
Towards Robust Robotic Affordances:

Including the Environment Semantics in Grasping Behaviours

By

PAOLA A. ARDÓN RAMÍREZ

Paola.Ardon@hw.ac.uk



HERIOT-WATT UNIVERSITY AND UNIVERSITY OF EDINBURGH

Master Thesis in accordance to the regulations of
Edinburgh Centre for Robotics.

AUGUST 2018

Project Supervisors:
Dr. Katrin Solveig Lohan
Dr. Subramanian Ramamoorthy

ABSTRACT

Artificial intelligence is essential to achieve a reliable human-robot interaction, especially when it comes to manipulation tasks. Most of the state-of-the-art literature explores robotics grasping methods by focusing on the target object or the robot’s morphology, without including the environment. When it comes to human cognitive development approaches, these physical qualities are not only inferred from the object, but also from the semantic characteristics of the surroundings.

The same analogy can be used in robotic affordances for improving objects grasps, where the perceived physical qualities of the objects give valuable information about the possible manipulation actions. This work proposes a framework able to reason on the object affordances and grasping regions. Each calculated grasping area is the result of a sequence of concrete ranked decisions based on the inference of different highly related attributes. The results show that the system can infer on suitable grasping areas depending on its affordance without having any *a-priori* knowledge on the shape nor the grasping points. To achieve such methodology, a combination of deep learning neural networks embedded in the form of a knowledge base along with geometrical object modelling techniques are used.

The designed framework is assessed not only by using standard learning evaluation metrics, but it is also tested on the zero-shot grasping affordance prediction scenario, obtaining a 81.3% accuracy on familiar objects. Moreover, it is compared with state-of-the-art methods that use labelled data to obtain the grasping region. The results demonstrate that the proposed method, in 88% of the cases, achieves the same grasping areas as the available current methodology without the need for labelled data.

Additionally, the outcome that is presented in this work allows the three years research proposal to continue. For which the final objective is to achieve a framework that provides a humanoid robot with autonomous capabilities to be able to help in the household. The framework mentioned above is primarily targeted to the elderly and people with health conditions or impairments.

ACKNOWLEDGEMENTS

I would first like to thank my thesis supervisors Dr. Katrin Lohan of the School of Mathematics and Computer Sciences at Heriot-Watt University and Dr. Subramanian Ramammorthy of the School of Informatics from Universtiy of Edinburgh. Their door office was always open whenever I ran into a trouble spot or had a question about my research or writing. They consistently allowed this paper to be my own work but steered me in the right direction whenever they thought I needed it.

I would also like to thank the experts who were involved in the validation for this research project: Dr. Ron Petrick and Dr. Helen Hastie of the School of Mathematics and Computer Sciences at Heriot-Watt University. Without their passionate participation and input, the validation of this work could not have been successfully conducted.

I would also like to acknowledge Ingo Keller and Eli Phoenix for their very valuable comments on this thesis.

Finally, I must express my very profound gratitude to Èric for providing me with his priceless peer reviews, unfailing support and continuous encouragement throughout this process of researching and writing this thesis.

This accomplishment would not have been possible without them. Thank you.

Paola Ardón Ramírez

DECLARATION

I Paola Ardón Ramírez declare that this dissertation is my own original work that is being submitted to Heriot-Watt University, Scotland in partial of the Degree of Master of Science in Robotics and Autonomous Systems. I acknowledge that the original work that is being submitted to Heriot-Watt University has properly been cited and referenced. Some elements of this work may have already been submitted to Heriot-Watt University as part of the dissertation preparatory work under Robotics Research Report (B31AP) and/or Robotics Research Proposal (B31AT). It has not been submitted to any other university or institute of higher learning.

Signature: _____



Paola Ardón Ramírez

August 15, 2018

TABLE OF CONTENTS

	Page
List of Tables	vi
List of Figures	vii
1 Introduction	1
1.1 Motivation	1
1.2 Problem Statement	2
1.3 Hypothesis	2
1.4 Objectives	3
1.5 Research Impact	3
2 Background	5
2.1 Modelling Objects for Grasping Behaviours	5
2.1.1 Related Work	6
2.2 Object Affordances	6
2.2.1 Related Work	7
3 Object Modelling	8
3.1 Superquadrics Overview	9
3.2 Delaunay Triangulation Overview	10
3.3 Visualising the Grasping Area	11
3.4 Object Modelling Results	12
4 Object Affordances	14
4.1 Collecting Data	15
4.2 Learning the Knowledge Base Using the Environment	16
4.2.1 Inside Each Deep Convolutional Neural Networks	18
4.2.2 Putting Together the Knowledge Base	19

4.3	Results on Affordances Learning	20
4.3.1	Each deep convolutional neural network (CNN) Performance . . .	20
4.3.2	Reasoning on the Object Affordance	21
5	Framework Results	25
5.1	Grasping Regions, Before and After Affordance Inference	26
5.2	Zero-shot Affordance	27
5.3	Similar Shape, Different Affordance	28
5.4	Compare with Similar Methods	30
5.5	Framework Limitations	31
6	Final Remarks	33
A	IEEE Workshop on ARSO 2018	35
B	Robotics Research Review 2018	38
	Bibliography	53

LIST OF TABLES

TABLE	Page
4.1 Used attributes and entities of the knowledge base (KB) graph.	17
4.2 Each of the deep CNN accuracy performance.	20
5.1 Objects modelling and grasping points before and after affordance reasoning.	27
5.2 Zero-shot affordance prediction on semantically similar objects.	29
5.3 Objects modelling and grasping points for objects with a similar shape.	30

LIST OF FIGURES

FIGURE	Page
1.1 Affordances relationship model.	1
1.2 Proposed affordances relationship model.	4
3.1 Project Scope for object modelling.	8
3.2 Superellipsoids examples for object modelling.	10
3.3 Delaunay triangulation example.	11
3.4 End-effector simulation.	12
3.5 Extracted grasping points examples.	13
4.1 Project Scope for exploring object affordances classification.	14
4.2 Sample of objects used for the framework from the Washington and MIT datasets and the different affordances groups.	15
4.3 Example of the different entities that build the designed KB.	16
4.4 KB representation used for the object affordance inference. Given an im- age, the model estimates the attributes features hierarchically following the stated inference rule.	18
4.5 Structure of a deep CNN.	18
4.6 Example of a cleaning object and the extracted attributes used to build the KB graph.	19
4.7 Posterior probability distribution for the different classifiers used in the KB.	21
4.8 Confusion matrices for the knowledge base representation.	22
4.9 Distributional posterior probabilities per class of the knowledge base.	23
4.10 Parallel coordinate plot of the features in the KB.	24
5.1 Proposed framework for grasping affordance inference.	25
5.2 Extracted grasping points examples on different objects compared with their ground truth as presented in Cornell’s dataset.	32

Acronyms

SVM	support vector machine
DOF	degrees-of-freedom
KB	knowledge base
CAD	computer-aided design
ARSO	Advanced Robotics and its Social Impacts
2-D	two-dimensional
3-D	three-dimensional
6-D	six-dimensional
CNN	convolutional neural network
SVM	support vector machine
MSE	mean square error

INTRODUCTION

1.1 Motivation

Humanoid robots are playing increasingly important roles when it comes to indoor applications. Consider a robot assisting humans by finding, collecting and delivering an object. In such complex and dynamic environments, it is hard to provide the system with every possible representation of objects. This limitation can confuse the system into reaching very similar objects with entirely different purposes, such as a candle for a glass full of liquid. Thus, the importance of a rich common sense library for robotic grasping behaviours based on the object affordance.

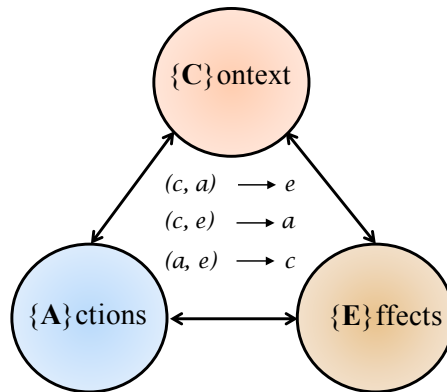


Figure 1.1: Affordances model presented in [28]. $\mathbf{C} = \{c_1, c_2, \dots, c_n\}$ is the set of attributes of the object, $\mathbf{A} = \{a_1, a_2, \dots, a_n\}$ is the set of available actions and $\mathbf{E} = \{e_1, e_2, \dots, e_n\}$ is the set of effects resulting from performing those actions as detected by the sensors.

Affordance is defined as “an opportunity for action” by Greeno [16]. Figure 1.1 shows an affordances model initially presented in Montesano et al. [28], which creates a correlation between the objects and their properties as being detected by the robot sensors. In this work, an object affordance is limited to its grasping action-effect pair that results from inferring on its context.

There is a wide range of object affordances theories in robotics. However, none of them uses a biologically inspired process as ground truth, as it is still unknown how the human thinking process works. Thus, it is not surprising that the development of artificial intelligence is still a vast area of research. Humans heavily rely on shapes and environments to identify and categorise objects in order to infer an action [4, 31]. As a result, humans succeed at generalising an action towards objects of the same category with significantly different shapes, e.g. glasses: wine, tumbler, martini, among others, and differentiate how to manipulate objects with similar shapes but for different purposes, e.g. bowling pin vs water bottle.

1.2 Problem Statement

In robotics, the most common approach to grasping affordances is to learn direct mappings to labels [7, 18, 23, 28]. However, this mapping accuracy is constrained by the amount of data needed to learn the grasping areas in each of the affordance groups. These learning methods do not reveal *what are the features that encode object affordances, especially for grasping behaviours*. Namely, these affordances do not strictly belong to the object itself. Instead, they are the result of the relationship established between them and the surroundings. Moreover, to engage in an interaction with humans, the robot has to be able to represent and reason with different sources of knowledge and decrease the already eminent uncertainty in the environment.

1.3 Hypothesis

Studies on the development of human cognitive methods demonstrate that humans improve the interactive learning process with objects not only based on previous experience with them (or similar ones) but also by inferring in the context of the environment where these objects reside [43]. Thereupon, creating a relationship between the object, the scenario where it is more likely to be found, and the object set of possible actions. Using the same analogy, this work hypothesises that in robotics the object affordances

for grasping behaviours can be improved by integrating semantic attributes of the object and the environment in which these objects are usually found, which is an approach not yet seen in the current literature.

1.4 Objectives

This research aims to investigate object affordances focussed on improving the grasping behaviours by including the environment. Thus, during the master thesis, the objectives are the following:

1. Visually exploring objects and extracting their model. Doing so without any *a-priori* acknowledge, as it will be explained in Chapter 3.
2. To implement and test a technique that allows harvesting a vast library of object features, including its surrounding environment, to prove they are valuable to deduce the object's affordance, as it is going to be discussed in Chapter 4.

By putting these two objectives together the goal is to achieve a framework that restrains the grasping behaviours of the objects depending on their affordances.

1.5 Research Impact

This work summarises an architecture that pursues to address the previously described challenges. The presented solution builds upon the assumption that, the robot visual feedback represents a good source of information. Thus, the focus on the improvement of affordances reasoning for improving the grasping areas. The work establishes its foundations on the affordances map presented in Figure 1.1 [28], particularly on its *context* element where the affordance identification resides. In this work the context $\mathbf{C} = \{c_1, c_2, \dots, c_n\}$ is modified to be the set of semantic attributes of the object and the environment that builds upon the affordance. The set of available actions, $\mathbf{A} = \{a_1, a_2, \dots, a_n\}$, is understood as two big groups: (i) the way in which the object can be approached, such as its suitable grasping areas and (ii) the usages that the object can achieve. For example a book, it is not only used to read but also as an ornament or in emergency cases as a table leg support. In the scope of this work, the action group of focus is the object best grasping area according to its affordance. And, the set of effects of performing those actions, $\mathbf{E} = \{e_1, e_2, \dots, e_n\}$, is kept in a simple discretisation

between positive or negative effects for those actions. Figure 1.2 shows a summary of the proposed object affordance approach. Where the original relationship among the map components is kept, however, they are divided into subsets:

- $(\{o\}bject \cup \{s\}urrounding) \subseteq \{C\}ontext$,
- $(\{g\}rasps \cup \{u\}sage) \subseteq \{A\}ctions$,
- $(\{p\}ositive \cup \{n\}egative) \subseteq \{E\}ffects$.

The achieved framework allows the system to model an unknown object and to reason on its affordance by correlating the semantic features of the object and its environment. This with the objective of calculating the best possible grasping region which is strongly related to the object's affordance group. Each grasping area is the result of a sequence of concrete ranked decisions based on the inference of different highly related attributes. Learning these ranked decisions allows a system with a more human-like grasping behaviour towards objects.

The system combines object reconstruction methods based on geometrical approaches and deep learning techniques that delivers an efficient knowledge base **KB** for object affordances grasping behaviours useful in indoor environments.

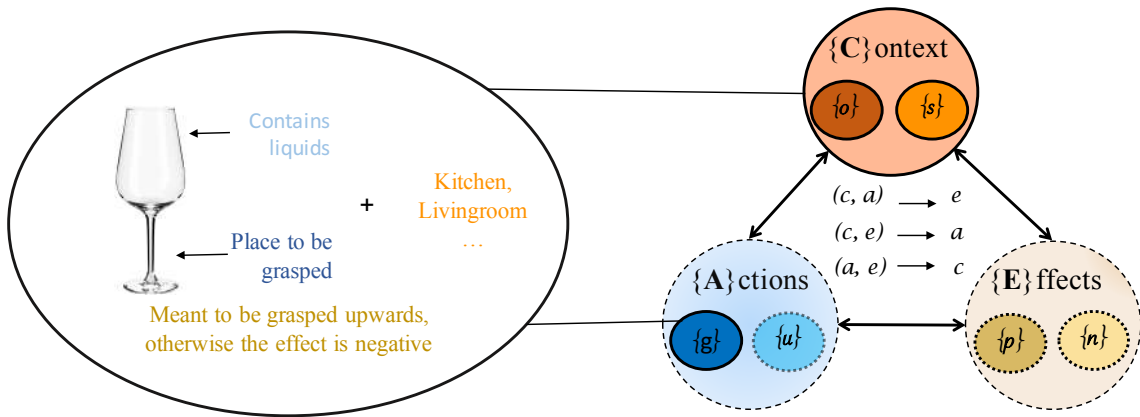


Figure 1.2: Proposed affordances relationship model, where: $\{o\}$ is the subset of object semantic features, $\{s\}$ the subset of the surroundings semantic features, $\{g\}$ the grasping regions subset, $\{u\}$ the usage of the object subset, and $\{p\}$ and $\{n\}$ the positive and negative effects respectively. The scope of this work comprises the $\{C\}$ ontext set and the $\{g\}$ rasp subset of the $\{A\}$ ctions.

BACKGROUND

The literature offers a wide range of approaches to address the grasping robotic task. Generally, each solution is designed according to different constraints such as: the object representation, end-effector, and the manipulator's degrees of freedom. As a result, creating a variety of solutions that cannot be generalized to every robotic platform nor to complete the full relationship model motivating the use of affordances.

This chapter summarizes the theory and related works that are considered to be the base of the proposed method, leading to the successful accomplishment of the objectives presented in [Chapter 1](#).

2.1 Modelling Objects for Grasping Behaviours

The first goal is to model objects without the need of *a-priori* three-dimensional (3-D) data acknowledge. This work explores one of two of the approaches that have gained popularity in the field, using superquadric models and Delaunay triangulation for object reconstruction. Both methods are based on basic geometrical shapes in order to achieve an approximation of the perceived target. Superquadrics are an extension of quadric surfaces and include supertoroids, superhyperboloids and superellipsoids. They are used in object modelling because they are able to define closed surfaces. On the other hand, Delaunay triangulation are an extension of the Voronoi diagram and offers a more uniform reconstruction of the object.

Even though these methods do not accurately represent the target, they offer a complete online schema for grasping that serve as the base for a robust framework.

2.1.1 Related Work

There is a wide range of methods for object modelling. However, the summarized works in this section involve only those methods that do not need any *a-priori* information about the object. Boissonnat and Geiger [6], Delingette [9] start by giving the fundamentals of Delaunay triangulation with the purposes of reconstructing objects. This method is the combined by numerous works that combine the technique with meshes and surface reconstruction to achieve a more complete model.

Goldfeder et al. [15], Pelossof et al. [33] and Vezzani et al. [41] are works that profit from superquadric modelling to then extract the possible grasps of an object. Pelossof et al. [33] represents the objects with superquadrics to find primitive shapes such as boxes and cylinders. Once the model is obtained, they apply a support vector machine (SVM)¹ classifier to select an optimal grasp from the object's grasping parameter space.

Goldfeder et al. [15] also integrate shape primitives and superquadrics, but their object representation is a multilevel superquadric tree. This tree is created using a decomposition of the initial model, which contains the shape primitives. After a pruning routine, a subspace containing a set of suitable grasps is obtained. Vezzani et al. [41] uses the superquadric modelling for both the object and the end-effector showing it to be successful at computing the grasping area of the object and the desired pose of the end-effector. This idea is expanded in Chapter 3 as it is going to be the basis of the proposed method for object modelling.

2.2 Object Affordances

Object affordances for grasping behaviours refer to organize and store the whole knowledge that an agent has about the grasping of an object, in order to facilitate reasoning on grasping solutions and their achievability. Some of the most popular methods in this area are deep learning and knowledge base (KB). Both of them facilitate the collection and analysis of large amounts of data.

Particularly, KB methods are growing in artificial intelligence. They pursue to learn a set of general rules and features that allow the system to infer about an object or

¹Supervised learning method that uses regression analysis and classification to analyse data [8].

an action. Moreover, this method is not restricted to the output task but it also allows the system to query a larger array of questions regarding the features involved in the process. Thus, giving it direct access to the key elements that define the output.

2.2.1 Related Work

In the field of object affordances there is a wide variety of works, where not all of them care about the target object categorization. There are many methods that extract viable grasping points on the objects, independently if the object is known or novel to the system, thus not explicitly considering the target's affordance. Examples of such works are Ardón et al. [1], Lenz et al. [23], Zech and Piater [44], to mention some. Some other focus on learning the robot's control and dynamic models to achieve a grasp, such as Stoytchev [38], Bonaiuto and Arbib [7]. The latter work learns grasp affordances from motor parameters to plan grasps using trial-and-error reinforcement learning. Stoytchev [38] follows psychology theories such as the ones presented in Greeno [16] and Piaget and Cook [34] to learn from exploratory behaviours the invariants in the resulting set of observations for the grasps.

Other works such as Moldovan et al. [26] implement a Bayesian network probabilistic method [5] to learn to differentiate affordances models among two objects. Their proposed method shows good results under uncertainty. All of these previously mentioned methods assume primitive shapes such as cylinders or boxes as the target objects.

Other methods such as Geib et al. [14] focus on the actions and objects relations in a single interface representation to capture the needs of planning and robot control. And, in their later work in Detry et al. [10] they use these action complexes to extract the best grasping points of the objects.

In the vast repertoire of learning methods connecting affordances, not necessarily limited to objects, there are works that try to mimic the human reasoning by building a KB of actions based on tasks built upon reinforcement learning [37, 45].

Instead, Kraft et al. [20], Madry et al. [24], Montesano and Lopes [27] learn the visual descriptors of the objects using classifiers such as SVM [8] and decision trees [13] to categorize the objects and obtain the possible grasps. Others such as Do et al. [11], Nguyen et al. [30] instead of using classifiers alone build a model using deep CNN based on the visual object's features, being able to generalize better given the robustness of the data.

OBJECT MODELLING

To achieve a complete framework that infers on the most suitable grasping areas depending on an object's affordance the system first needs to model the target object. Thus, Figure 3.1 presents a complete flow scheme of the framework with the object modelling as the focus of this chapter.

Learning techniques have become part of the state-of-the art when it comes to grasping. However, it brings some limitations such as to collect or find a suitable dataset that

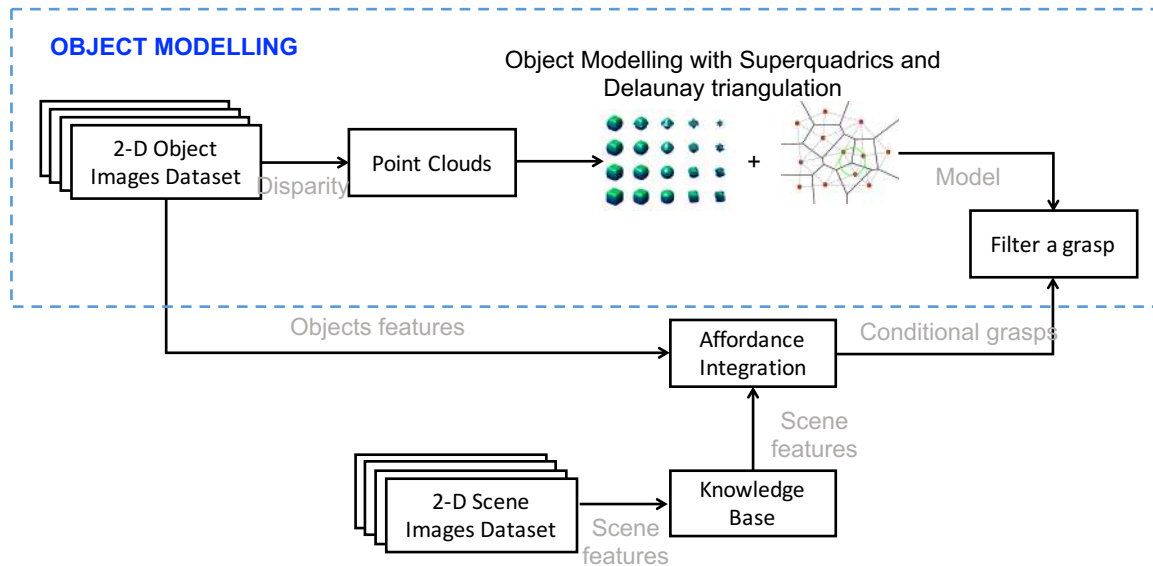


Figure 3.1: Project scoped to connecting objects with the environment in which they are more likely to be found in. The highlighted part scopes the object modelling.

includes the two-dimensional (2-D) with the mapped 3-D information about the object. Hence, the proposed approach is to model the object without any *a-priori* knowledge of the object using superquadric models [19]. One of the greater problems of using superquadrics is their limitation on uniformly sampling points around all the surface. These methods provide a denser sampling around the objects curvatures [19]. In order to avoid this issue the proposed framework combines Delaunay triangulation [22].

