one type of input. For instance, Mahler *et al.* [38] presented a model called GQ-CNN, that predicts the probability of grasps success from depth images. In such a manner, this framework estimated a grasp pose, by merely depth data, without any additional algorithm of pose estimation. In summary, GQ-CNN retrieves and explores the depth map indicated, and predicts several grasps poses, in particular, locations and angles (in the case of a parallel gripper) to pick up the object. The grasp with a higher success rate is selected, and it is calculated the depth of the object to plan the gripper trajectory.

In some cases can be used only RGB or grayscale images, accomplishing a plane exclusive system approach. This strategy is less common due to the diverse constraints, turning the approach into rare cases, and low accuracy estimations.

Turning away from gripper oriented approaches, Tuan-Tang le *et al.* [6] implemented the pose estimation stage with simple algorithms. The procedure is divided into two stages. On the first one, Objects-Of-Interest (OOI) data handling processes all the data acquired by the system, in this case RGB image and depth image. This goes through a procedure of preprocessing the data, converting the RGB image point to depth image point, and selecting the best target object. The selection with the best target is collected by the second stage, called appropriate 3D pose estimation, where the information is queried. Finally, a 3D coordinate system is constructed on the expected target and a final pose is estimated and refined.

Shin *et al.* [5] also focused on simple algorithms to determine the localization and orientation of an object. After the process of segmentation, done by a Mask R-CNN, object poses are measured. First, their centers of gravity are calculated and afterwards is applied an algorithm that selects straight lines from the center point to detect the edge points. This line is rotated by 30 degrees to find the line with the shortest distance between edge points.

VII. Discussion

Table I presents an overview of several works that were developed with the intention of improving the current state of the art for bin picking problems. These approaches are all based on deep learning methodologies.

As observed in Table I, the different methods can be discriminated by specific characteristics, such as neural networks, bin picking domains, data modalities and types of learning. In terms of neural networks are mainly presented two different methods: instance and semantic segmentation.

An interesting procedure is to apply brand new networks, that are developed from a specific neural network backbone, for example, Resnet-101. The other approach is to implement a custom neural network with concrete objectives. For instance, Mask-RCNN is a framework for object instance segmentation, therefore it can be used for detecting and identifying different objects or persons in the image. The head architecture of this network is based on Faster R-CNN and has Residual neural network (ResNet) C4 and Feature Pyramid Network (FPN) backbones [52]. In essence, these networks are developed with different features, to obtain particular results.

Bin picking domains and learning approaches were referred in Section IV. In Table I, the identification system of bin picking domains that were used on each work was processed through the study of the architecture presented on each paper. At last, the variety of the applied data modalities is identified, between RGB image, depth image, point cloud data and multi-modal modalities, such as RGB-D. It is worth noting that the majority of point cloud data is produced by RGB-D modality.

VIII. Conclusions

Currently, bin picking based on deep learning approaches is far from being a robust automated method, despite the progress made in the last few years. Therefore, it is perceptible that in recent years, there is a vast research and development of new methodologies in this area, mainly regarding deep learning techniques, since they have a great future in non-systematic areas, where the environment is always changing. As described

TABLE I
MOST RELEVANT BIN PICKING WORKS BASED ON DEEP LEARNING.

Authors	Neural Networks (Year)	Bin Picking Domain	Data modality	Learning
Lenz *et al.* [47]	(2014)	Gripper Oriented	RGB-D image	Labelling
Mahler *et al.* [38]	DexNet v2 (2017)	Gripper Oriented	Depth images	Labelling
Asif *et al.* [12]	GraspNet (2018)	Gripper Oriented	RGB-D images	Labelling
Xiang *et al.* [9]	PoseCNN (2018)	Object Oriented-Model Based	RGB image	Labelling
Lee *et al.* [41]	Yolo V2 (2018)	Object Oriented-Model Based	IR and depth image	Labelling
Shin *et al.* [5]	Mask-RCNN (2019)	Object Oriented-Model Free	RGB-D image	Labelling
Le *et al.* [6]	Mask-RCNN (2019)	Object Oriented-Model Free	RGB-D image and point cloud	Labelling
Shao *et al.* [7]	Resnet w/U-net (2019)	Object Oriented-Model Free	RGB-D image	Self-supervised learning
Blank *et al.* [8]	Yolo and Feature base (2019)	Object Oriented-Model Based	RGB-D image	Labelling
Wu *et al.* [10]	Faster R-CNN (2019)	Object Oriented-Model Free	RGB-D images	Labelling
Jiang *et al.* [17]	DCNN model (2020)	Gripper Oriented	Depth Image	Labelling
Tachikake *et al.* [14]	G^*_{base} (2020)	Gripper Oriented	RGB image	Labelling
Sukhan Lee *et al.* [4]	Hybrid Deep Learning (2020)	Object Oriented-Model Based	RGB image and point cloud	Labelling
Zhuang *et al.* [15]	Part Mask RCNN (2021)	Object Oriented-Model Based	RGB image	Labelling
Kumra *et al.* [32]	(GR-ConvNet) (2021)	Gripper Oriented	RGB-D image	Labelling
Iriondo *et al.* [3]	Graph Convolutional Network(GCN) (2021)	Gripper Oriented	RGB-D image	Labelling
Jiang *et al.* [16]	suction graspability U-Net++ (SG-U-Net++) (2021)	Gripper Oriented	Point Cloud	Labelling
Mohammed *et al.* [11]	Multi-view change observation-based approach(MV-COBA) (2021)	Object Oriented-Model Free	RGB-D images	Self-supervised learning