3.1 Superquadrics Overview

Superquadric functions are an extension of quadric surfaces and include supertoroids, superhyperboloids and superellipsoids. Superellipsoids are most commonly used in object modelling because they define closed surfaces. Jaklic et al. [19] defines a superquadric in an object-centred coordinate system represented as the inside-outside function:

$$(3.1) \quad F(x, y, z, \lambda) : \left(\left(\frac{x}{\lambda_1} \right)^{\frac{2}{\lambda_5}} + \left(\frac{y}{\lambda_2} \right)^{\frac{2}{\lambda_5}} \right)^{\frac{\lambda_5}{\lambda_4}} + \left(\frac{z}{\lambda_3} \right)^{\frac{2}{\lambda_4}},$$

where (x, y, z) is a 3-D point in the superquadric model and $\lambda = [\lambda_1, \dots, \lambda_5]$ defines the superquadric shape. Equation 3.1 provides a simple test of whether a given point lies inside or outside a superquadric, such that

$$(3.2) \quad P(x, y, z) = \begin{cases} F < 1, & \text{inside} \\ F = 0, & \text{on surface} \\ F > 0, & \text{outside} \end{cases}$$

The inside-outside description can be expressed in a generic coordinate system by adding six further variables representing the superquadric pose, e.g., three for translation and three for Euler angles, with a total of eleven independent variables (i.e. $\lambda = [\lambda_1, \dots, \lambda_{11}]$). Some examples for superellipsoids and their definition are depicted in Figure 3.2. The object modelling via superquadrics consists on finding the values of the parameter vector $\lambda \in \mathbb{R}^{11}$ so that most of the 3-D points in \mathbf{N} , where \mathbf{N} is the total number of points in the space, lie inside, on, or close to the superquadric surface. The minimization of the distance from points to the model is seen as an optimization problem in the literature.

The minimization of the algebraic distance from points to the model can be solved by defining a least-squares minimization problem [19]:

$$(3.3) \quad \min_{\lambda} = \sum_{i=1}^N \left(\sqrt{\lambda_1 \lambda_2 \lambda_3} (F(\mathbf{s}_i, \lambda) - 1) \right)^2$$

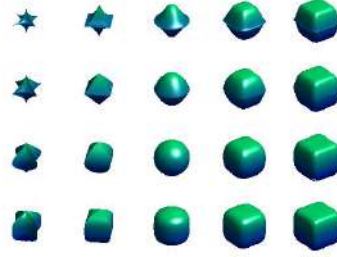


Figure 3.2: Superquadrics examples for object modelling. The surface of a superquadric is described as: $|x|^r + |y|^s + |z|^t = 1$, where r, s, t are positive real numbers that determine the main features of the superquadric. From left to right, these values are varied from less than one (pointy octahedron) to greater than two (cube modified with rounded corners and edges).

where $(F(\mathbf{s}_i, \boldsymbol{\lambda}) - 1)^2$ is the point-superquadric distance minimization and $\lambda_1 \lambda_2 \lambda_3$ is proportional to the superquadric volume. Equation 3.3 allows to solve the object modelling as an optimization problem without the need to have any *a-priori* information about the object or its shape. In the literature, this optimization problem is solved by using Levenberg-Marquardt [29] and Ipopt [42]. Among these two options, the framework uses the former method.

Nonetheless, one of the known problems of superquadrics is that it samples more points around the curvatures of the perceived shape [19]. Thus, in order to extract the grasping points along the whole surface of the object, the superquadric is combined with a Delaunay triangulation.

3.2 Delaunay Triangulation Overview

For modelling an object given a set of sample points, the Delaunay triangulation provides a convenient set of triangles to use as polygons in the model. In particular, the Delaunay triangulation avoids narrow triangles, as they have large circumcircles compared to their area. A Delaunay triangulation, considers a set \mathbf{P} of points in the (D-dimensional) Euclidean space to form a triangle and discern if the triangulation is Delaunay. An example is shown in Figure 3.3. For a triangulation to be Delaunay, no point in \mathbf{P} should be inside the circumcircle shaped by the D-dimensional triangulation DT, with the angle vectors composed by the points in \mathbf{P} , $\text{DT}(\mathbf{P})$ is formed by four chosen points inside \mathbf{P} [22]. In two dimensions, one way to detect if a point \mathbf{d} lies in the circum-

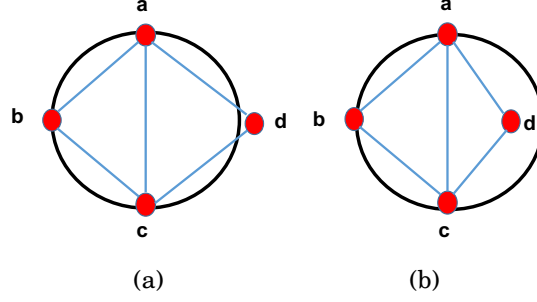


Figure 3.3: Delaunay triangulation example. (a) Delaunay triangulation, (b) not a Delaunay triangulation.

circle of points **a**, **b**, **c** is to evaluate the determinant

$$(3.4) \quad \begin{vmatrix} a_x & a_y & a_x^2 + a_y^2 & 1 \\ b_x & b_y & b_x^2 + b_y^2 & 1 \\ c_x & c_y & c_x^2 + c_y^2 & 1 \\ d_x & d_y & d_x^2 + d_y^2 & 1 \end{vmatrix} > 0,$$

where **a**, **b** and **c** are sorted counterclockwise, as depicted in Figure 3.3. This determinant is then positive, if and only if, **d** is inside the circumcircle, indicating is not a Delaunay triangulation. One of the properties of the Delaunay triangulation is that each triangle of a set of points corresponds to a face of a convex hull of the projection of points. In order to ensure a robust grasping area, the framework takes into account only those triangles that are enclosed in a diameter threshold. In this manner, avoiding considering a grasping area where only one convex hull lies, thus jeopardising the grasping task.

3.3 Visualising the Grasping Area

Given that the dynamics and control of the robotic end-effector are out of the scope of this work, the end-effector is simulated with a fictitious superellipsoid. This with the solely purpose of visualising the calculated grasping area. This superellipsoid is nominated \mathcal{H} . The hand pose is represented with a 6-D vector $\mathbf{x} = [x_h, y_h, z_h, \phi_h, \theta_h, \psi_h]$, where (x_h, y_h, z_h) are the coordinates of the hand's origin and $(\phi_h, \theta_h, \psi_h)$ the Euler Angles with respect to the world frame. This superellipsoid, \mathcal{H} , is built using the dimensions of iCub humanoid robot end-effector [25]. \mathcal{H} is sub-sampled to a set of points located in the centre with the purpose of being placed on the calculated grasping area of the object \mathcal{O} , resulting on a better visualisation of the calculated grasp region. At this

point of the framework development, the grasping region is considered to be where the density of grasping points is higher than a set threshold. This threshold is calculated by setting a perimeter of 10mm of diameter to check for a density of grasping points. Where a region is considered graspable if its points density is greater than 15 points, $P(x, y, z)$, per perimeter.

3.4 Object Modelling Results

This section shows the results of the object modelling. Up to this stage, it is of interest to verify the quality of the modelled object as well as the extracted grasping points.

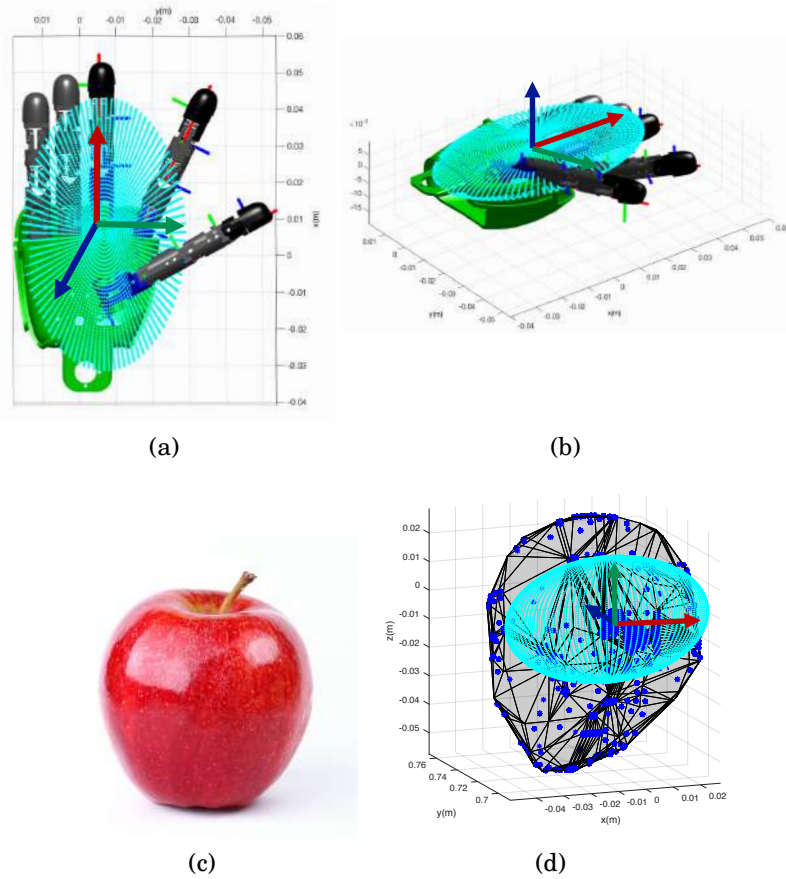


Figure 3.4: End-effector simulation. (a)-(b) show iCub humanoid Robot end-effector CAD model with its superellipsoid in cyan colour (axis colours: x is red, y is green and the z is blue); (c) 2-D image of the target object used for the sample reconstruction; (d) point cloud reconstruction using superquadrics and Delaunay triangulation, the detected grasping points are shown in blue and the final location of the end-effector in cyan colors.

Subfigure 3.4a shows the superellipsoid described in Section 3.3 on the CAD model of iCub end-effector. Subfigure 3.4b shows a profile version of both models. From now on, as referenced on these two images, the axis colours are represented as x-axis is red, the y-axis is green, and the z-axis is blue. Thus, not only allowing the visualisation of the end-effector but also the region where a grasp is acceptable (cyan ellipsoid), which is going to be evaluated in Chapter 5.

Subfigure 3.4d illustrates the modelled object with the grasping points highlighted in blue. Figure 3.5 portrays an example of this case. Where for objects which affordance action does not result in an adverse effect, such as an apple, the grasping area is inconsequential. Nonetheless, this is not the case for objects which effect could be negative, such as a meant to pour item. In order to tackle down this issue, the next chapter presents a learning affordances solution.

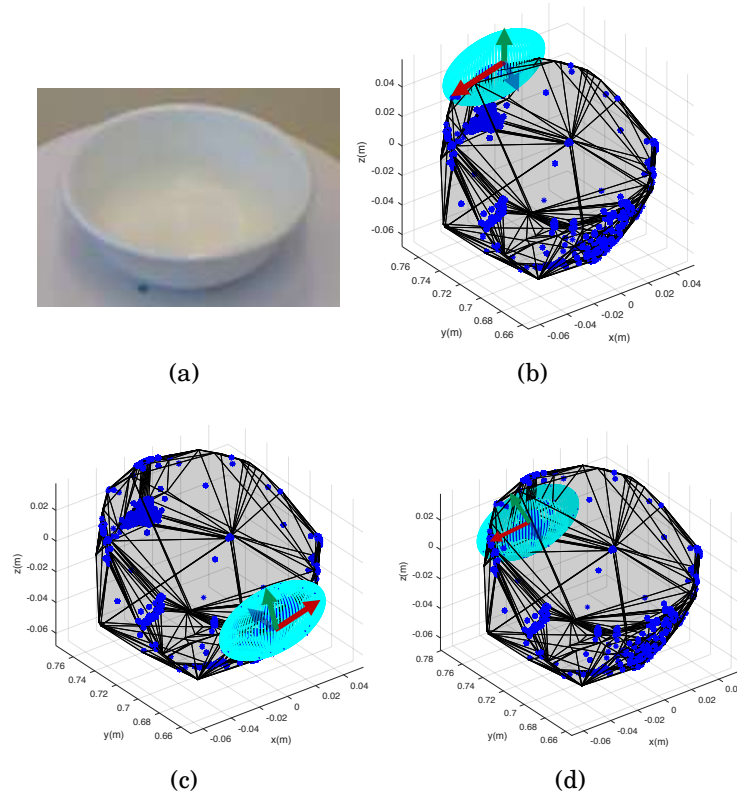


Figure 3.5: Extracted grasping points examples. (b) to (d) show different options for the grasping area of a bowl without considering the object affordance.

OBJECT AFFORDANCES

While the previous module does not need any *a-priori* information on the object to obtain a model, reasoning about the object affordance needs a library of features that gives some background about its correct affordance. In this work an affordance is considered as the relation between the object, its surrounding environment and a grasping behaviour that is convenient depending on the object's usage purpose.

The created library follows the steps highlighted in Figure 4.1. This part is enclosed

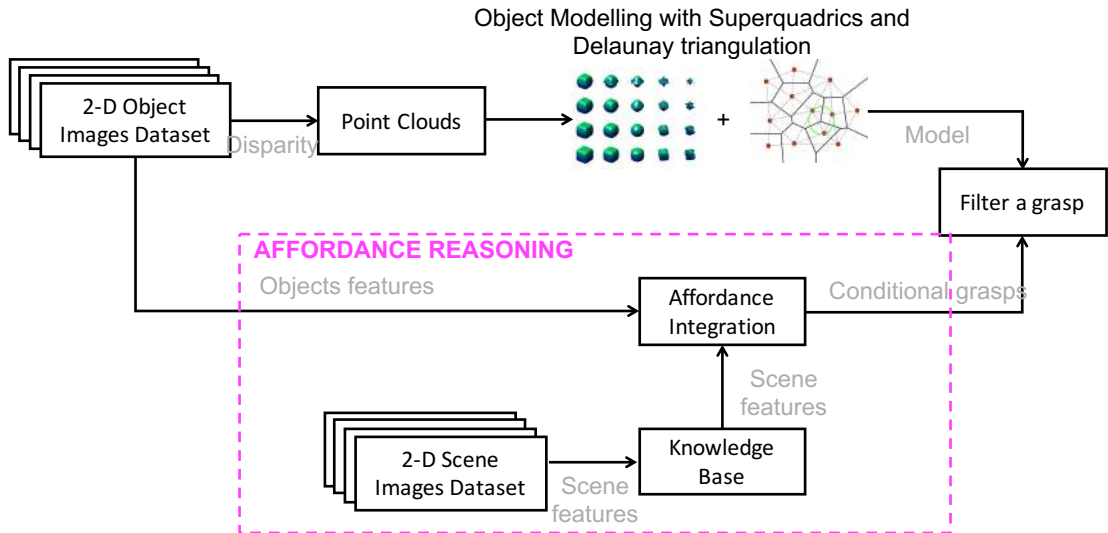


Figure 4.1: Project scoped to connecting objects with the environment in which they are more likely to be found in. The featured part scopes the object affordance classification.

in the form of a knowledge base (KB). KB methods are growing in artificial intelligence. They pursue to learn a set of general rules and features that allow the system to infer about an object or an action. Moreover, this method is not restricted to the output task but it also allows the system to query a broader array of questions regarding the features involved in the process.

In this work, a KB graph is used as a predictive model to an object affordance. The system collects a set of attributes about the objects and the environment, to then connect them in a graph style based on a set of general rules that defines the relationship among these attributes. Consequently, allowing the system to reason about the affordance group and the previously calculated grasping points. This designed KB consists of two steps: (i) collecting data and (ii) learning this data relationship to reason on the affordance for grasping.

4.1 Collecting Data

Collecting data refers to the repository of images collected from two different datasets that are finally organized in the affordance categories shown in Figure 4.2. The first one is the Washington-RGB dataset that contains 300 objects and 51 different classes,

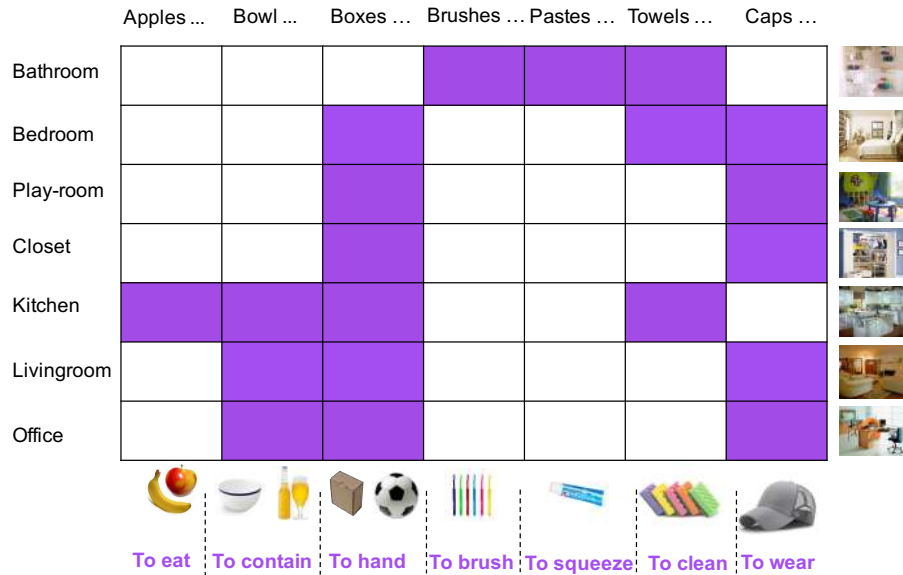


Figure 4.2: Sample of objects used for the framework from the Washington (objects depicted in columns) and MIT datasets (scenes depicted in rows) scenes depicted and the different affordances groups.

providing the point clouds and the 2-D images for each one of the instances [21]. The second dataset is the MIT Indoor scene recognition that contains 15,620 different images of 67 different indoor scenes from which this work uses seven of those classes [35]. By unifying these two datasets the objects are correlated to the environment in which they are more likely to be found in. The columns represent objects in the Washington dataset, while the rows the scenes in the MIT one, resulting in the inferred affordance shown in the bottom of the table.

Both datasets are split into 70% for training and the remaining 30% for testing. These subsets are used to train and test a battery of classifiers that help defining good object affordances features.

4.2 Learning the Knowledge Base Using the Environment

A KB is visualised as a graph representation as illustrated in Figure 4.3 where the entities (nodes) are connected by general rules (edges). In this proposed solution, the entities include the target object, the visual semantic attributes of the object and its surrounding, and the resulting affordances groups. The general rules are the attribute to attribute relation. The relation between attributes is weighted accordingly, where the higher the weight, the higher the correlation between the two entities.

The previously described repertoire of images (Section 4.1) is used to define the at-

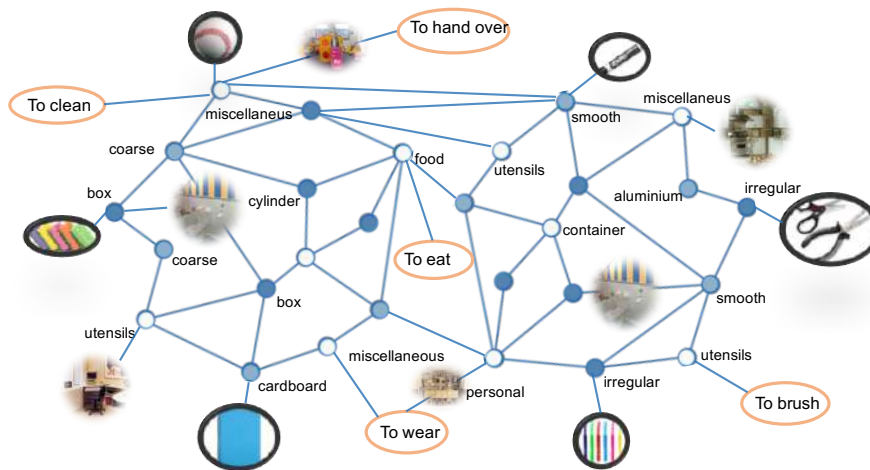


Figure 4.3: Example of the different entities that build the designed KB.

Table 4.1: Used attributes and entities of the KB graph.

Attribute	Entities per Attribute
Shape	box, cylinder, irregular, long, round
Texture	aluminium, cardboard, coarse, fabric, glass, plastic, rubber, smooth
Categorical	container, food, personal, miscellaneous, utensils
Environment	bathroom, bedroom, play-room, closet, kitchen, livingroom, office

tributes portrayed in Table 4.1 about the object. Farhadi et al. [12] offers a robust guide on how to describe objects. They divide the features into three main types: base, semantic and discriminative. In this work, the base features, such as edges and colours, are extracted using CNN as explained in Subsection 4.2.1. The semantic features are visual characteristics of the object. From now on, these features will be referred to as visual semantic features. They are the result of a deep learning CNN and are divided as:

- Shape attributes: This is defined as the set of visual attributes that describes the objects geometrical appearance,
- Texture attributes: Are a set of categories based on visual characteristics of the objects materials,
- Categorical attributes: Reflecting the semantic understanding of the object. For example, an apple is within the category of food, and
- Environment attributes: The scenarios in which the objects are more likely to be found in. This attribute is added with the purpose of facilitating the object affordances reasoning. The implemented KB considers two scenarios in which the object can be located, thus the object is not restricted to a particular environment. For example, a glass containing liquids is more likely to be found in a kitchen and a living room.

Finally, the discriminative features, those that offer a comprehensive understanding of the semantic features, are achieved through a predictive decision trees model as explained in Subsection 4.2.2.

Figure 4.4 illustrates the inference procedure followed in the KB, to arrive in an affordance group. This KB is composed of four different deep learning neural networks that, through the pre-trained CNN, resnet50 [17], extract features from the perceived images. These four different deep learning CNN correspond to the four different visual

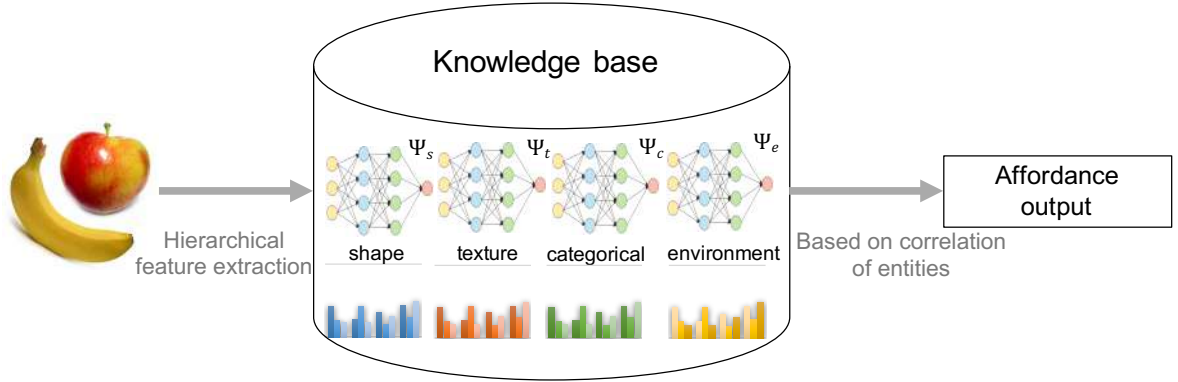


Figure 4.4: KB representation used for the object affordance inference. Given an image, the model estimates the attributes features hierarchically following the stated inference rule. These attributes are then available information on the KB. A predictive model is then applied to select the object affordance.

semantic attributes described in Table 4.1 that result in the preferred set of entities in a graph for a given affordance grasp behaviour.

4.2.1 Inside Each Deep Convolutional Neural Networks

Convolutional neural networks are a class of artificial neural networks that have successfully been applied in many fields, one of them computer vision. It is currently one of the most popular feature extraction methods used in deep learning techniques. In summary, instead of feeding the entire image as an array of numbers, the image is broken up into many tiles, the machine then tries to predict what each tile is. Finally, the computer determines what is in the picture based on the prediction of all the tiles. This procedure allows to parallelise the operations and detect the object regardless of where it is located in the image. Figure 4.5 illustrates the learning components of a deep CNN, which are: (i) *convolution* is a series of filters applied in a layered fashion

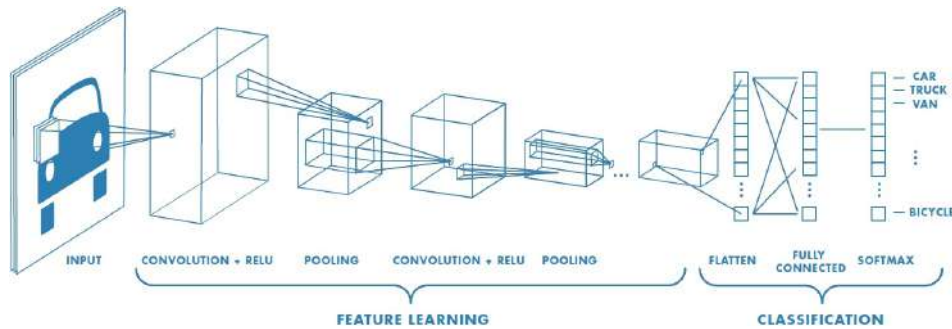


Figure 4.5: Structure of a Deep CNN [40].

to extract features from the input; (ii) *pooling* reduces data dimensionality. It applies a function summarising neighbouring information; and (iii) *fully connected layer* where each neuron in the input is connected to each neuron in the output. Depending on the task, a regression or classification algorithm can be applied to build the desired output.

4.2.2 Putting Together the Knowledge Base

The KB is then a predictive model based on the hierarchical information obtained from the different semantic attributes of the object (visualized as nodes in Figure 4.6) and the defined general rule that correlates attributes (the edges in Figure 4.6 from now on referred as weights). From each of the attributes, $\forall a \in \mathbf{A} : \mathbf{A} \in [1, \dots, K]$, where K is the total number of visual semantic features as described in Table 4.1, a set of weights represented as a vector $\Psi_{a_k} = [\psi_1, \psi_2, \dots, \psi_n]$ is extracted, where n is the total number of entities in that attribute. These Ψ_{a_k} are hierarchically connected with the next attribute a_{k+1} . Then Ψ_{a_k} offers a way to rank on the next best entity candidate.

The higher the ψ , the higher the probability that the connected two entities among attributes result in a better affordance inference. These weights are proportional to the posterior probability distribution obtained from the classification task. Such that the posterior probability distribution is defined as the Bayes rule:

$$(4.1) \quad \hat{P}(a|x) = \frac{P(x|a)P(a)}{P(x)},$$

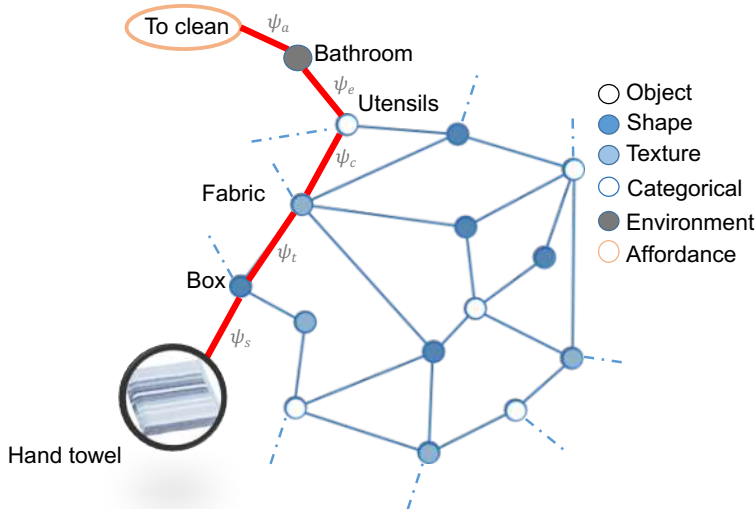


Figure 4.6: Example of an object and the extracted attributes used to build the KB graph learning a raking of weights Ψ (shown in red) that result in an affordance group.

where x is an image belonging to an attribute a , $P(a)$ is the posterior distribution and $P(x)$ is a normalisation constant that consists of the sum over a of the multivariate normal density. Figure 4.6 depicts an example of an object which grasping affordance can be to clean or to hand over. In this example, the weights deduce the best ranking (shown in red) to the *to clean* grasping affordance.

The collected information from each of the deep CNN is then learned using a decision tree as a predictive model,

$$(4.2) \quad (\mathbf{y}, Z) = (y_1, y_2, y_3, \dots, y_n, Z),$$

where Z is the affordance group that the system is trying to infer, and the vector \mathbf{y} is the set of features $y_1, y_2, y_3, \dots, y_n$ used for the inference task. Thus, the model learns the ranking that infers on the affordance grasping task $R(x) = \Psi_A^\top \mathbf{y}(x)$ where Ψ_A is the transpose of the model parameters from all the attributes and $\mathbf{y}(x)$ is the set of visual features of a given image x .

4.3 Results on Affordances Learning

The results of the presented KB for object affordances including the environment features are presented in this section. As a reminder, the proposed framework can reason on the object affordance. In this work, affordance is understood as the action-effect relation of an object, with the purpose of discerning a suitable grasp region.

4.3.1 Each deep convolutional neural network (CNN) Performance

The first tests are done individually on each of the deep learning CNN that build up the KB. 30% of the images from the Washington-RGB dataset were used for testing the battery of classifiers. Table 4.2 presents a summary of their accuracies, whereas exhaustively presented in literature, the scene recognition (environment) is the hardest

Table 4.2: Each of the deep CNN accuracy performance.

Classifier	Accuracy
Shape	95.71%
Texture	98.83%
Categorical	99.91%
Environment	76.50%

classification to boost. Even though the aim of the proposed framework is not to improve the performance of the individual classifiers, these illustrated accuracies match the state-of-the-art results shown in [17, 21]. Figure 4.7 presents a summary of these classifiers posterior probability distribution. As observed in the plots, the ones presenting a consistent distribution and mean among different classes are those with better accuracy performance in Table 4.2.

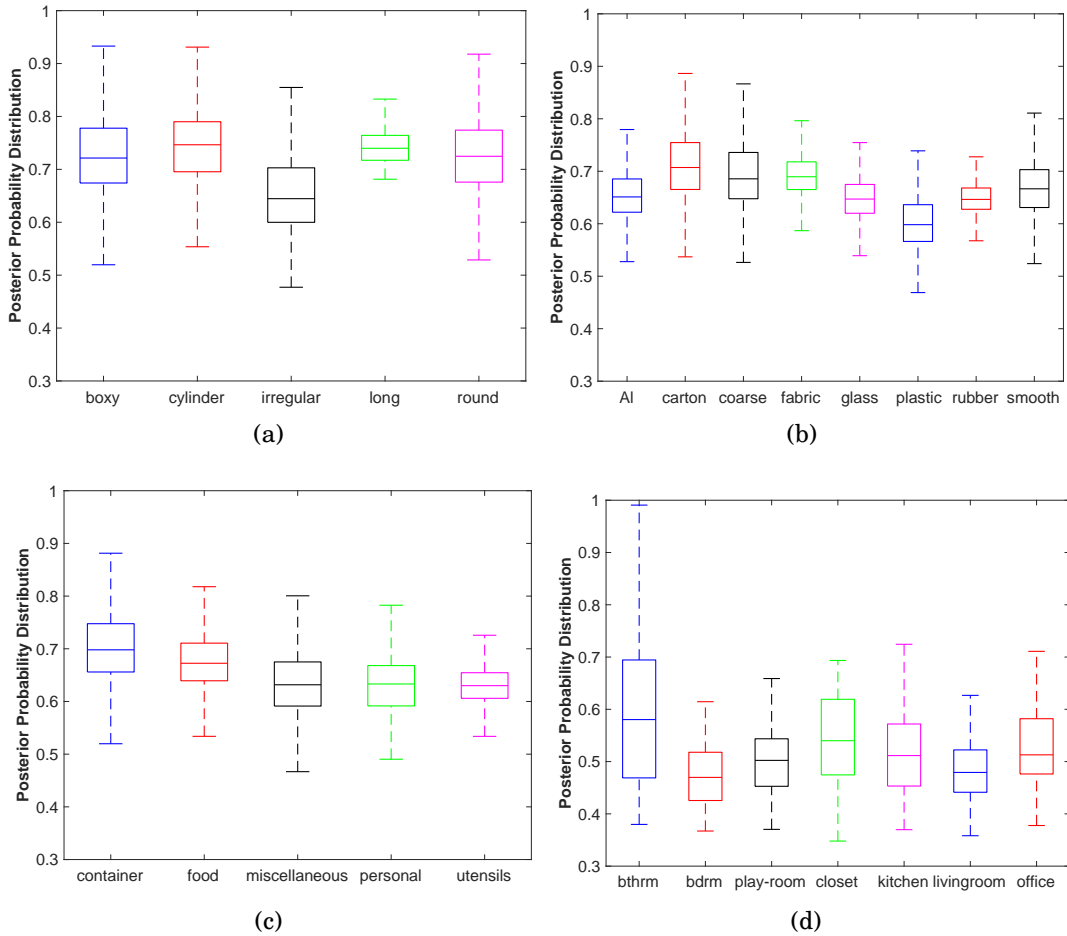


Figure 4.7: Posterior probability distribution for the different classifiers used in the KB. The box plots illustrate the performances for each deep CNN: (a) shape, (b) texture, (c) categorical and (d) environment. Note: *Al* stands for aluminium, *bthrm* for bathroom, and *bdrm* for bedroom.

4.3.2 Reasoning on the Object Affordance

In order to evaluate the overall performance of the KB the accuracy and probabilities distributions before and after adding the environment features were collected. Fig-

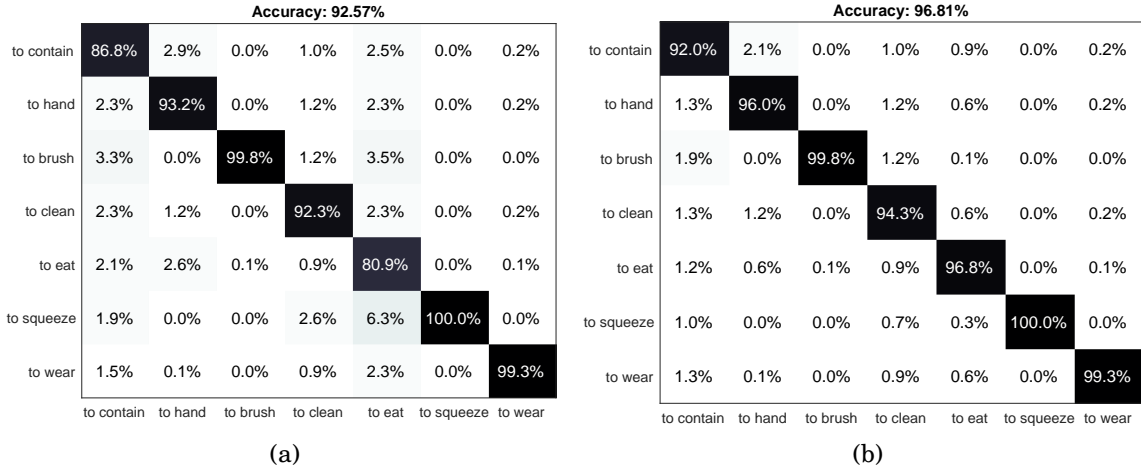


Figure 4.8: Confusion matrices for the knowledge base representation: (a) before adding environment features, showing an average diagonal accuracy of 92.57%; (b) after including the environment, showing a diagonal average accuracy of 96.81%.

Figure 4.8 and Figure 4.9 show the data for both cases. Not including the environment in the affordances has lower accuracy than adding these features to the KB, as illustrated in Subfigure 4.8a and Subfigure 4.8b. Furthermore, Subfigure 4.8a also shows a slightly higher spread among different affordance classes. For example, the case of affordances which objects have a general semantic categorical attribute such as “miscellaneous” or “container”. A percentage of objects are miss-classificated among the *to contain*, *to brush*, *to eat*, and *to squeeze* categories. Concerning grasping, this miscue represents a significant negative effect, especially for objects which real affordance is *to contain* and its miss-classification results in the system lifting up the object from any point risking dropping its content. This risk is reduced by 4.24% when adding the environment features, as portrayed in Subfigure 4.8b, especially in categories such as *to contain*, *to hand over* and *to eat*.

The posterior probability distribution of the objects among each category is also improved. Subfigure 4.9a and 4.9b show the overall increase in the median probability of the objects in the different affordances categories. While there is a decrement in the distribution for categories such as *to hand*, there is an increment for others such as *to clean*, *to squeeze* and *to wear*. This change in the distribution is accredited to the variation in environments where these objects can be found. Whereas for categories such as *to eat*, *to contain* and *to hand over* are directly related to spaces such as living rooms and kitchens, the rest are spread over more than just two contexts of surroundings.

To check the role of the environment features in solving classification discrepancies,

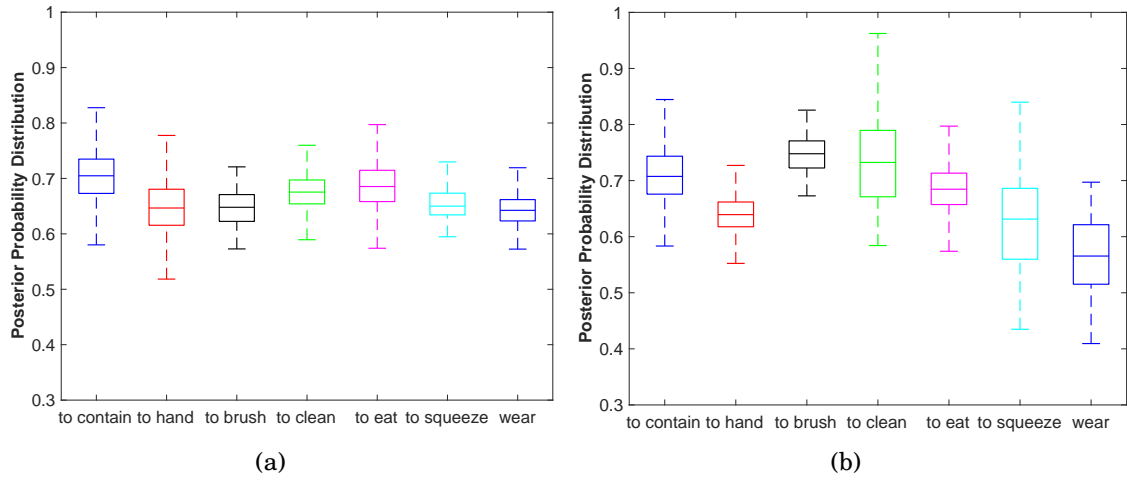
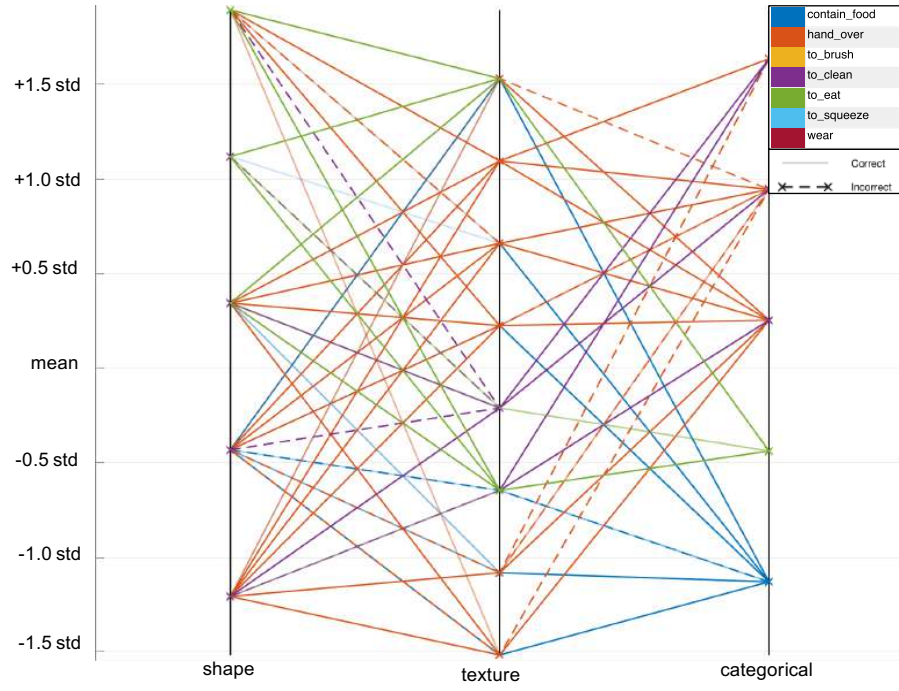
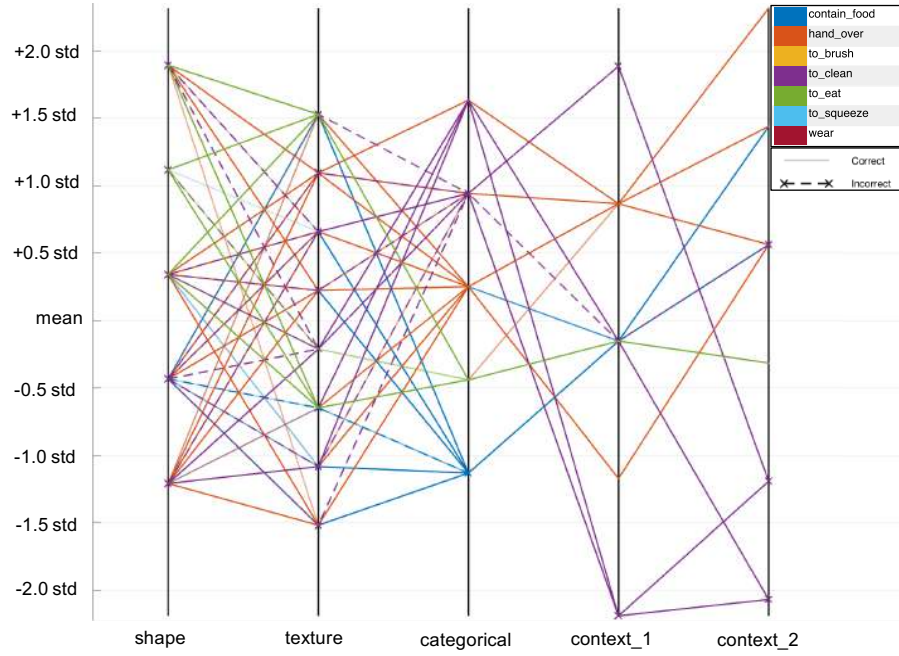


Figure 4.9: Distributional posterior probabilities per class of the knowledge base: (a) before the environment inclusion, and (b) after the environment features are included.