in this paper, bin picking procedures can be divided into different areas, such as deep learning, pose estimation, data acquisition, grasp planning, among others. Nevertheless it is noticeable that in some of these areas described, specific methods have a higher influence compared to others. For example, in the data acquisition sub-procedure almost all the works use RGB-D techniques. However, there are also sub-procedures, such as deep learning, where in several works are applied different methodologies, to either object oriented or gripper oriented domains.

Future developments for bin picking, or pick and place problems, are mainly focused on the software and hardware of the processing unit, and less on the mechanical structure of the system (robot arm). To progress is important to develop neural networks with higher accuracy and precision, capable of executing segmentation and identification of several objects in a cluttered environment. However, in counterpart, to accomplish this objective it is necessary for a high quantity of computer power, to establish a good ratio between accuracy and response time. It is also crucial to develop new databases, to provide a high number of data already labelled, such as the Common Objects in Context Microsoft COCO dataset [51].

REFERENCES

[1] D. Buchholz "Bin-Picking - New Approaches for a Classical Problem," July 2015, http://www.digibib.tu-bs.de/?docid=00060699

[2] J. Marvel, K. Saidi, R. Eastman, T. Hong, G. Cheok & E. Messina "Technology Readiness Levels for Randomized Bin Picking," in Proceedings of the Workshop on Performance Metrics for Intelligent Systems, 2012, pp. 109–113. doi: 10.1145/2393091.2393114.

[3] A. Iriondo, E. Lazkano & A. Ansuategi "Affordance-Based Grasping Point Detection Using Graph Convolutional Networks for Industrial Bin-Picking Applications," Sensors, vol. 21, no. 3, 2021, doi: 10.3390/s21030816.

[4] S. Lee and Y. Lee, "Real-Time Industrial Bin-Picking with a Hybrid Deep Learning-Engineering Approach," 2020 IEEE International Conference on Big Data and Smart Computing (BigComp), 2020, pp. 584-588, doi: 10.1109/BigComp48618.2020.00015.

[5] H. Shin, H. Hwang, H. Yoon and S. Lee, "Integration of deep learning-based object recognition and robot manipulator for grasping objects," 2019 16th International Conference on Ubiquitous Robots (UR), 2019, pp. 174-178, doi: 10.1109/URAI.2019.8768650.

[6] T. Le & C. Lin "Bin-Picking for Planar Objects Based on a Deep Learning Network: A Case Study of USB Packs," Sensors, vol. 19, no. 16, 2019, doi: 10.3390/s19163602.

[7] Q. Shao, J. Hu, W. Wang, Y. Fang, W. Liu, J. Qi & J. Ma, "Suction Grasp Region Prediction using Self-supervised Learning for Object Picking in Dense Clutter," arXiv, 2019. doi: 10.48550/ARXIV.1904.07402.

[8] A. Blank et al., "6DoF Pose-Estimation Pipeline for Texture-less Industrial Components in Bin Picking Applications," 2019 Euro-

pean Conference on Mobile Robots (ECMR), 2019, pp. 1-7, doi: 10.1109/ECMR.2019.8870920.

[9] Y. Xiang, T. Schmidt, V. Narayanan, & D. Fox, "PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes," 2018, doi:10.15607/RSS.2018.XIV.019.