Figure 4.10 shows a parallel coordinates plot of the KB. This parallel coordinate plot maps the different obtained standard deviation values as a point on the line connecting the different attributes (the vertical delimiters). Each of the lines represents a tested object in the KB with their posterior probability distance from the mean value (normalised as 0) represented by the standard deviation on the y-axis. Each of the objects are categorised by a colour indicating their affordance group, as portrayed in both plots of Figure 4.10. There are many classification errors, which are illustrated as the dashed connecting lines, in the first three attributes. However, as depicted in Subfigure 4.10b, despite the fact that the standard deviation increases once the environment is added (context_1 and context_2), the miss-classification is notably reduced trying to place the input in an affordance group.



(a)



(b)

Figure 4.10: Parallel coordinate plot of the features in the KB: (a) plot with only object attributes, and (b) plot including the environment features.

FRAMEWORK RESULTS

Once the object is classified into an affordance category, the grasping region is limited accordingly. Figure 5.1 is a reminder of the proposed solution which final objective is to extract the grasping areas of the objects depending on their affordance. The system selects from the set of grasping points obtained in the object reconstruction module (Chapter 3) and limits the grasps depending on the affordance classification (Chapter 4).

In order to impose such constraints to obtain the grasping area, the space of the previously obtained grasping points is discretised into ten sub-spaces in the third dimension, z , so that the following decision on the grasping area can be made:

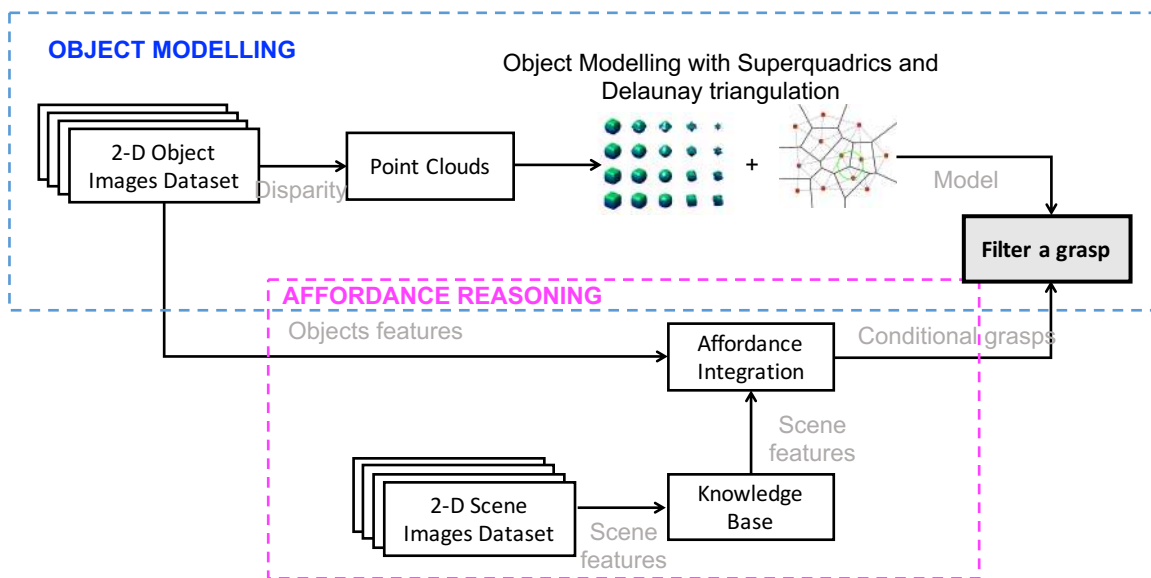


Figure 5.1: Proposed framework for grasping affordance inference.

- The grasping region should lie on those points located in the four central subspaces of the discretised space for objects that are meant to contain edibles.
- For the rest of objects, it is considered as the grasping region those subspaces where the density of grasping points is higher than a threshold (as mentioned in Section 3.3), given that the affordance action-effect is not critical (i.e., hand over, to clean, among others.).


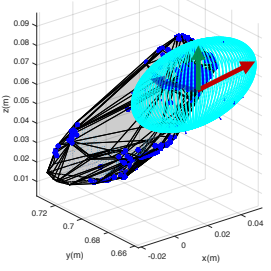
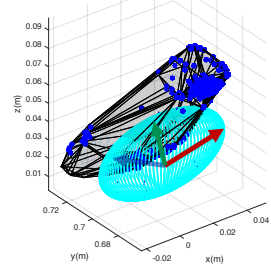

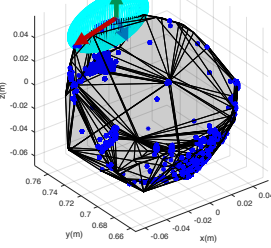
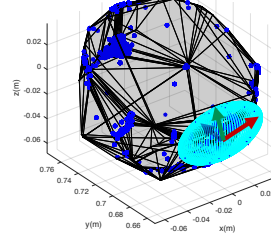

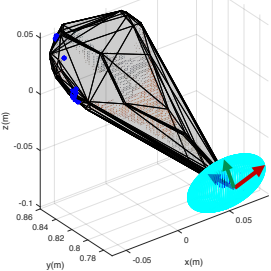
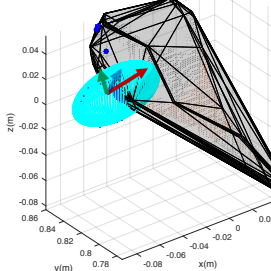
Combining the object modelling, affordances learning and grasping restriction allows for a more human-like method for object interaction. Many approaches recognise novel objects online. They are based on visual similarity or interactive learning mechanisms. An exhaustive summary on learning objects for grasping is presented in Ardón et al. [2] (Appendix B). In contrast to these methods, the one proposed in this work is not only able to: (a) infer on the object affordance of known and semantically familiar objects, but also (b) to extract a suitable grasping region of the target depending on the interpreted affordance. This chapter summarises the collection of experiments done on the proposed framework that help on assert the efficiency of the method.

5.1 Grasping Regions, Before and After Affordance Inference

Some sample objects are taken from the Washington dataset in order to evaluate if the obtained grasping on an object improves if the affordance is known. Table 5.1 shows some of these objects from which the grasping areas obtained before and after inferring on the affordance are compared. These grasp regions are analysed qualitatively according to the most likely action that a human would take in order to obtain the less negative effect.

For example, the first row shows the obtained model from a water bottle. The achieved grasp before deducing the affordances results in being placed on the lid of the bottle, which would result in an adverse effect if the bottle contained liquid and the lid was not secure. The last column of the table shows the calculated grasping area after the affordance has been inferred, which shows to be a more suitable solution given the risk of the object containing edibles. The same case can be pleaded for the second row object, a bowl. In a slightly different case, the third row shows two different grasping regions for the scissors which affordance has been determined as *hand over*. Thus both grasping choices seem acceptable given that there is no critical effect involved.

Table 5.1: Objects modelling and grasping points before and after affordance reasoning.

Object	Grasps Before	Affordance	Grasps After
		to contain	
			
			

5.2 Zero-shot Affordance

Given a novel object, it is often useful to predict its affordance either for grasping or usage actions. In this work, the object affordance is limited to its grasping action and is seen as the combination of the action-effect pair that results from the observations of the object and its environment. Zero-shots affordance, in this case, refers to the affordance prediction of a familiar object. For this part of the experiments, a set of semantically similar objects has been chosen from a third dataset, Cornell [39]. This dataset is used to learn how to grasp objects in other works such as Lenz et al. [23], Sung et al. [39]. These works exploit the fact that the dataset contains the 3-D point cloud of the objects and their corresponding labelled grasping regions in the form of rectangles.

From the Cornell dataset, 22 semantically similar objects to the ones used for the training of the KB are chosen, obtaining an average accuracy of 81.3% on the object affordance inference. In order to deduce the affordance of an unknown object o the same

hierarchical procedure explained in Subsection 4.2.2 is followed. The set of weights Ψ_A has ranked a connection of attributes that result in an affordance, depending on the perceived semantics. Furthermore, this pattern of connection has been learned in a predictive model to result in the grasping areas of the object. Table 5.2 shows a sample of the familiar objects tested using the KB with their affordance group and deduced grasping area. Where, the most critical case is shown by the ones which affordance is to contain edibles, a cup in the second row, for which the grasping area is correctly calculated. Regarding affordance inference, two objects which first affordance option should be *to wear* are classified as *to hand over*. For these particular objects, shoe and sunglasses, this miss-classification does not result in a critical action for the objects. However, tracing back the scores in the KB in both cases the objects have been categorised as a "container" and miscellaneous objects, respectively, instead of personal ones. Moreover, the connected environment is a living room (see Table 4.1 for entities per attribute categories). This miss-classification hints the need for adding either more data to the training set or more attributes to the KB. Furthermore, it reflects the results shown in Figure 4.10 in Subsection 4.3.2, where the categorical and environment attributes show to be the ones that contribute the most on discerning the affordance category.

5.3 Similar Shape, Different Affordance

One of the most significant arguments for building this framework is to help a robot to analyse on affordances in a human-like manner. That is to say, just as humans succeed at generalising an action towards objects of the same category with significantly different shapes, e.g. glasses: wine, tumbler, martini, and differentiate how to manipulate objects with similar shapes but for different purposes, e.g. candle vs water bottle. Given the database objects, this section shows the different affordances and grasping regions obtained for objects with similar shape but different affordances thus different preferred grasping regions.

Table 5.3 shows examples of two different shapes that are very common to objects with considerably different affordances. Interestingly, the second row shows examples of cylindrical objects with two different grasping affordances actions (hand over and to contain edibles) where the located grasping regions considerably differ according to the deduced affordance of the object.

Table 5.2: Zero-shot affordance prediction on semantically similar objects. The original images contain the labels (rectangles) for the preferred grasping regions from [23, 39].

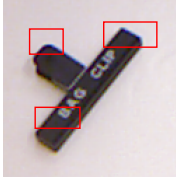
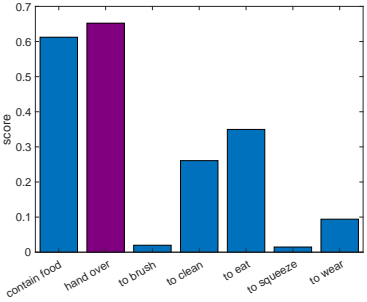
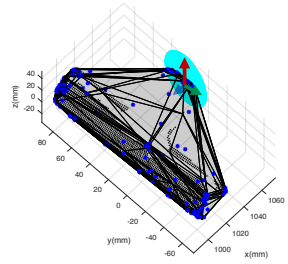
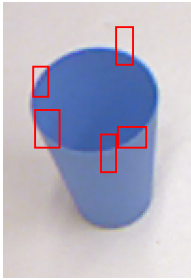
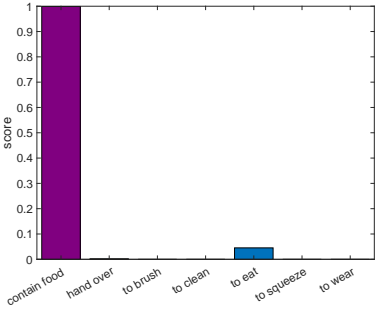
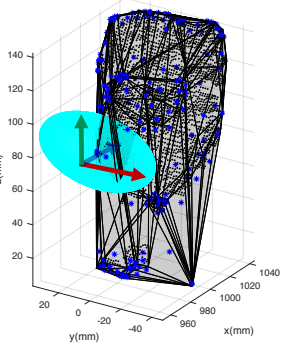
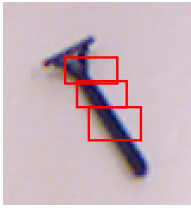
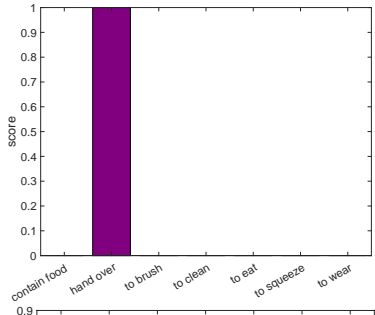
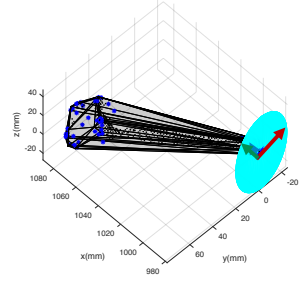

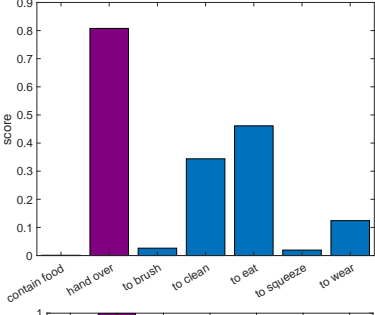
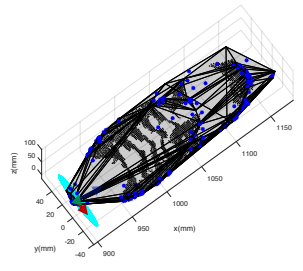
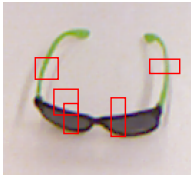
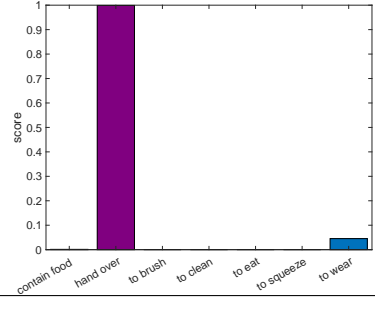
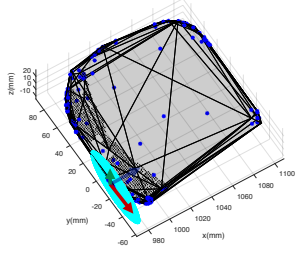

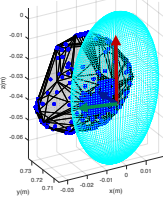

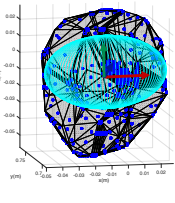
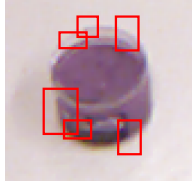
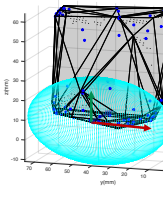
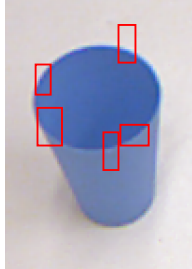
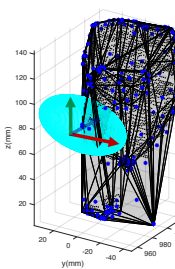
Object	Affordance	Grasp Region
	hand over 	
	to contain 	
	hand over 	
	hand over 	
	hand over 	

Table 5.3: Objects modelling and grasping points for objects with a similar shape. The objects from the Cornell dataset contain the labelled ground truth (rectangles) for the preferred grasping regions as used in [23, 39].

Shape	Affordance 1	Grasps	Affordance 2	Grasps
round	 hand over		 to eat	
	 hand over		 to contain	

5.4 Compare with Similar Methods

Different works have been done in the field of affordance detection and grasping. However, the popular available methods learn a labelled set of data in order to be able to identify the grasping regions. Contrary to these techniques, the method presented in this work deduces the grasping region without any *a-priori* information about the grasping points, using the affordance categorisation and shape of the object solely.

Given that the presented method does not train on grasp labels, in order to evaluate its output, it is compared to the ground truth labels of the Cornell dataset (used to evaluate the Zero-shot affordance in Section 5.2). Lenz et al. [23], Saxena et al. [36], Sung et al. [39] are works that use deep learning techniques to learn the grasping points of the objects mapped in the Cornell dataset images. It is worth mentioning that these works do not account for affordances learning but for object recognition. They simulate the end-effector with a rectangle, allowing it to account for the end-effector orientation, and use point and rectangle metrics to measure the mean square error (MSE) between their ground truth and the obtained grasps. Their proposed point metric computes the centre point of the predicted rectangle and considers the grasp is a success if it is within some distance from at least one ground truth rectangles. This metric does not account

for orientation as the rectangle metric does. Contrary to the method proposed in this work, their labelled grasping regions are based on the robot's end-effector control, and kinematic constraints and not on object affordances. Thus, a direct quantitative comparison is not viable. However, it is possible to use a modified version of their proposed point metric. The results of this work can be qualitatively evaluated by visually inspecting the resulting area. Moreover, quantified by the percentage of grasping regions that coincide between both sets of data, i.e., the ground truth rectangles of the Cornell dataset and the superellipsoids of this proposal.

In order to obtain such percentage, the Euclidean distance from the centre point of the labelled rectangles, observation a , to the centre point of the superellipsoid, observation b , is measured and expected to be below a set threshold (set to $10mm$ on the point cloud projection images). From the Cornell dataset, a subset of 65 random images was taken, including images from different perspectives of the same object. These images were categorised into an affordance group, illustrating their provided grasping label as a red rectangle on the 2-D image, as seen in Table 5.2 and Table 5.3. By measuring the Euclidean distance, as previously explained, 88% of the calculated grasps using the KB proposed in this work fall inside the labelled grasping regions. The other 12% falls either close to a valid region, as it is the case for Subfigure 5.2i to Subfigure 5.2l or entirely in a new area given that it has followed the constraints of the grasping regions depending on the affordance. For example, this is the case with the glass in Table 5.3. Figure 5.2 shows more examples of objects comparing both grasping areas. The ground truth is represented in the 2-D images with red rectangles and the obtained ones using the KB method are projected on the point clouds with the cyan region.

5.5 Framework Limitations

Although the framework shows overall satisfactory results it is worth mentioning that the calculated grasping area does not account with the manipulator orientation to be fitted on the object. At the moment this orientation is handled manually. The orientation and object surface matching are going to be handled through reaching and grasping behaviour techniques that are out of the scope of this work. Moreover, as it can be observed on objects such as the candle in Table 5.3 the calculated grasping area, although is correct, it would be difficult to reach given its closeness to the surface that supports the object. In order to solve this issue, more than one suitable grasping area will need to be provided for the reaching and grasping behaviour module to approach the object.

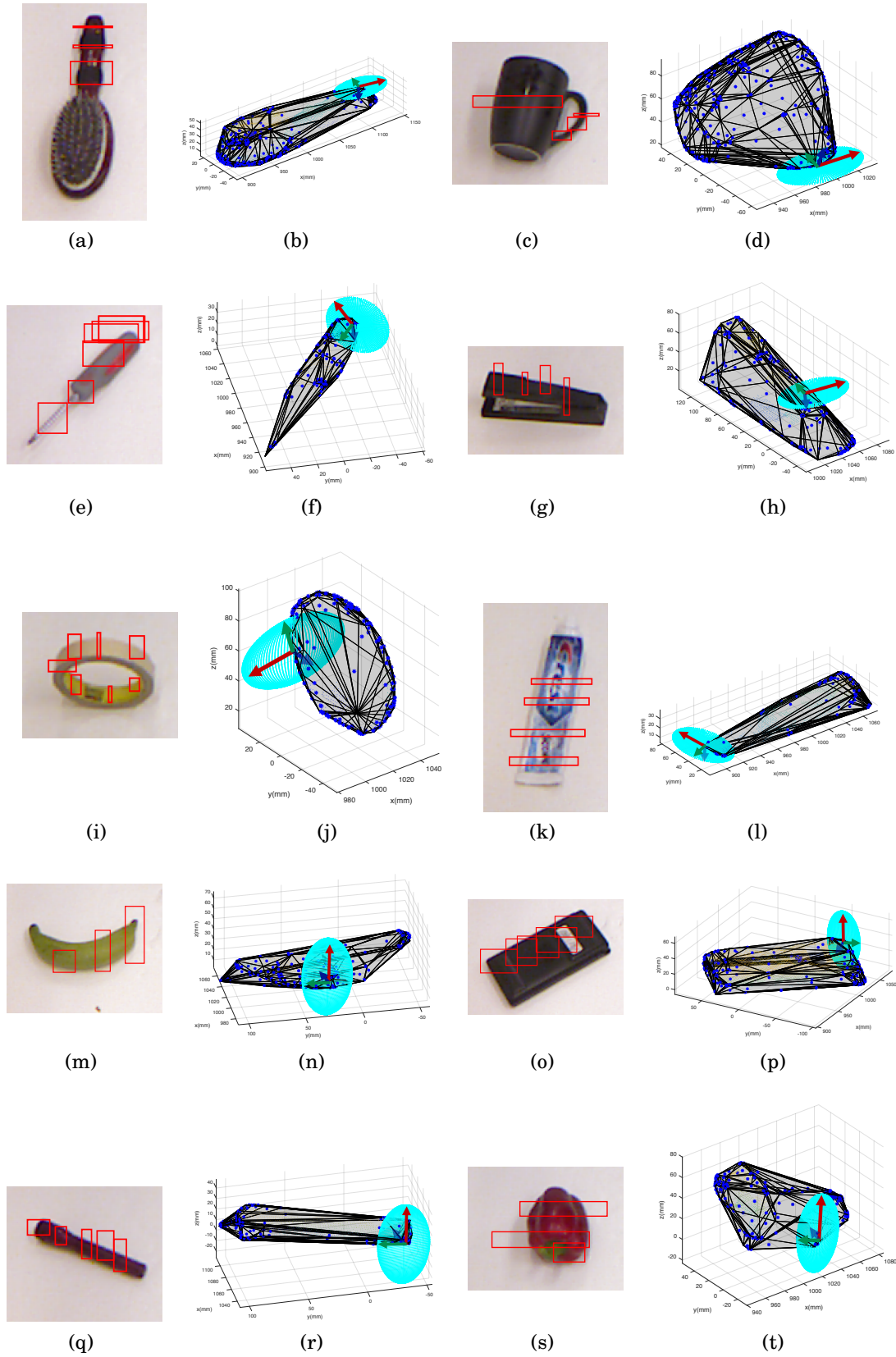


Figure 5.2: Extracted grasping points examples on different objects compared with their ground truth as presented in Cornell's dataset.

FINAL REMARKS

Past research has presented approaches to the grasping problem extensively. However, grasping behaviours depending on the object affordances is still an open challenge due to the vast variety of object shapes and robotic platforms. Furthermore, the current approaches need considerable amounts of data to train a model without being able to generalise among different classes of objects successfully, nor to distinguish the best grasp area depending on the object's purpose of use.

Thus, this work presents the base of a cognitive affordances framework that can identify and encapsulate the good affordance features of an object to deduce on a suitable grasping behaviour.

The results of the evaluation performed on the framework support the hypothesis presented at the beginning of this work. Namely, that the affordance task is not only limited to the relationship that can be built between the target object and the agent but that it also considers the surrounding environment. The results show that without any *a-priori* awareness on the grasping area of the object, the designed KB is able to induce on the object's affordance grasping points. The affordance classification is further improved by the incorporation of the environment in which these objects likely reside. Thus, allowing the system to have a better chance at deducing correctly the grasping area of the object.

Furthermore, by building a KB the system does not only learn the final predictive affordances model, but it can also access high-level information that allows it to distinguish different visual semantic attributes of the objects and their related environment.

Regarding the effectiveness of the framework as a whole, throughout Chapter 5 different comparisons were made. Especially with similar techniques that depend on labelled data in order to obtain the grasping areas. The outcome of this comparison positively asserts the effectiveness of the designed KB. By obtaining an 81.3% of accuracy at inferring the affordance of semantically similar data, and 88% of similarity on grasping areas with labelled, not previously seen, data.

It is worth to mention that one of the most significant challenges during this stage of the research was to find the right dataset to train the predictive model and test the obtained grasping areas with ground truth labels for comparison purposes. The lack of datasets that contained all the needed elements for the training and testing inspired a method that offered a hierarchical solution, i.e., the KB ranked predictive model.

Withal, the presented framework has room for improvement, which is facilitated by its modularity. Overall, the performance of the KB can be increased by adding more attributes to the base, as well as modifying the predictive model to deal with more than one affordance classification at the time (for example, an object's affordance can be *to hand over* as well as *to clean*). Furthermore, the dynamics and system control schemes of the humanoid robot and the environment are considered out of the scope of the presented work. Nonetheless, Pairet et al. [32] offers a learning-based framework that comprises relative and absolute robotic skills for dual-arm manipulation suitable for dynamic environments, that together with a self-learning mechanism for grasping, as proposed in Ardón et al. [3] (Appendix A), it offers a solution to achieve a complete human-robot interaction platform for indoor environments.



IEEE WORKSHOP ON ARSO 2018

The paper in this appendix is to be presented in the Advanced Robotics and its Social Impacts (ARSO) workshop on September in Genova, Italy. The focus of the workshop is particularly on the impact of artificial intelligence and empowered autonomous systems.

Object Affordances by Inferring on the Surroundings*

Paola Ardón Ramírez¹, Subramanian Ramamoorthy² and Katrin Solveig Lohan³

Abstract—Robotic cognitive manipulation methods aim to imitate the human-object interactive process. Most of the of the state-of-the-art literature explore these methods by focusing on the target object or on the robot’s morphology, without including the surrounding environment. Most recent approaches suggest that taking into account the semantic properties of the surrounding environment improves the object recognition. When it comes to human cognitive development methods, these physical qualities are not only inferred from the object but also from the semantic characteristics of the surroundings. Thus the importance of affordances. In affordances, the representation of the perceived physical qualities of the objects gives valuable information about the possible manipulation actions. Hence, our research pursuits to develop a cognitive affordances map by (i) considering the object and the characteristics of the environment in which this object is more likely to appear, and (ii) achieving a learning mechanism that will intrinsically learn these affordances from self-experience.

Index Terms—Humanoid robot, affordances, object recognition, learning, grasping

I. INTRODUCTION

A. Motivation

Humanoid robots are playing increasingly important roles when it comes to indoor applications, for which object affordances are vital to succeed in the human-robot interaction task. Some of these applications include assisting humans in daily activities such as cooking, cleaning, shopping, among others, thus the importance of improving robotic grasp affordances, especially in dynamic environments.

Affordance is defined as “an opportunity for action”, [7]. In robotics, we are interested in object affordances; investigating the best procedure to imitate the cognitive human development on how to interact with objects, [9]. There is a wide range of theories that try to explain the human thinking, none of them taken as the ground truth one, thus it is not surprising that the development of robotic cognitive techniques is still a wide area of research. Humans heavily rely on shapes and environments to identify and categorize objects in order to infer an action ([4], [13], [6]). As a result, we succeed at generalizing an action towards objects of the same category with significantly different shapes, e.g, glasses: wine, tumbler, martini, etc., and to differentiate how to manipulate objects with similar shapes but for different purposes, e.g, bowling pin vs. water bottle or a candle vs. a

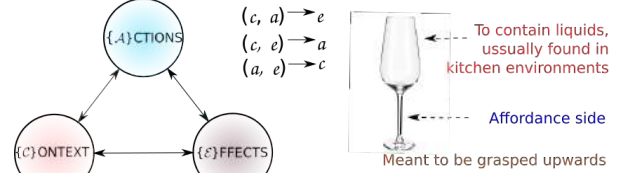


Fig. 1. Affordances model originally presented in [12], which creates a correlation between the objects and their properties as being detected by the robot sensors. We consider a slightly modified setting using reinforcement learning where: $\mathcal{C} = \{c_1, c_2, \dots, c_n\}$ will be the set of semantic attributes of the object and the environment, $\mathcal{A} = \{a_1, a_2, \dots, a_n\}$ the set of available actions and $\mathcal{E} = \{e_1, e_2, \dots, e_n\}$ the effects of performing those actions as detected by the sensors. In this model, the relationship among components of two sets infers on the best match component from the third set.

glass full of liquid. In robotics, the most common approach to affordance learning is to learn direct mappings from sensor measurements to affordance labels ([2], [3], [8], [10], [12]). However, the accuracy of this mapping is constrained by how good the perception and reconstruction of the object is, not to mention the robotic morphology constraints.

B. Problem Statement and Hypothesis

In order to achieve cognitive grasping processes, there are two main approaches in the literature. On one hand, some of the methods focus on extracting viable grasping points on the objects, independently if the object is known, familiar or novel to the system. Examples of such works are [10], [3], [1], [16], [5], among many others. These data-driven methods use these extracted features to improve their grasping success rate. However, because of the need to constantly keep learning they require large amounts of data and are not well generalized among objects belonging to different categories. On the other hand, some works focus on learning the grasping task based on the robot’s morphology using simple object primitive shapes such as spheres and boxes ([2], [8], [11]). These two different procedures consider an isolated target or many objects on a planar surface, which do not reflect real-world scenarios. Additionally, these two different approaches perform well independently, however, the literature does not put together *what are the features that encode the good object affordances?* These affordances do not belong strictly to the object nor to the robotic agent, instead, they are the result of the relationship established between them.

Social research studies on the development of human cognitive methods demonstrate that we humans improve our interactive learning with objects not only based on our previous experience with them (or similar ones) but also by inferring in the context of the environment where these objects reside ([15], [14]). Thus, we create a relationship

*Thanks to ORCA Hub EPSRC (EP/R026173/1, 2017-2021) and consortium partners

¹Paola Ardón is with the School of Mathematical and Computer Science at Heriot-Watt University and with the School of Informatics at University of Edinburgh, Edinburgh, UK paola.ardon@ed.ac.uk

²Subramanian Ramamoorthy is with the School of Informatics, University of Edinburgh, Edinburgh, UK s.ramamoorthy@ed.ac.uk

³Katrin Solveig Lohan is with the School of Mathematical and Computer Science, Heriot Watt University, Edinburgh, UK k.lohan@hw.ac.uk

between the object, the scenario where is more likely to find it, and the set of possible actions to interact with it. Using the same analogy, in robotics, the object affordances can be improved by integrating semantic attributes of the object and the environment in which these objects are usually found, which is an approach not yet seen in the current literature.

C. Objectives

This research project aims to investigate object affordances to improve the manipulation success rate by including the context of the environment when building the relationship map between the target object and the agent, e.g. humanoid robot. For this purpose, we want to create a learning mechanism based on previous experience that intrinsically generates the reward of a successful grasp, with the purpose of avoiding the use of external datasets. Figure 1 is the common used affordances model, [12], modified for our proposal along with a toy affordances example. In our case, the set of semantic properties will be composed by the object and the environment. And, the set of actions and effects will be the result of the robot's own experience.

II. METHOD

The project comprises the following sequential stages:

A. Visual Features

This stage will explore how to improve object recognition, by correlating it with the environment it is most likely located in. It will be based on early cognitive vision (ECV) descriptors containing information about shape, texture and categorical classification of the objects, as well as to give valuable information on segmenting the foreground (unknown object) and the background (environment). Thus, it is twofold: (i) the robot first interacts (visually) with the object in order to acquire a model, and (ii) once the model has been obtained it can be used for segmenting the background and learn the relationship affordances map.

B. Affordances Learning

For learning affordances, we will explore the use of reinforcement learning techniques. Instead of relying on extrinsic reward signals we will explore the usage of intrinsic ones in order for the system to experience the *success of grasping*, just as living creatures learn the skill hierarchies [14]. This approach aims to overcome the large number of samples needed for the same task using methods such as Bayesian networks [11] and learning by demonstration [3], [8].

C. Reach and Grasp Planning

This stage will be achieved by using a motion planner that will guide the end-effector towards the automatically computed grasping point. Using an on-hand camera will allow readjusting the grasping point, which will lead to a motion planner with online capabilities able to work in a dynamic environment.

D. Testing our Method

This stage aims to answer the following questions: (i) can the system identify the right object? We will use object recognition benchmarking metrics to address this question (ii) does it choose the right action for the object? for which we will measure the grasp success based on the grasp stability.

III. FINAL REMARKS

Past research has presented approaches to the affordance problem extensively. Nonetheless grasping is still an open challenge due to the large variety of object shapes and robotic platforms. The current state of the art methods is limited to specific robot manipulator, grasping scenarios, and objects. Further, the current approaches need a large amount of data to train the learning model without being able to successfully generalize among different classes of objects. Thus we aim to build a cognitive grasping framework that is able to identify and encapsulate the good features of an object that give valuable information about its affordances while learning from its own experience. This task should not only be limited to the relationship that can be built between the target object and the agent but also considering the environment surrounding the object.

REFERENCES

- [1] R. ALA, D. H. KIM, S. Y. SHIN, C. KIM, AND S.-K. PARK, *A 3D-grasp synthesis algorithm to grasp unknown objects based on graspable boundary and convex segments*, Information Sciences, 295 (2015), pp. 91–106.
- [2] J. BONAIUTO AND M. A. ARBIB, *Learning to grasp and extract affordances: the Integrated Learning of Grasps and Affordances (ILGA) model*, Biological cybernetics, 109 (2015), pp. 639–669.
- [3] H. DANG AND P. K. ALLEN, *Robot learning of everyday object manipulations via human demonstration*, IEEE/RSJ 2010 International Conference on Intelligent Robots and Systems, IROS 2010 - Conference Proceedings, (2010), pp. 1284–1289.
- [4] H. P. O. DE BEECK, K. TORFS, AND J. WAGEMANS, *Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway*, Journal of Neuroscience, 28 (2008), pp. 10111–10123.
- [5] C. DUNE, E. MARCHAND, C. COLLOUET, AND C. LEROUX, *Active rough shape estimation of unknown objects*, in Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on, IEEE, 2008, pp. 3622–3627.
- [6] L. FADIGA, L. FOGASSI, V. GALLESE, AND G. RIZZOLATTI, *Visuomotor neurons: Ambiguity of the discharge or motorperception?*, International journal of psychophysiology, 35 (2000), pp. 165–177.
- [7] J. G. GREENO, *Gibson's affordances*, American Psychological Association, (1994).
- [8] T. HERMANS, J. M. REHG, AND A. BOBICK, *Affordance prediction via learned object attributes*, in IEEE International Conference on Robotics and Automation (ICRA): Workshop on Semantic Perception, Mapping, and Exploration, Citeseer, 2011, pp. 181–184.
- [9] T. E. HORTON, A. CHAKRABORTY, AND R. S. AMANT, *Affordances for robots: a brief survey*, AVANT. Pismo Awangardy Filozoficzno-Naukowej, 2 (2012), pp. 70–84.
- [10] I. LENZ, H. LEE, AND A. SAXENA, *Deep learning for detecting robotic grasps*, International Journal of Robotics Research, 34 (2015), pp. 705–724.
- [11] B. MOLDOVAN, P. MORENO, M. VAN OTTERLO, J. SANTOS-VICTOR, AND L. DE RAEDT, *Learning relational affordance models for robots in multi-object manipulation tasks*, in Robotics and Automation (ICRA), 2012 IEEE International Conference on, IEEE, 2012, pp. 4373–4378.
- [12] L. MONTESANO, M. LOPES, A. BERNARDINO, AND J. SANTOS-VICTOR, *Learning object affordances: From sensory-motor coordination to imitation*, IEEE Trans. Robotics, 24 (2008), pp. 15–26.
- [13] E. OZTOP, N. S. BRADLEY, AND M. A. ARBIB, *Infant grasp learning: a computational model*, Experimental brain research, 158 (2004), pp. 480–503.
- [14] J. PIAGET AND M. COOK, *The origins of intelligence in children*, vol. 8, International Universities Press New York, 1952.
- [15] J. V. WERTSH AND P. TULVISTE, *Apprenticeship in thinking: Cognitive development in social context*, Science, 249 (1990), pp. 684–686.
- [16] P. ZECH AND J. PIATER, *Active and transfer learning of grasps by sampling from demonstration*, (2016).

ROBOTICS RESEARCH REVIEW 2018

The paper in this appendix is to be submitted as a journal this academic year. It is a summary of the commonly used object reconstruction and grasping methods used in robotics.

Learning Object Reconstruction for Grasping: A Review

Paola Ardón¹, Subramanian Ramamoorthy² and Katrin Solveig Lohan³

Abstract—Humanoid robots are playing increasingly important roles when it comes to indoor applications, being grasping one of them. There is a repertoire of methodologies that attempt to make the grasping task as human-like as possible, varying from manipulating specific known objects to applying an online learning method to grasp novel ones. In this review we highlight the object reconstruction techniques for data-driven grasp synthesis approaches based on whether the recognition is being done in an isolated or cluttered environment. We also point out to different datasets for object reconstruction that are available online that give useful information when it comes to object manipulation.

This with the motivation to facilitate on the search or creation of a dataset that encapsulates, in an efficient manner, the most relevant information regarding the object and the environment to achieve a successful grasp.

Index Terms—Robotics, humanoid robot, object recognition, dataset, learning, grasping,

I. INTRODUCTION

Grasping is considered a simple human task, yet it is not so simple for a robot. It requires many skills. On one hand the object perception and reconstruction, on the other, the hardware used for the manipulation to which the planning and reaching approach depend on to accomplish the task successfully. In the search to emulate human behaviour, specially on humanoid-robots, these two aspects are then combined to create autonomous grasping techniques. To reach this autonomy it is not enough to simply approach and grab the object, factors such as discerning between grasps among different objects and relating at some level with the environment are also needed. This ability to distinguish among grasps and to infer the objects utility in an environment is known as *affordances* in robotics [41]. *Cognitive developmental robotics (CDR)* aims to provide new understanding of how human higher cognitive functions are developed through synthetic approaches [5]. Learning these cognitive functions is one of the greatest challenges in artificial systems, and manipulation in robotics is not the exception.

In the search of developing a grasping synthetic approach for humanoid robots the literature offers a wide range of approaches. However these methods are designed according to a series of constraints such as the available sensors for data acquisition and the system's manipulators. Thus, creating

¹Paola Ardón is with the School of Mathematical and Computer Science, Heriot Watt University and with the School of Informatics, University of Edinburgh, Edinburgh, UK paola.ardon@ed.ac.uk

²Subramanian Ramamoorthy is with the School of Informatics, University of Edinburgh, Edinburgh, UK s.ramamoorthy@ed.ac.uk

³Katrin Solveig Lohan is with the School of Mathematical and Computer Science, Heriot Watt University, Edinburgh, UK k.lohan@hw.ac.uk

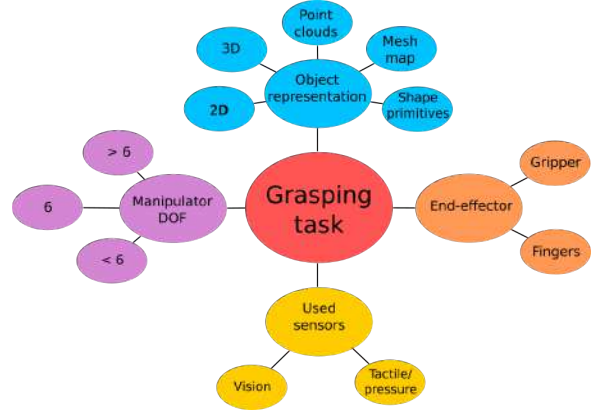


Figure 1. Components considered in this review to achieve the grasping task.

a variety of solutions that cannot be generalized to every robotic platform. Despite of this diversity, all the works consider the same factors depicted in Figure 1, where the grasping task depends on:

- The *object representation*, which can be *two-dimensional (2-D)*, *three-dimensional (3-D)*, represented using point-clouds or mesh maps or a combination of one or more of the previously mentioned models.
- The *end-effector*, which can be either gripper or multi-fingered hand. Common examples are Armar III [7] and Barret robot [82] which are a gripper and multi-finger hand respectively. There is a wide field that focuses on the betterment of the manipulators. Works such as Homberg et al. [47] focuses on the robotic hand in order to improve the end-effector sensor-movement properties for the manipulation task.
- The *used sensors* for the data acquisition, which can vary from vision (monocular, stereo vision, depth cameras) to tactile/pressure sensors that help on the object exploration to asses on the grasp closure.
- The *manipulator's degrees-of-freedom (DOF)*, where the ideal case is having 6-DOF, but it is also common to find end-effectors that have less or more than 6-DOF. In the first case, the system can suffer from singularities, in the latter, it produces redundancy when solving for the inverse kinematic model.

Although there are many aspects that need to be considered in the manipulation task, Rosman and Ramamoorthy [74] agree that is the qualitative structure of the object in an environment and the relationships between them that allow

the efficient construction of a manipulation task. Having this in mind, in this review paper we highlight the object reconstruction techniques used on the available manipulation literature, based on whether these methods consider or not the surrounding environment and other objects. The purpose of this type of review is to facilitate on the process of choosing or creating an object dataset that concisely contains valuable information about the target and how to manipulate it in a given environment. With this objective, we consider the representation in Figure 1 to be the base of the summary presented in Tables I to III where we also indicate if the grasping task considers the object to be isolated or in a cluttered environment.