[10] P. Wu, W. Chen, H. Liu, Y. Duan, N. Lin and X. Chen, "Predicting Grasping Order in Clutter Environment by Using Both Color Image and Points Cloud," 2019 WRC Symposium on Advanced Robotics and Automation (WRC SARA), 2019, pp. 197-202, doi: 10.1109/WRC-SARA.2019.8931929.

[11] M. Q. Mohammed et al., "Deep Reinforcement Learning-Based Robotic Grasping in Clutter and Occlusion," Sustainability, vol. 13, no. 24, 2021, doi: 10.3390/su132413686.

[12] U. Asif, J. Tang, & S. Harrer, "GraspNet: An Efficient Convolutional Neural Network for Real-time Grasp Detection for Low-powered Devices," in Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, Jul. 2018, pp. 4875–4882. doi: 10.24963/ijcai.2018/677.

[13] J. Bohg, A. Morales, T. Asfour and D. Kragic, "Data-Driven Grasp Synthesis—A Survey," in IEEE Transactions on Robotics, vol. 30, no. 2, pp. 289-309, April 2014, doi: 10.1109/TRO.2013.2289018.

[14] H. Tachikake, & W. Watanabe, "A Learning-based Robotic Bin-picking with Flexibly Customizable Grasping Conditions," 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020, pp. 9040-9047, doi: 10.1109/IROS45743.2020.9340904.

[15] Z. Chungang, Z. Wang, H. Zhao, & H. Ding, "Semantic part segmentation method based 3D object pose estimation with RGB-D images for bin-picking," Robotics and Computer-Integrated Manufacturing, vol. 68, p. 102086, Apr. 2021, doi: 10.1016/j.rcim.2020.102086.

[16] P. Jiang et al., "Learning suction graspability considering grasp quality and robot reachability for bin-picking," arXiv, 2021. doi: 10.48550/ARXIV.2111.02571.

[17] P. Jiang et al., "Depth Image–Based Deep Learning of Grasp Planning for Textureless Planar-Faced Objects in Vision-Guided Robotic Bin-Picking," Sensors, vol. 20, no. 3, 2020, doi: 10.3390/s20030706.

[18] . Shao, J. Hu, W. Wang, Y. Fang, W. Liu, J. Qi, & J. Ma "Suction Grasp Region Prediction using Self-supervised Learning for Object Picking in Dense Clutter," arXiv, 2019. doi: 10.48550/ARXIV.1904.07402.

[19] M. Alonso, A. Izaguirre, & M. Graña, "Current Research Trends in Robot Grasping and Bin Picking," 2019, pp. 367–376. doi: 10.1007/978-3-319-94120-2_35.

[20] M. Fujita et al., "Bin-picking Robot using a Multi-gripper Switching Strategy based on Object Sparseness," 2019 IEEE 15th International Conference on Automation Science and Engineering (CASE), 2019, pp. 1540-1547, doi: 10.1109/COASE.2019.8842977.

[21] H. Park, & D. Kim, "An Open-source Anthropomorphic Robot Hand System: HRI Hand," HardwareX, Feb. 2020, pp. e00100, Vol. 7, doi: 10.1016/j.ohx.2020.e00100

[22] M. B. Shaikh, & D. Chai, "RGB-D Data-Based Action Recognition: A Review," Sensors, vol. 21, no. 12, 2021, doi: 10.3390/s21124246.

[23] X. Yang, & Y. Tian, "Effective 3D action recognition using EigenJoints," Journal of Visual Communication and Image Representation, vol. 25, no. 1, pp. 2–11, 2014, doi: https://doi.org/10.1016/j.jvcir.2013.03.001.

[24] M. Zollhöfer, "Commodity RGB-D Sensors: Data Acquisition," arXiv, 2019. doi: 10.48550/ARXIV.1902.06835.

[25] K. Kleeberger, R. Bormann, W. Kraus, & M. F. Huber "A Survey on Learning-Based Robotic Grasping," Curr Robot Rep 1, 239–249 2020, doi: 10.1007/s43154-020-00021-6

[26] M. Laskey et al., "Robot grasping in clutter: Using a hierarchy of supervisors for learning from demonstrations," 2016 IEEE International Conference on Automation Science and Engineering (CASE), 2016, pp. 827-834, doi: 10.1109/COASE.2016.7743488.

[27] L. Pinto and A. Gupta, "Supersizing Self-supervision: Learning to Grasp from 50K Tries and 700 Robot Hours," 2015.

[28] A. Depeursinge, "Multiscale and Multidirectional Biomedical Texture Analysis: Finding the Needle in the Haystack," 2017.