A. Overview on Grasping Methods

Given the variety of robotic morphologies and applications of the grasping task (pick and place, folding cloth, assemble etc.) we find a variety of approaches in the literature that try to find a solution to the grasping problem. There are many surveys that help to summarize the available literature in the area. Sahbani et al. [76] offers an overview of 3-D object grasp synthesis algorithms. In the field, grasp synthesis is understood as the series of aspects that are considered to carry successfully the grasping task. Sahbani et al. [76] proposes to divide the methods into: analytical and empirical. Analytic formulations have been reviewed by Bicchi and Kumar [15] and the empirical ones by Bohg et al. [17].

1) **Analytical approaches**: focus on the kinematics and dynamics of the robotic system in order to determine the grasp. These methods usually avoid the computation of mathematical and physical models obtained from imitating human grasping strategies. Figure 2 shows the division of these approaches proposed by Sahbani et al. [76]. On the down side, these methods suffer from the variety of errors that rise from the noisy sensor readings and the inaccurate models of the robot kinematics and dynamic. As a result, the pose of the object with respect to the end-effector is not accurate and the task is compromised. The review presented in Bicchi and Kumar [15] reveals there is a lack of literature in this area that deals with positioning errors. According to Sahbani et al. [76], in this type of approaches during the grasping task execution the fingers must be controlled considering the following criteria:

- The end-effector *dexterity* which in robotics is still a wide-open field of research. Cheng [24] offers an overview on the different challenges in the robotics community, where there is a great effort to build end-effectors that look more like human hands.
- The grasp *equilibrium*, which is seen as the immobilization of the grasped object against the possible external disturbance.
- A grasp is considered *stable* if once the object is grasped at equilibrium a small disturbance is applied on the object or fingers and the system comes back to its original configuration.
- The *dynamic behaviour* of the forces acting on the manipulator and the accelerations they produce on the

grasping task.

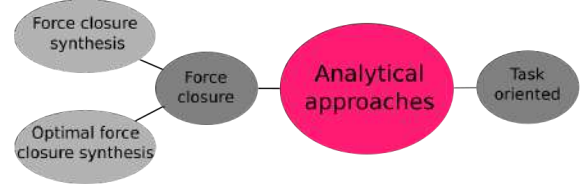


Figure 2. A synthetic view of existing analytical approaches as presented in Sahbani et al. [76]. Analytical approaches can be either force closure or task oriented.

2) **Empirical or data-driven approaches**: data-driven approaches differ on how the set of grasps candidates is sampled and how they discern between grasps. Some of these methods are based on analytic formulations and others are open to human demonstrations, perceptual information or even on heuristics. Figure 3 shows a synthetic view of the data-driven approaches as presented in Bohg et al. [17]. Contrary to analytic approaches, these methods place more weight on the object representation and perceptual processing, e.g. feature extraction, similarity metrics, object recognition or classification and pose estimation. Thus, works falling into this category are the focus of this review.

Ekvall and Kragic [35] and Morales et al. [65] define some criterion that needs to be taken into account to succeed in data-driven approaches:

- The *end-effector centre point* needs to be aligned to the grasping point of the object. Where a grasping point is defined as the place in the object's surface with highest probability to generate a successful grasp.
- The *approach vector* should describe the 3-D angle at which the robotic end-effector should approach the grasping point.
- The *wrist orientation* of the robotic hand.
- The *initial* end-effector configuration.

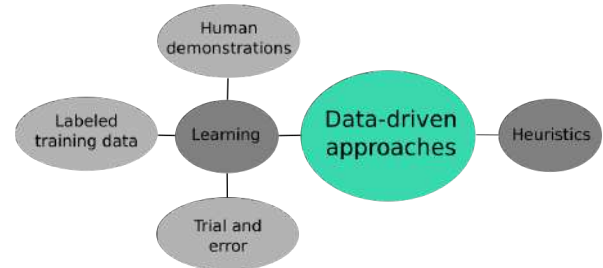


Figure 3. A synthetic view of existing data-driven approaches as presented in Bohg et al. [17]. Data-driven approaches can be either based on heuristics or on learning from data. This data can be provided by labelled data, human demonstration or trough experience of trial and error.

There are different subdivisions to this type of approaches. Sahbani et al. [76] proposes to divide them based on whether they use object features alone or **learning by demonstration (LBD)**, based on humans grasping objects. Bohg et al. [17] on the other hand proposes to divide them according to

the level of *a priori* information regarding the object. In this review we follow this scheme, however we focus on the object reconstruction methodologies that lead to create efficient grasps hypotheses based on whether the objects are recognized in an isolated or in cluttered environment. Following Bohg et al. [17], empirical or data-driven approaches are then sub-divided in:

- *Known objects*: usually in this group the system counts with a database (built offline) containing the object models and a series of good grasps associated with these objects. Thus, to accomplish the grasping task they online estimate the pose of the object.
- *Familiar objects*: in this group it is assumed that the perceived object is similar to one existent in the dataset either in terms of shape, colour, texture or category. Therefore the grasping points associated with the known object can be used on the perceived one.
- *Unknown objects*: in this group there is neither a dataset to associate the object with nor grasping points, thus involving heuristics to extract the local or global features of the objects and their corresponding grasping hypotheses.

Hence, it seems suitable for our object reconstruction review on grasping methods to use these three groups. Once the object has been recognized the grasping task is concerned with the pose estimation, generation of grasping hypotheses, discerning between grasp and other essential aspects for which techniques such as motion planning and machine learning are used, however these others fall out of the scope of this survey.

B. Overview on Datasets

Nowadays datasets are considered a key aspect to link neuroscience and robotics to develop cognitive methods that help towards autonomous behaviours, specially on humanoid robots. However, finding or creating the right dataset can be an extensive process. Huang et al. [49] surveys datasets for the manipulation task, presenting a summary of datasets no older than 10 years. For each reviewed dataset they report on modalities, activities, and annotations. However, these datasets focus on capturing the motion and human actions towards the designated grasping activity and contain little details regarding the object reconstruction. Bianchi et al. [14] explores the datasets presented in the workshop on Grasping and Manipulation. It includes human motion datasets, *instrumental activities of daily living (IADL)*, other activities, object geometry and motion, and haptic interaction datasets. Given to the modularity of the grasping task, there is not benchmarking system or dataset that helps on evaluating the methodologies in the state-of-the-art. The datasets presented in Bianchi et al. [14] are expected to serve as references in the future in this comparative process.

On object reconstruction, object model databases account for information such as object shape and other characteristics such as material, and object weight. Nevertheless, existing methods for constructing 3-D object recognition databases are time and resource consuming, often requiring specialized

equipment. There are online datasets available for this purpose however they are built for specific application therefore might contain limited information.

Additional to the different information that they might contain datasets can differ in the nature of the images. They can be either *synthetic*, computer generated graphics or *real images* which are the ones obtained from the robotic sensors. Some examples of manipulation datasets available online are [3, 53, 69], and Levine et al. [57]. Throughout this review we indicate on Tables I to III the works that have their database online.

C. Outline

This review is divided into: Section II expands on methods that assume the object is known and therefore are handled with a database; Section III shows some of the works that assume the object to be familiar either at a low level (texture, colour, shape) or at a high level e.g. category; Section IV contains methodologies that do not work based on a database, neither for object nor for motion models; and Section V summarizes this work and identifies some issues found in the review literature.

II. REPRESENTING KNOWN OBJECTS

Considering the target to grasp is known, then the task reduces to identifying the pose and analysing the best plan to reach and successfully grab the object. Table I is a summary of the different works presented in this section with the most relevant considered aspects to achieve a grasping task.

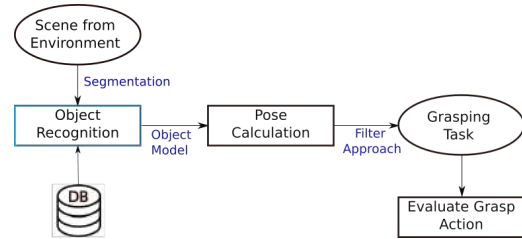


Figure 4. Abstract process flow to achieve a successful grasping with known objects. The perceived object is compared to a database to extract the model so that the system can extract its pose and create the best grasping approach.

Figure 4 is an abstract representation of the grasping task on works that focus on grasping known objects. Generally after perceiving the scene, the obtained image is compared with an offline built database that contains the object model. This model, regardless of its nature, serves to extract the pose of the object. Once the pose has been estimated the system creates a series of grasping hypothesis. Even though the object to grasp is already known, these grasping hypotheses vary from learning techniques such as *LBD*, which the system learns the motion model to approach the object (Argall et al. [4]), to solely motion planning based on the dynamics and kinematics of the robot. Once the system has filtered the best reaching approach it executes the grasping task, which can be classified as successful or unsuccessful according to the applied method metrics.

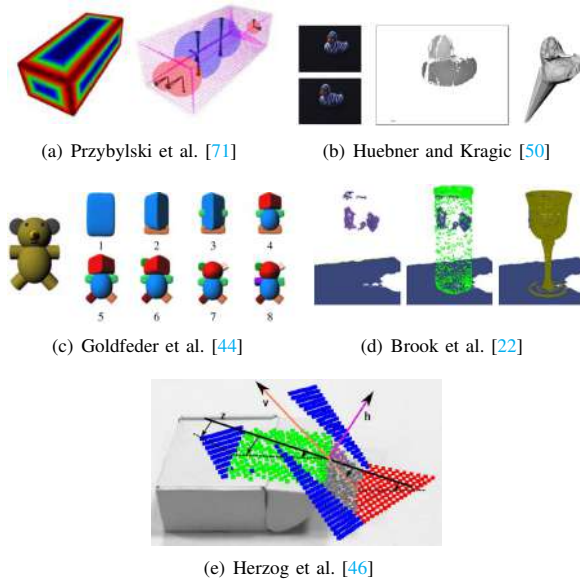


Figure 5. Some examples of object representation for grasping a known target in an isolated environment. a) shape approximation by inscribing spheres. Representation of a rectangular box with object angle spheres ($\alpha = 180^\circ$ in blue and $\alpha = 90^\circ$ in red); b) 3-D points for disparity taken from two stereo images (a duck image) and the resulting polygonal structure; c) decomposition tree with $n=8$, the leaf nodes are merged pairwise to form a binary tree from the bottom up; d) different object representations, one wrong (middle) and two right using point cloud; e) red describes the background, blue are void regions, green is the surface.

A. Isolated Objects

There are many works that present interesting approaches to grasping known objects, specially objects in an isolated environment. This is when there are no more objects or distracting features in the field of view of the robot. Specially in this area, there is an important number of works that emphasize the importance of finding geometric symmetry in the target object, given that it helps on creating an efficient grasp planning.

A grid of medial spheres is the basis for the work presented in Przybylski et al. [71], in which they create a volumetric 3-D model based on the medial axis transform. A representation of their method is presented in Figure 5(a). To create this model they use two different type of datasets: a synthetic one, obtained from Chen mesh dataset [23], and a dataset with objects of their own acquisition. In their work they identify spheres in their object representation that can be used to generate a good grasp. For these spheres they consider two key parameters: the angle and the diameter. The angle is used as an indicator of a sphere's significance for grasp planning, where a big angle describes the object shape and small angles describe the surface. For their grasp planning they pay more attention to the shape. An extension of their work is presented in Asfour et al. [6] where they extend the work to a sliced-map of spheres representation of the object that proves to have higher grasping accuracy than their previous method.

Instead of spheres Huebner and Kragic [50] focuses on box primitives. They present a method that wraps given 3-D data points of an object into primitive box shapes by a fit-and-split algorithm based on minimum volume bounding boxes. Although these box shapes are not able to approximate arbitrary data in a precise manner, it gives efficient clues on planning grasps on arbitrary object parts. An example is shown in Figure 5(b). Goldfeder et al. [44] extends the approach of shape primitives to superquadrics¹ decomposition tree. Their object representation is then a multilevel superquadrics tree created using a decomposition of the initial model. An example of their method is shown in Figure 5(c). However, depending on geometry of the target, this method represents some object parts poorly.

There is a series of works that emphasize the quality of grasping. Borst et al. [20] uses a geometrical representation of the target and take advantage of the tactile and pressure sensors to explore the object and extract the force-closure parameters for the grasping. Even though they succeed at generating different grasping points they limit their dataset to basic objects such as balls, cylinders and boxes. Brook et al. [22] presents a framework that uses the 3-D model and point cloud representations of an object to find a consensus on how the object should be grasped. Their method shows to be robust to adjust to incorrect object recognition, and takes into account the potential grasp executions due to imperfect robot calibration. An example of their method of object modelling is shown in Figure 5(d). Combining learning methods, Pelosof et al. [68] uses support vector machine (SVM)² to select an optimal grasp from the space of grasping parameters of an object. They represent the objects with superquadrics model.

Not many methods work on objects of non-rigid materials. Grasping soft objects brings along a whole set of new challenges such as an increment in the dimensionality configuration and a greater variety of visual appearances for the same object. Maitin-Shepard et al. [62] and Ramisa et al. [72] try to address these issues by focusing on clothing. Maitin-Shepard et al. [62] presents a vision based grasping point detection with the aim to be used at picking up cloth by using solely geometric cues that show to be robust to texture variations. They extract the 2-D corners found using feature extraction and a sampling method to fit the 3-D points extracted using stereo vision. Even though their method shows high reliability, it is limited to folding towels material. In Ramisa et al. [72] they deal with wrinkled clothing. They specially focus on detecting the collars in deformed polo t-shirts. For their method, Ramisa et al. [72], uses a bag-of-features³ based detector that combines the appearance and 3-D geometry features extracted using a Kinect camera.

A considerable amount of literature focuses on learning

¹Superquadrics are a geometrical representation of ellipsoids, parabolas and hyperbolas to an arbitrary power in different dimensions [11].

²Supervised learning method that uses regression analysis and classification to analyse data [27].

³It is a vector of occurrence counts of a vocabulary of local image features [85].

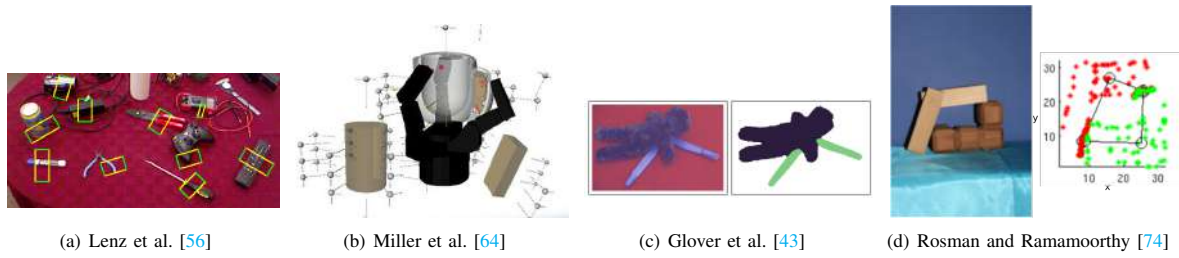


Figure 6. Examples of object representation for the different off-line grasping methods in cluttered environments. a) green lines correspond to robotic gripper plates with detected grasps; b) the balls represent the starting position for the centre of the palm. The long arrow shows the grasp approach direction and the short arrow the thumb direction; c) original and the segmented image into (unknown) objects that are then matched against a known model; d) using contact point networks to discover topological structure in the scene, the algorithmic output is shown alongside, where the points are the support vectors for each object (two wooden blocks connected by a rotary joint), and the open circles are the contact points.

methods that give the robot a more human-like behaviour. A commonly used approach is for a robot to learn from a human how to grasp, these methods fall into the **LBD** category. They use a database of recordings of humans manipulating specific objects to teach a robot the motion model. Herzog et al. [46] (see Figure 5(e)) and Pastor et al. [67] assume that objects with similar shapes can use the same grasping templates. The object model is built with a local shape descriptor constructed using **3-D** information from depth sensors. Instead of focussing on **3-D** data extraction, Balasubramanian et al. [10] and Romero et al. [73] use **2-D** camera to detect objects based on their sizes. The objects are categorized as small, medium and large. In order to plan the grasping, the human teacher manually places the robot's end-effector in the desired final pose commanding it to close the hand while recording the action. In order to add robustness Romero et al. [73] adds exploration with tactile sensors. In their work Faria et al. [38] also take advantage of the use of tactile sensors to reassure the hand closure. They record how humans manipulate simple daily objects and construct probabilistic representation models for the task. The objects to be manipulated are modelled using feature extraction on colour images and stereo depth map. Using **2-D** features to extract **3-D** information Dang and Allen [29] focusses on getting the rotation axis of everyday object manipulation. However, this information is obtained by placing trackers on the handles, in this way extracting the **3-D** model of the object. Li and Pollard [58] proposes a shape matching algorithm that accommodates the sparse shape information associated with the hand pose and the relative placement of the contact points and normals of the object. Focussed on moving objects, Ekvall and Kragic [35] addresses the problem of automatic grasp generation for robotic hands where the shape primitives of the objects are used to provide the basis not only for the grasp generation but also for a grasp evaluation process when there is uncertainty in the object pose. Using affordances cues Stark et al. [79] learns object models based on the hints on these object affordances by extracting the **2-D** local geometric features. Their database focuses on objects with handles taken from one of the ETHZ datasets [1] and object models of their own acquisition.

B. Objects in a Cluttered Environment

A more real-like scenario to perform the grasping task is a cluttered or one where the complete object shape is not visible to the robot. There are some works that try to overcome this challenge, however there are not many.

Lenz et al. [56] proposes a new method for handling multi-modal data in the context of feature learning using deep learning instead of just object detection. In their work they pay special attention to the modality information obtained from the first layer of their network. For it they use **RGB-D** data and not just **2-D** image. The system obtains an **red-green-blue depth (RGB-D)** image from a Kinect sensor mounted on the robot and searches over a large space of possible grasps. For each of these grasps, the method extracts a set of raw features corresponding to the colour and depth of the images as well as the surface normals, then uses these as inputs to a deep network that scores the rectangles, as shown in Figure 6(a). Finally the top-ranked rectangle is selected and the corresponding grasp is executed using the rectangle parameters and the surface normal at its centre. In Miller et al. [64] they simplify the object modelling to shape primitives using spheres, cylinders, cones and boxes. These shapes provide the guidance to a set of grasps starting positions that are tested on the object model, as shown in Figure 6(b). Their algorithm manages to avoid obstacles during the grasping task with a **4-DOF** robotic hand. Their experiments are run in a simulation provided by GraspIt! [63] using the Barret robot hand Townsend and Salisbury [82]. Due to the morphology of the Barret arm, their pre-grasping positions are limited cylindrical and spherical shapes.

An interesting work is presented by Detry et al. [33] which uses probabilistic spatial relations between **3-D** features, organizing these features in a hierarchy. Features at the bottom are bounded to local **3-D** descriptors and the higher-level features are encoded in a probabilistic spatial configuration of more elementary features. Along the same line, Collet et al. [26] takes advantage of a **3-D** feature extraction method to which they apply a **Random sample consensus (RANSAC)** [31] and mean shift algorithm to register multiple instances of an object, thus extracting its pose in a cluttered scene. Glover et al. [43] extends this

method to model partially visible objects, completing their shape using a probabilistic model of shape geometry and a graphical model for performing correspondence between shape descriptors. An example of their method is shown in Figure 6(c).

Ciocarlie et al. [25] focus on household objects. They combine scene interpretation from 3-D range data and tactile sensors to recover from grasp failure identification. The object models are built using a mixture of 3-D point cloud data with 3-D meshes for the recognition. Azad et al. [9] focuses on kitchen objects. Their method deals with textured objects as well as global objects that can be segmented with their shape, using as inputs a 3-D computer-aided design (CAD) model to render the target combined with feature extraction methods. Madry et al. [60] takes advantage of the use of 2-D and 3-D object representation and uses it to generate transferable grasps between objects using categorical knowledge. In their work they segment the scene and from it use feature extraction, contour shape, and 3-D shape descriptor algorithms. Their database contains 14 categories representing everyday objects. Using this database they show to have good results at calculating the possible grasps and transfer these to similar objects among that dataset, however they did not make experiments with a robotic arm.

An interesting work is presented by Rosman and Ramamoorthy [74], where the algorithm does not only consider a cluttered environment but also attempts to learn the relationship between the environment and the object, although they do not experiment on the manipulation task. In their work they are less concerned with the detailed object identification but more interested in separating the scene into potential objects that can be manipulated. They rely on the idea that for a robot in order to efficiently grasp objects in some environment it needs to know something about how these objects relate to each other and to the background information. They restrict to measure the qualitative relationships to the contacts between objects to be *on* and *adjacent*, an example of their method is shown in Figure 6(d). They work with point clouds to represent the scene as a set of layers to extract these contact points. The object segmentation is based on colour information from the point cloud, which is only possible if the objects in the scene are of different colours.

III. REPRESENTING FAMILIAR OBJECTS

In this section we discuss those works that have some level of uncertainty when it comes to represent the objects to grasp. These methods train over a dataset that contains a limited number of objects belonging to specific categories and then are tested on new objects belonging to one of those categories. These type of methods rely on the assumption that objects that are meant for the same purpose (e.g. such as pouring, writing, cutting, etc.) share geometric symmetry, spacial configuration and contain the same of very similar grasping points. Table II shows a summary of the works presented in this section.

Figure 7 is an abstract flow diagram of the common steps observed on architectures that grasp familiar objects.

Table I
Works with off-line learning and their approach to grasp the target object.

Publication	DOI	DB		Recognition				Extraction				EE		
		S	R	2-D	3-D	SP	PC	MM	M	SV	D	E	G	F
Przybylski et al.	7 & 4	✓	✓		✓	✓					✓		✓	✓
Huebner and Kragic	n	✓			✓	✓	✓			✓				✓
Asfour et al.	7	✓	✓		✓	✓		✓			✓		✓	✓
Goldfeder et al.	7	✓			✓	✓							✓	
Maitin-Shepard et al.	7		✓	✓						✓			✓	
Herzog et al.	7 & 4		✓	✓	✓		✓				✓		✓	✓
Pastor et al.	7 & 4		✓	✓	✓				✓			✓		✓
Balasubramanian et al.	7 & 4	✓	✓	✓	✓				✓					✓
Romero et al.	6	✓	✓	✓	✓				✓			✓		✓
Faria et al.	6	✓	✓	✓	✓		✓		✓	✓		✓		✓
Dang and Allen	7		✓	✓					✓			✓		✓
Ekvall and Kragic	7		✓	✓	✓	✓			✓				✓	✓
Stark et al.	7	✓	✓	✓						✓			✓	
Borst et al.	7	✓	✓	✓	✓	✓			✓			✓		✓
Pelosssof et al.	7 & 4	✓	✓		✓		✓	✓						✓
Ramisa et al.	7		✓	✓	✓				✓				✓	
Brook et al.	7	✓	✓		✓		✓		✓				✓	
Lenz et al.***	7		✓	✓	✓				✓		✓		✓	
Miller et al.***	4	✓			✓	✓					✓	✓		
Detry et al.**	6		✓		✓					✓			✓	
Collet et al.**	6		✓	✓	✓				✓				✓	
Glover et al.**	n		✓	✓	✓		✓		✓				✓	✓
Ciocarlie et al.**	6		✓	✓	✓	✓	✓			✓	✓	✓		
Azad et al.**	7	✓	✓		✓	✓		✓		✓				✓
Rosman and Ramamoorthy**	n		✓		✓		✓			✓				
Madry et al.***	n		✓	✓	✓				✓					

data base (DB); synthetic (S); real (R); shape primitives (SP); point clouds (PC); mesh map (MM); monocular camera (M); stereo vision (SV); depth camera (D); exploration with tactile or pressure sensors (E); end-effector (EE); gripper (G); multi-fingered (F); not mentioned in the original work (n); have a dataset available online (*); consider obstacles or cluttered environment (**); consider environment and have an available dataset online (***).

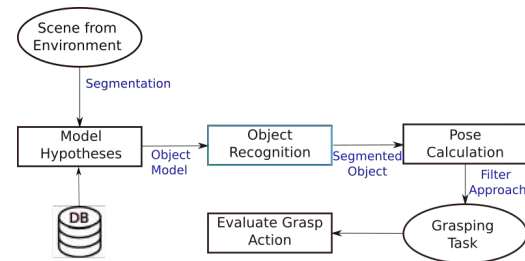


Figure 7. Abstract process flow to achieve a successful grasping with partially known objects. The main difference in this process is that the object recognition is the result of a partial match in the dataset combined with a probabilistic technique that determines the best way to complete the perceived target model.

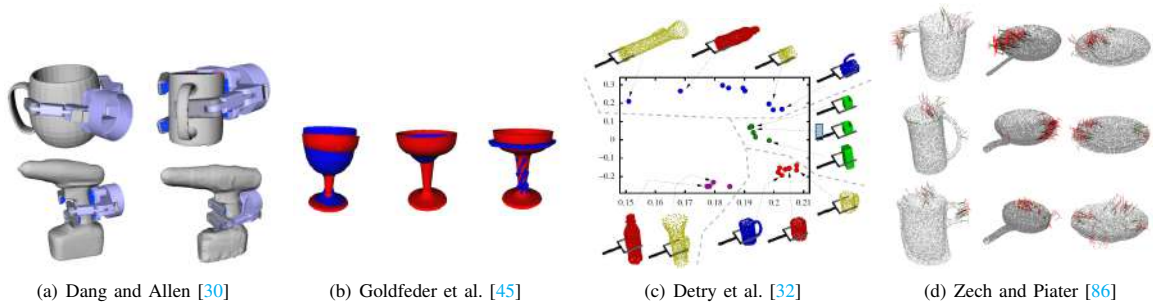


Figure 8. Examples of object representation for the different hybrid grasping methods. a) predefined semantic grasps on target object generated using the proposed method; b) transfer the pre-grasps from the neighbour objects to the sensed objects; c) 2-D approximation of candidates geometric configuration computed with their method, dot colours indicate the data cluster to which a part belongs to; d) results for learning grasps with similar objects, the grasps are rather unevenly distributed when it comes to the familiar object.

A. Isolated Objects

We start with the works that focus on reconstructing familiar objects without any occlusion or other distraction from the environment. Jiang et al. [51] addresses the problem of grasping partially known objects by drawing a rectangle on what the system thinks it is the correct grasping position based on the previously learned models in a dataset. On the obtained image of the perceived object they represent an oriented rectangle that encodes the configuration of their gripper. Their dataset contains nine categories and do the testing on a new object belonging to one of those categories. Dang and Allen [30] proposes to use partial object geometry, tactile contacts and hand kinematics to encode semantic constraints to grasp the objects. In their approach a semantic affordances map is associated with a specific object class therefore it requires similarities between objects as seen in Figure 8(a).

Works such as Goldfeder et al. [45] base their approach solely on 3-D models instead of combining it with tactile exploration. They have a dataset of known 3-D models and use their precomputed grasp data to suggest new grasps on familiar objects. To achieve the task they introduce a new shape descriptor for partial 3-D data range along with partial 3-D models to shape the globally similar, but not identical objects, as shown in Figure 8(b). Bohg et al. [18] model the objects based on the geometric information. They apply a semi-global matching to the stereo views to obtain a dense 3-D reconstruction of the scene. In Bohg and Kragic [16] they extend their approach by applying the concept of shape context. To learn the grasping task they use a supervised learning approach in which the classifier is trained with synthetic images and then tested on images of their own acquisitions.

Detry et al. [32] creates a dictionary of object parts. These parts are identified as the ones containing the best grasping points of the objects. In their method, Detry et al. [32] apply dimensionality reduction and unsupervised clustering algorithms to obtain the size and shape of the part of interest of the object. This learned dictionary allows the agent to grasp familiar objects from the parts the system considers to

be similar. An example of their objects extraction is shown in Figure 8(c). They do the extraction using point clouds to create the 3-D model. The surface segments are extracted using a set of predefined regions of interest (ROI) [84], which are then centred on the gripper. Mahler et al. [61] focus on 3-D object classification between similar objects. They represent the objects using a height-map that is then rendered across the grasp axis.

Saxena et al. [77] proposes an algorithm that does not require neither builds a 3-D model of the object. Instead, they directly predict the shape of the target as a function of 2-D images, focusing on extracting the point to grasp. Their method is trained via supervised learning using synthetic images for the training set. For each object two (or more) images are taken from different camera positions from which they predict the grasping point mainly using three local cues: edges, textures and colour. Avoiding 3-D models as well, Kehoe et al. [52] base their method on cloud computing with shape uncertainty among a class of objects that can be modelled as structured polygons. Their method takes as input a polygon produced from an image of the object using an image contouring algorithm to which they add a Gaussian uncertainty around each vertex and centre of mass of the model to calculate the best grasping approach.

Curtis and Xiao [28] use a dataset that takes into account the geometrical and physical information of the grasping, such as: the rough shape of the object, rough size, weight, material type and combine it with a set of good representative grasps to automate the learning process as much as possible when the familiar object comes along.

Focusing on discerning grasping points on familiar objects Boularias et al. [21] proposes a probabilistic approach where they learn a function that predicts the success grasping probability of a known object represented using 3-D point clouds. Their work is motivated by the fact that points that are geometrically close to each other tend to have similar grasp success probabilities, thus also successfully performing on familiar objects.

Methods such as El-Khoury and Sahbani [36] use LBD techniques to obtain the motion model. El-Khoury and

Sahbani [36] calculates the grasp stability among familiar objects with handles. This stability is obtained by computing contact points, ensuring force-closure on known objects. They use a 3-D model of the target based on the combination of Gaussian curvature-concaveness and watershed segmentation⁴ algorithms. Zech and Piater [86] try to imitate the human-learning process and divide their method into active and transfer learning to grasp familiar objects. Their method is grounded on kernel adaptive, mode-hopping Markov Chain Monte Carlo (MCMC), [42]. In order to represent the objects they use point clouds to extract the 3-D model and calculate the transferable grasps between similar objects, similar in size and shape as seen in Figure 8(d).

B. Objects in a Cluttered Environment

In the works that take into account a cluttered environment for familiar objects we find Hsiao et al. [48] which extends from Brook et al. [22]. They use Bayesian theory [13] to predict the level of success of a grasping action as well as the shape and pose of an object based on an existent dataset. They consider the problem of uncertainty in the acquired data due to noisy sensors. For each hypothesis about the geometry or pose of the object to be grasped they create a different set of grasp plans. To reconstruct the object they use a combination of point clouds and 3-D model, just as in Brook et al. [22], except that now they apply a probabilistic framework to predict the shape of a familiar object.

Ciocarlie et al. [25] presents a framework based on the work presented on Hsiao et al. [48] and Goldfeder et al. [45], where in order to improve the performance of the object recognition in a cluttered place they focus in applying a tight coupling between vision and tactile sensors. Saxena et al. [78] extends from [77] but now assuming noisy readings, this is that only partial faces of the object are visible. They propose a probabilistic model that uses point clouds and the image taken from the object to infer the best configuration of the robotic hand to proceed with the grasp. Following Boularias et al. [21] work, Le et al. [55] extracts the successful grasping points using a point cloud map model of the object and segment the scene using depth information from a depth camera.

IV. REPRESENTING NOVEL OBJECTS

This section presents works that manipulate novel objects, referring to objects that the system has not seen before therefore there is not available model to compare with. These methods do not have a database neither for objects nor for motion models.

Figure 9 is an abstract representation of the general process followed by works in this section. Where these methods build the object shape and grasping hypotheses online based on heuristics. Table III shows a summary of the works presented in this section.

⁴It is a transformation that treats the image as a topographic surface where high intensity denotes peaks and hills while low intensity denotes valleys [83]

Table II

Works with off-line learning and their approach to grasp the target object.

Publication	DOI	DB		Recognition				Extraction				EE		
		S	R	2-D	3-D	SP	PC	MM	M	SV	D	E	G	F
Jiang et al.	7		✓		✓				✓				✓	
Dang and Allen	7		✓			✓		✓			✓	✓		
Goldfeder et al.	7	✓	✓			✓					✓		✓	
Bohg et al.	7	✓	✓	✓	✓					✓				✓
Bohg and Kragic	7	✓	✓	✓	✓					✓				✓
Detry et al.	6	✓	✓	✓	✓		✓		✓				✓	
El-Khouiry and Sahbani	n	✓	✓	✓	✓				✓			✓		
Mahler et al.	n	✓			✓			✓					✓	
Saxena et al.*	6	✓			✓		✓		✓				✓	
Kehoe et al.*	7	✓			✓		✓		✓			✓		
Curtis and Xiao	7	✓	✓	✓	✓	✓			✓				✓	
Zech and Piater	n	✓	✓	✓	✓				✓			✓		
Boularias et al.	7	✓	✓	✓	✓		✓		✓		✓			✓
Hsiao et al.**	7	✓	✓		✓				✓				✓	
Ciocarlie et al.**	7	✓	✓		✓		✓	✓			✓	✓		
Le et al.**	5		✓	✓	✓	✓			✓		✓		✓	
Saxena et al.**	6		✓	✓	✓	✓					✓		✓	✓
	& 7													

data base (DB); synthetic (S); real (R); shape primitives (SP); point clouds (PC); mesh map (MM); monocular camera (M); stereo vision (SV); depth camera (D); exploration with tactile or pressure sensors (E); end-effector (EE); gripper (G); multi-fingered (F); not mentioned in the original work (n); have a dataset available online (*); consider obstacles or cluttered environment (**).

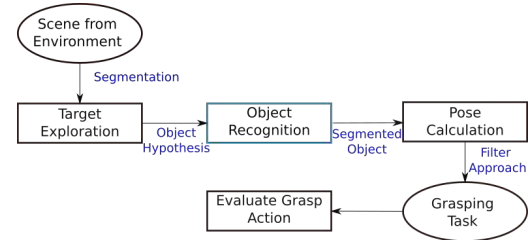


Figure 9. Abstract process flow to achieve a successful grasping with unknown objects. Where now the object modelling and grasping approach are built online based on exploration.

A. Isolated Objects

There are some works such as Faria et al. [39] and Dune et al. [34] that focus solely on object reconstruction without any specific application. Faria et al. [39] uses a 3-D representation of the object using a probabilistic volumetric map derived from in hand exploration. Their procedure uses contour following on the object's surface using the fingertip sensors. This data is then combined with a 3-D point cloud map. Given these two methods, for each voxel they obtain a probability distribution of the occupancy grid and shape of the target. Dune et al. [34] also approaches the problem of modelling unknown objects by choosing the quadric (e.g, a generalization of conic sections such as ellipsoids [11]) that best approximates the object's shape, an example is shown in Figure 10(a). In order to do so, they use multi-view measurements of the object applying a non-linear optimization technique to determine the next best view based on the estimated parameters. Once the view has been selected

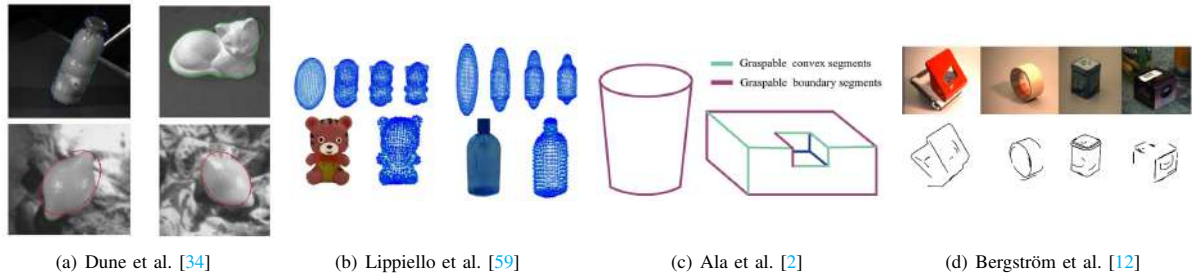


Figure 10. Examples of object representation for the different unknown objects grasping methods. a) some results of contour extraction and conic fitting using active contours; b) steps of the object model reconstruction algorithm for different objects; c) model of objects with graspable boundary and convex segments; d) objects with their match contours to then calculate the grasping hypotheses.

they apply their shape approximation as a combination of contour extraction and conic fitting.

Considering explicitly the grasping application, Lippiello et al. [59] proposes an algorithm composed of the object surface reconstruction and a local grasp planner that evolves in parallel with the object modelling. The reconstruction algorithm uses images taken by a camera placed in the robot's arm and applies a virtual elastic reconstruction surface method around the object. This surface shrinks towards the object until some points intercept the visual hull, eventually taking the object shape as seen in Figure 10(b). Interestingly, their grasp planner runs in parallel with the object reconstruction moving the fingers towards the points that the method considers to be optimal. Also running processes in parallel, Bone et al. [19] integrates online computer vision-based 3-D modelling with an online grasp planning. The object silhouettes are extracted from these images and used to form a 3-D solid model of the object. Their algorithm analyses the model and generates the force closure and pose to grasp the object. Popović et al. [70] builds a hierarchical architecture where the representation of the object is based on edge and texture information. This representation is then used to generate and edge-based and surface-based grasps. Ala et al. [2] algorithm is based on the concept of graspable boundary and convex segments. These are obtained from a single 3-D image from depth sensors. Their method provides graspable segments analysing them geometrically and incorporating memory of grasping experience. An example of their algorithm is shown in Figure 10(c).

Reinforcement learning (RL) [81] is a commonly used technique in artificial intelligence, thus also being applied for grasping tasks. Stulp et al. [80] proposes a model-free based on RL to shape and set the goal parameter positions of an object. The shape of the objects is modelled in 2-D from which the 3-D parameters are obtained and modified along the process until the system considers that the motion primitives are robust enough to proceed with the object pose calculation.

Bergström et al. [12] is based on visual input from stereo camera and their work is more concerned about the quality of the grasp. They reconstruct a wire frame object model through curve matching, as seen in Figure 10(d). From this

model they predict the grasping points to generate a full grasp configuration.

B. Objects in a Cluttered Environment

Representing objects and being able to extract their grasping points is a specially hard task when it comes to many unknown objects in the environment. This still represents a wide field for research. Kroemer et al. [54] presents a hybrid architecture where a controller uses various machine learning methods, including LBD and RL, that cope with a large amount of uncertainty regarding where and how to grasp an object. They focus on cluttered scenes where the scene is represented with early cognitive vision (ECV) [66] descriptors to detect the different objects. Along the same line, Eppner and Brock [37] considers the effect of shape adaptability of the object from visual sensors, however their method uses only range data with which they apply a flood fill segmentation to separate the geometry of the objects at depth discontinuities. Fischinger et al. [40] also takes advantage of a hybrid architecture to detect objects using a point cloud from a single depth camera. They propose a shape-based method that promises to reduce the scene description complexity.

V. DISCUSSION

In the process of reviewing the object reconstruction methods used on data-driven grasping approaches we can find that most of them have the following three major constraints:

A. Generalizing a Grasping Method

There are few works that search on a grasping application beyond pick and place objects. Examples of this minority are Maitin-Shepard et al. [62] and Ramisa et al. [72] who specifically design their methods to manipulate soft materials. Maitin-Shepard et al. [62] focuses on towels and Ramisa et al. [72] on identifying the collars of wrinkled polo t-shirts, both methods with the purpose of folding cloth. Even though these works show to be successful in their tasks, they use a special method for the feature extraction and a specialized motion model that allows the robotic hand to fold, making it difficult to generalize a given methodology to other applications.

Table III

Works with off-line learning and their approach to grasp the target object.

Publication	DOI	DB		Recognition				Extraction				EE		
		S	R	2-D	3-D	SP	PC	MM	M	SV	D	E	G	F
Stulp et al.	7 & 4	✓	✓		✓				✓				✓	
Faria et al.	7	✓	✓		✓			✓	✓			✓		✓
Dune et al.	n	✓		✓	✓	✓			✓					
Lippiello et al.	7		✓	✓	✓				✓			✓		✓
Bone et al.	6		✓		✓			✓	✓				✓	
Popović et al.*	n	n	✓		✓			✓		✓			✓	
Ala et al.	7		✓		✓						✓			✓
Bergström et al.	6 & 7		✓	✓	✓				✓	✓			✓	
Levine et al.*	7		✓	✓	✓				✓					✓
Fischinger et al.**	7		✓	✓	✓		✓				✓		✓	
Eppner and Brock**	7		✓	✓	✓			✓		✓	✓	✓		✓
Kroemer et al.**	7		✓	✓	✓			✓		✓	✓	✓		✓

data base (DB); synthetic (S); real (R); shape primitives (SP); point clouds (PC); mesh map (MM); monocular camera (M); stereo vision (SV); depth camera (D); exploration with tactile or pressure sensors (E); end-effector (EE); gripper (G); multi-fingered (F); not mentioned in the original work (n); have a dataset available online (*); consider obstacles or cluttered environment (**).

Another example of this limitation is shown at the hardware level. In the works presented in this review, we observe that some methods concerned with grasping known objects generalize their technique to the usage of a gripper and a multi-fingered end-effector. Contrary to the ones working with familiar or unknown objects that stay consistent with a type of robotic hand.

B. Benchmarking

This *ad-hoc* characteristic of the different grasping methods also makes it difficult to have a benchmarking process that quantitatively measures the grasping success in the data-driven approaches. Nonetheless, in the analytical ones there are methods such as the one presented in Russell [75] that introduces metrics to categorize a grasping method. They argue that regardless of the actual task, any grasping and manipulation problem can be broken down into kinematic and kinetics, which they define as motion and effort. They make available online the different metrics and how to test a given grasping method that falls into this category.

An alternative for the data-driven approaches is to use benchmark datasets to measure the quality for the object recognition stage. Throughout this work we have indicated a set of benchmark and object reconstruction datasets that are available online. Despite of the existence of these datasets, they cover only part of the process and do not give a deeper insight for the grasping task as a whole.

C. Object Reconstruction and its Context

Through out this review we observe that point cloud, either on 2-D or 3-D, are one of the most commonly used techniques, specially on the methods where there is uncertainty about the model of the object. Reconstructing an

object using this methodology gives already an insight about the grasping points, making the process more efficient.