[29] University of California, Berkeley GQ-CNN Training Datasets. (https://berkeley.app.box.com/s/6mnb2bzi5zfa7qpwyn7uq5atb7vbztng)

[30] S. Levine, P. Pastor, A. Krizhevsky, & D. Quillen, "Learning Hand-Eye Coordination for Robotic Grasping with Deep Learning and Large-Scale Data Collection," arXiv, 2016. doi: 10.48550/ARXIV.1603.02199.

[31] A. Depierre, E. Dellandrea, & L. Chen, "A Large-Scale Dataset for Robotic Grasp Detection," 2018, doi: 10.48550/ARXIV.1803.11469.

[32] S. Kumra, S. Joshi, & F. Sahin, "Antipodal Robotic Grasping using Generative Residual Convolutional Neural Network," arXiv, 2019. doi: 10.48550/ARXIV.1909.04810.

[33] Cornell University Robot Learning Lab: Learning to Grasp. (https://www.kaggle.com/oneoneliu/cornell-grasp)

[34] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes," arXiv, 2017. doi: 10.48550/ARXIV.1711.00199.

[35] B. Rajoub, "Supervised and unsupervised learning," 2020, pp. 51–89. doi: 10.1016/B978-0-12-818946-7.00003-2.

[36] J. Mahler et al., "Dex-Net 1.0: A cloud-based network of 3D objects for robust grasp planning using a Multi-Armed Bandit model with correlated rewards," 2016 IEEE International Conference on Robotics and Automation (ICRA), 2016, pp. 1957-1964, doi: 10.1109/ICRA.2016.7487342.

[37] A. T. Miller, S. Knoop, H. I. Christensen and P. K. Allen, "Automatic grasp planning using shape primitives," 2003 IEEE International Conference on Robotics and Automation (Cat. No.03CH37422), 2003, pp. 1824-1829 vol.2, doi: 10.1109/ROBOT.2003.1241860.

[38] J. Mahler et al., "Dex-Net 2.0: Deep Learning to Plan Robust Grasps with Synthetic Point Clouds and Analytic Grasp Metrics," arXiv, 2017. doi: 10.48550/ARXIV.1703.09312.

[39] M. Grard, "Generic instance segmentation for object-oriented bin-picking," phdthesis, 2019.

[40] S. Hinterstoisser et al., "Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes," 2011 International Conference on Computer Vision, 2011, pp. 858-865, doi: 10.1109/ICCV.2011.6126326.

[41] J. Lee, S. Kang, & S.-Y. Park, "3D Pose Estimation of Bin Picking Object using Deep Learning and 3D Matching," 2018.

[42] T. Birdal and S. Ilic, "Point Pair Features Based Object Detection and Pose Estimation Revisited," 2015 International Conference on 3D Vision, 2015, pp. 527-535, doi: 10.1109/3DV.2015.65.

[43] J. Rodrigues, & J. Jerónimo, "3D pose estimation for bin-picking: A data-driven approach using multi-light images," Aug. 2018.

[44] P. J. Besl and N. D. McKay, "A method for registration of 3-D shapes," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 14, no. 2, pp. 239-256, Feb. 1992, doi: 10.1109/34.121791.

[45] Y. Li, Q. Lei, C. Cheng, G. Zhang, W. Wang, & Z. Xu, "A review: machine learning on robotic grasping," Mar. 2019, p. 54. doi: 10.1117/12.2522945.

[46] A. Zeng et al., "Multi-view Self-supervised Deep Learning for 6D Pose Estimation in the Amazon Picking Challenge," arXiv, 2016. doi: 10.48550/ARXIV.1609.09475.

[47] I. Lenz, H. Lee, & A. Saxena, "Deep Learning for Detecting Robotic Grasps," arXiv, 2013. doi: 10.48550/ARXIV.1301.3592.

[48] RectLabel: An image annotation tool to label images for bounding box object detection and segmentation [Computer software]. Retrieved from https://rectlabel.com/

[49] LabelMe [Computer software]. (2017). Retrieved from http://labelme2.csail.mit.edu/Release3.0/index.php

[50] VGG Image Annotator (VIA) [Computer software]. (2017). Retrieved from https://www.robots.ox.ac.uk/ vgg/software/via/

[51] Lin, T., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. & Dollár, P. Microsoft COCO: Common Objects in Context. (2015)

[52] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," arXiv, 2017. doi: 10.48550/ARXIV.1703.06870.