Most of the approaches assume that the object to be grasped is already segmented from the background. Some of them consider the objects in a cluttered space however none of these methods tries to learn the context of the environment in which they appear. There is a notable lack of works in the manipulation field that try to solve the grasping problem using this approach. Nevertheless, in the object recognition alone we find works that try to tackle the issue. For example, Rosman and Ramamoorthy [74] proposes to split the scene into layers based on point clouds with the purpose to learn the relationship between objects. In their process, they extract the contact points among objects but not of the objects in relation to a scene. Another example is Aydemir and Jensfelt [8], which deals with the problem that most of the existing approaches assume the objects to be on a planar surface. To solve this they propose to extract the correlation between local 3-D structure and object placement in everyday scenes which proves to boost the results of the recognition. Even though they do not apply their method to accomplish other tasks besides the object identification, they make their dataset to be available online so that it can be used on other applications.

These two previously mentioned methods assume the object models to be known, however, a more realistic scenario is when the object model is familiar or unknown. This situation is more likely comparable to the way in which we humans interact with new tools in a given situation. Even though we do not always acknowledge the utility of a partially or completely unknown object we, generally, deduce how to interact with it based on our own experience and the environment in which the object is at. Thus, we are interested on the idea of allowing the robotic system to infer on the target object, either using a dataset (in the case of familiar objects) or through exploration (in the case of novel objects), to create a foundation about the object and then use this as a precedent to learn the context of the environment in which is located.

D. Final Remarks

In this review we discussed object reconstruction techniques for data-driven grasp synthesis approaches based on whether the recognition is being done in an isolated or cluttered environment. For the organization of the methodologies we followed the schema proposed by Bohg et al. [17] where the techniques are divided according to the level of *a priori* information that is available about the object, resulting in known, familiar and novel objects.

The purpose of this review is then to highlight the most popular methods used for object reconstruction that ease on the extraction of good grasping points. This with the motivation of facilitating on the search or creation a dataset that encapsulate in an efficient manner the most relevant information regarding the object and the environment in order to achieve a successful grasping process.

REFERENCES

- [1] ETHZ. <http://www.vision.ee.ethz.ch/en/datasets/>. Accessed: 2018-01-30.
- [2] Ala, R., Kim, D. H., Shin, S. Y., Kim, C., and Park, S.-K. (2015). A 3d-grasp synthesis algorithm to grasp unknown objects based on graspable boundary and convex segments. *Information Sciences*, 295:91–106.
- [3] Alt, N., Xu, J., and Steinbach, E. (2016). A dataset of thin-walled deformable objects for manipulation planning. In *Grasping and Manipulation Datasets (ICRA Workshop)*, Stockholm, Sweden.
- [4] Argall, B. D., Chernova, S., Veloso, M., and Browning, B. (2009). A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483.
- [5] Asada, M., Hosoda, K., Kuniyoshi, Y., Ishiguro, H., Inui, T., Yoshikawa, Y., Ogino, M., and Yoshida, C. (2009). Cognitive Developmental Robotics : A Survey. *IEEE Transactions on Autonomous Mental Development*, 1(1).
- [6] Asfour, T., Przybylski, M., and Dillmann, R. Unions of Balls for Shape Approximation in Robot Grasping.
- [7] Asfour, T., Regenstein, K., Azad, P., Schroder, J., Bierbaum, A., Vahrenkamp, N., and Dillmann, R. (2006). Armar-iii: An integrated humanoid platform for sensory-motor control. In *Humanoid Robots, 2006 6th IEEE-RAS International Conference on*, pages 169–175. IEEE.
- [8] Aydemir, A. and Jensfelt, P. (2012). Exploiting and modeling local 3d structure for predicting object locations. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 3885–3892. IEEE.
- [9] Azad, P., Asfour, T., and Dillmann, R. (2007). Stereo-based 6d object localization for grasping with humanoid robot systems. In *Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on*, pages 919–924. IEEE.
- [10] Balasubramanian, R., Xu, L., Brook, P. D., Smith, J. R., and Matsuoka, Y. (2014). Physical human interactive guidance: Identifying grasping principles from human-planned grasps. *Springer Tracts in Advanced Robotics*, 95(4):477–500.
- [11] Barr, A. H. (1981). Superquadrics and angle-preserving transformations. *IEEE Computer graphics and Applications*, 1(1):11–23.
- [12] Bergström, N., Bohg, J., and Kragic, D. (2009). Integration of visual cues for robotic grasping. In *International Conference on Computer Vision Systems*, pages 245–254. Springer.
- [13] Bernardo, J. M. and Smith, A. F. (2001). Bayesian theory.
- [14] Bianchi, M., Bohg, J., and Sun, Y. (2016). Latest datasets and technologies presented in the workshop on grasping and manipulation datasets.
- [15] Bicchi, A. and Kumar, V. (2000). Robotic grasping and contact: A review. In *Robotics and Automation, 2000. Proceedings. ICRA'00. IEEE International Conference on*, volume 1, pages 348–353. IEEE.
- [16] Bohg, J. and Kragic, D. (2010). Learning grasping points with shape context. *Robotics and Autonomous Systems*, 58(4):362–377.
- [17] Bohg, J., Morales, A., Asfour, T., and Kragic, D. (2013). Data-Driven Grasp Synthesis - A Survey. pages 1–21.
- [18] Bohg, J., Welke, K., León, B., Do, M., Song, D., Wohlkinger, W., Madry, M., Aldóma, A., Przybylski, M., Asfour, T., et al. (2012). Task-based grasp adaptation on a humanoid robot. *IFAC Proceedings Volumes*, 45(22):779–786.
- [19] Bone, G. M., Lambert, A., and Edwards, M. (2008). Automated modeling and robotic grasping of unknown three-dimensional objects. pages 292–298.
- [20] Borst, C., Fischer, M., and Hirzinger, G. (2003). Grasping the dice by dicing the grasp. In *Intelligent Robots and Systems, 2003.(IROS 2003). Proceedings. 2003 IEEE/RSJ International Conference on*, volume 4, pages 3692–3697. IEEE.
- [21] Boularias, A., Kroemer, O., and Peters, J. (2011). Learning robot grasping from 3-d images with markov random fields. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 1548–1553. IEEE.
- [22] Brook, P., Ciocarlie, M., and Hsiao, K. (2011). Collaborative grasp planning with multiple object representations. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 2851–2858. IEEE.
- [23] Chen, X., Golovinskiy, A., and Funkhouser, T. (2009). A benchmark for 3d mesh segmentation. In *Acm transactions on graphics (tog)*, volume 28, page 73. ACM.
- [24] Cheng, G. (2014). *Humanoid robotics and neuroscience: Science, engineering and society*. CRC Press.
- [25] Ciocarlie, M., Hsiao, K., Jones, E. G., Chitta, S., Rusu, R. B., and ucan, I. A. (2014). Towards Reliable Grasping and Manipulation in Household Environments. pages 241–252. Springer, Berlin, Heidelberg.
- [26] Collet, A., Berenson, D., Srinivasa, S. S., and Ferguson, D. (2009). Object recognition and full pose registration from a single image for robotic manipulation. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pages 48–55. IEEE.
- [27] Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.
- [28] Curtis, N. and Xiao, J. (2008). Efficient and effective grasping of novel objects through learning and adapting a knowledge base. In *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, pages 2252–2257. IEEE.
- [29] Dang, H. and Allen, P. K. (2010). Robot learning of everyday object manipulations via human demonstration. *IEEE/RSJ 2010 International Conference on Intelligent Robots and Systems, IROS 2010 - Conference Proceedings*, pages 1284–1289.
- [30] Dang, H. and Allen, P. K. (2012). Semantic grasping: Planning robotic grasps functionally suitable for an object manipulation task. In *Intelligent Robots and Systems*

- (IROS), 2012 IEEE/RSJ International Conference on, pages 1311–1317. IEEE.
- [31] Derpanis, K. G. (2010). Overview of the ransac algorithm. *Image Rochester NY*, 4(1):2–3.
- [32] Detry, R., Ek, C. H., Madry, M., Piater, J., and Kragic, D. (2012). Generalizing Grasps Across Partly Similar Objects. *IEEE International Conference on Robotics and Automation*.
- [33] Detry, R., Pugeault, N., and Piater, J. H. (2009). A probabilistic framework for 3d visual object representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1790–1803.
- [34] Dune, C., Marchand, E., Collowet, C., and Leroux, C. (2008). Active rough shape estimation of unknown objects. In *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, pages 3622–3627. IEEE.
- [35] Ekvall, S. and Kragic, D. (2007). Learning and evaluation of the approach vector for automatic grasp generation and planning. In *Robotics and Automation, 2007 IEEE International Conference on*, pages 4715–4720. IEEE.
- [36] El-Khoury, S. and Sahbani, A. (2008). Handling objects by their handles. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, number EPFL-TALK-168926.
- [37] Eppner, C. and Brock, O. (2013). Grasping unknown objects by exploiting shape adaptability and environmental constraints. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 4000–4006. IEEE.
- [38] Faria, D. R., Martins, R., Lobo, J., and Dias, J. (2010). Probabilistic representation of 3D object shape by in-hand exploration. *IEEE/RSJ 2010 International Conference on Intelligent Robots and Systems, IROS 2010 - Conference Proceedings*, pages 1560–1565.
- [39] Faria, D. R., Martins, R., Lobo, J., and Dias, J. (2012). Extracting data from human manipulation of objects towards improving autonomous robotic grasping. *Robotics and Autonomous Systems*, 60(3):396–410.
- [40] Fischinger, D., Vincze, M., and Jiang, Y. (2013). Learning grasps for unknown objects in cluttered scenes. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 609–616. IEEE.
- [41] Gaver, W. W. (1991). Technology affordances. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 79–84. ACM.
- [42] Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*. CRC press.
- [43] Glover, J., Rus, D., and Roy, N. (2008). Probabilistic models of object geometry for grasp planning. *Proceedings of Robotics: Science and Systems IV, Zurich, Switzerland*, pages 278–285.
- [44] Goldfeder, C., Allen, P. K., Lackner, C., and Pelossof, R. (2007). Grasp planning via decomposition trees. In *Robotics and Automation, 2007 IEEE International Conference on*, pages 4679–4684. IEEE.
- [45] Goldfeder, C., Ciocarlie, M., Peretzman, J., Dang, H., and Allen, P. K. (2009). Data-driven grasping with partial sensor data. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pages 1278–1283. IEEE.
- [46] Herzog, A., Pastor, P., Kalakrishnan, M., Righetti, L., Asfour, T., and Schaal, S. (2012). Template-based learning of grasp selection. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 2379–2384. IEEE.
- [47] Homberg, B. S., Katzschmann, R. K., Dogar, M. R., and Rus, D. (2015). Haptic identification of objects using a modular soft robotic gripper. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 1698–1705. IEEE.
- [48] Hsiao, K., Ciocarlie, M., Brook, P., et al. (2011). Bayesian grasp planning. In *ICRA 2011 Workshop on Mobile Manipulation: Integrating Perception and Manipulation*.
- [49] Huang, Y., Bianchi, M., Liarokapis, M., and Sun, Y. (2016). Recent data sets on object manipulation: A survey. *Big data*, 4(4):197–216.
- [50] Huebner, K. and Kragic, D. (2008). Selection of robot pre-grasps using box-based shape approximation. In *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, pages 1765–1770. IEEE.
- [51] Jiang, Y., Moseson, S., and Saxena, A. (2011). Efficient grasping from RGBD images: Learning using a new rectangle representation. *Proceedings - IEEE International Conference on Robotics and Automation*, pages 3304–3311.
- [52] Kehoe, B., Berenson, D., and Goldberg, K. (2012). Toward cloud-based grasping with uncertainty in shape: Estimating lower bounds on achieving force closure with zero-slip push grasps. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 576–583. IEEE.
- [53] Kent, D., Behrooz, M., and Chernova, S. (2016). Construction of a 3d object recognition and manipulation database from grasp demonstrations. *Autonomous Robots*.
- [54] Kroemer, O., Detry, R., Piater, J., and Peters, J. (2010). Combining active learning and reactive control for robot grasping. *Robotics and Autonomous Systems*, 58(9):1105–1116.
- [55] Le, Q. V., Kamm, D., Kara, A. F., and Ng, A. Y. (2010). Learning to grasp objects with multiple contact points. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 5062–5069. IEEE.
- [56] Lenz, I., Lee, H., and Saxena, A. (2015). Deep learning for detecting robotic grasps. *International Journal of Robotics Research*, 34(4-5):705–724.
- [57] Levine, S., Pastor, P., Krizhevsky, A., and Quillen, D. (2016). Learning hand-eye coordination for robotic grasping with large-scale data collection. In *International Symposium on Experimental Robotics*, pages 173–184. Springer.

- [58] Li, Y. and Pollard, N. S. (2005). A shape matching algorithm for synthesizing humanlike enveloping grasps. In *Humanoid Robots, 2005 5th IEEE-RAS International Conference on*, pages 442–449. IEEE.
- [59] Lippiello, V., Ruggiero, F., Siciliano, B., and Villani, L. (2013). Visual Grasp Planning for Unknown Objects Using a Multifingered Robotic Hand. *IEEE/ASME TRANSACTIONS ON MECHATRONICS*, 18(3).
- [60] Madry, M., Song, D., and Kragic, D. (2012). From object categories to grasp transfer using probabilistic reasoning. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 1716–1723. IEEE.
- [61] Mahler, J., Pokorny, F. T., Hou, B., Roderick, M., Laskey, M., Aubry, M., Kohlhoff, K., Kroger, T., Kuffner, J., and Goldberg, K. (2016). Dex-Net 1.0: A cloud-based network of 3D objects for robust grasp planning using a Multi-Armed Bandit model with correlated rewards. *Proceedings - IEEE International Conference on Robotics and Automation*, 2016-June:1957–1964.
- [62] Maitin-Shepard, J., Cusumano-Towner, M., Lei, J., and Abbeel, P. (2010). Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 2308–2315. IEEE.
- [63] Miller, A. T. and Allen, P. K. (2004). Graspit! a versatile simulator for robotic grasping. *IEEE Robotics & Automation Magazine*, 11(4):110–122.
- [64] Miller, A. T., Knoop, S., Christensen, H. I., and Allen, P. K. (2003). Automatic grasp planning using shape primitives. In *Robotics and Automation, 2003. Proceedings. ICRA'03. IEEE International Conference on*, volume 2, pages 1824–1829. IEEE.
- [65] Morales, A., Chinellato, E., Fagg, A. H., and Del Pobil, A. P. (2004). Using experience for assessing grasp reliability. *International Journal of Humanoid Robotics*, 1(04):671–691.
- [66] Nagel, H.-H. (2004). Steps toward a cognitive vision system. *AI magazine*, 25(2):31.
- [67] Pastor, P., Righetti, L., Kalakrishnan, M., and Schaal, S. (2011). Online movement adaptation based on previous sensor experiences. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 365–371. IEEE.
- [68] Pelossof, R., Miller, A., Allen, P., and Jebara, T. (2004). An svm learning approach to robotic grasping. In *Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on*, volume 4, pages 3512–3518. IEEE.
- [69] Pokorny, F. T., Bekiroglu, Y., Pauwels, K., Butepage, J., Scherer, C., and Kragic, D. (2017). A database for reproducible manipulation research: Capridb-capture, print, innovate. *Data in brief*, 11:491–498.
- [70] Popović, M., Kootstra, G., Jørgensen, J. A., Kragic, D., and Krüger, N. (2011). Grasping unknown objects using an early cognitive vision system for general scene understanding. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 987–994. IEEE.
- [71] Przybylski, M., Asfour, T., and Dillmann, R. (2011). Planning grasps for robotic hands using a novel object representation based on the medial axis transform. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 1781–1788. IEEE.
- [72] Ramisa, A., Alenya, G., Moreno-Noguer, F., and Torras, C. (2012). Using depth and appearance features for informed robot grasping of highly wrinkled clothes. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 1703–1708. IEEE.
- [73] Romero, J., Kjellstrom, H., and Kragic, D. (2009). Modeling and evaluation of human-to-robot mapping of grasps. In *Advanced Robotics, 2009. ICAR 2009. International Conference on*, pages 1–6. IEEE.
- [74] Rosman, B. and Ramamoorthy, S. (2011). Learning spatial relationships between objects. *The International Journal of Robotics Research*, 30(11):1328–1342.
- [75] Russell, D. L. (2014). Performance metrics and benchmarks to advance the state of robotic grasping.
- [76] Sahbani, A., El-Khoury, S., and Bidaud, P. (2012). An Overview of 3D Object Grasp Synthesis Algorithms. *Robotics and Autonomous Systems*. 60(3):326–336.
- [77] Saxena, A., Driemeyer, J., Kearns, J., and Ng, A. Y. (2007). Robotic grasping of novel objects. In *Advances in neural information processing systems*, pages 1209–1216.
- [78] Saxena, A., Wong, L. L. S., and Ng, A. Y. (2008). Learning Grasp Strategies with Partial Shape Information. *Aai*, 3(2):1491–1494.
- [79] Stark, M., Lies, P., Zillich, M., Wyatt, J., and Schiele, B. (2008). Functional Object Class Detection Based on Learned Affordance Cues. In *Computer Vision Systems*, pages 435–444. Springer Berlin Heidelberg, Berlin, Heidelberg.
- [80] Stulp, F., Theodorou, E., Kalakrishnan, M., Pastor, P., Righetti, L., and Schaal, S. (2011). Learning motion primitive goals for robust manipulation. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 325–331. IEEE.
- [81] Sutton, R. S. and Barto, A. G. (1998). *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.
- [82] Townsend, W. T. and Salisbury, J. K. (1993). Mechanical design for whole-arm manipulation. In *Robots and Biological Systems: Towards a New Bionics?*, pages 153–164. Springer.
- [83] Vincent, L. and Soille, P. (1991). Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):583–598.
- [84] Vu, K., Hua, K. A., and Tavanapong, W. (2003). Image retrieval based on regions of interest. *IEEE Transactions on knowledge and data engineering*, 15(4):1045–1049.
- [85] Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, pages 977–984. ACM.

- [86] Zech, P. and Piater, J. (2016). Active and transfer learning of grasps by sampling from demonstration.

BIBLIOGRAPHY

- [1] Ardón, P., Dragone, M., and Erden, M. S. (2018a). Reaching and grasping behaviours by humanoid robots through visual servoing. In *Haptics: Science, Technology and Applications, Springer International Publishing AG*, pages 353–365. Springer Nature.
- [2] Ardón, P., Ramamoorthy, S., and Lohan, K. S. Learning object reconstruction for grasping: a survey. To be submitted.
- [3] Ardón, P., Ramamoorthy, S., and Lohan, K. S. (2018b). Object affordances by inferring on the surroundings. In *Workshop on Advanced Robotics and its Social Impacts (ARSO)*. IEEE. Forthcoming.
- [4] Beeck, H. d., Torfs, K., and Wagemans, J. (2008). Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway. *Journal of Neuroscience*, 28(40):10111–10123.
- [5] Bernardo, J. M. and Smith, A. F. (2001). Bayesian theory.
- [6] Boissonnat, J.-D. and Geiger, B. (1993). Three-dimensional reconstruction of complex shapes based on the delaunay triangulation. In *Biomedical Image Processing and Biomedical Visualization*, volume 1905, pages 964–976. International Society for Optics and Photonics.
- [7] Bonaiuto, J. and Arbib, M. A. (2015). Learning to grasp and extract affordances: the Integrated Learning of Grasps and Affordances (ILGA) model. *Biological cybernetics*, 109(6):639–669.
- [8] Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.
- [9] Delingette, H. (1999). General object reconstruction based on simplex meshes. *International journal of computer vision*, 32(2):111–146.

-
- [10] Detry, R., Baseski, E., Popovic, M., Touati, Y., Kruger, N., Kroemer, O., Peters, J., and Piater, J. (2009). Learning object-specific grasp affordance densities. In *Development and Learning, 2009. ICDL 2009. IEEE 8th International Conference on*, pages 1–7. IEEE.
- [11] Do, T.-T., Nguyen, A., and Reid, I. (2018). Affordancenet: An end-to-end deep learning approach for object affordance detection. In *International Conference on Robotics and Automation (ICRA)*.
- [12] Farhadi, A., Endres, I., Hoiem, D., and Forsyth, D. (2009). Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1778–1785. IEEE.
- [13] Freund, Y. and Mason, L. (1999). The alternating decision tree learning algorithm. In *icml*, volume 99, pages 124–133.
- [14] Geib, C., Mourao, K., Petrick, R., Pugeault, N., Steedman, M., Krüeger, N., and Wörgötter, F. (2006). Object action complexes as an interface for planning and robot control. In *IEEE RAS International Conference on Humanoid Robots*.
- [15] Goldfeder, C., Allen, P. K., Lackner, C., and Pelossof, R. (2007). Grasp planning via decomposition trees. In *Robotics and Automation, 2007 IEEE International Conference on*, pages 4679–4684. IEEE.
- [16] Greeno, J. G. (1994). Gibson’s affordances. *American Psychological Association*.
- [17] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- [18] Hermans, T., Rehg, J. M., and Bobick, A. (2011). Affordance prediction via learned object attributes. In *IEEE International Conference on Robotics and Automation (ICRA): Workshop on Semantic Perception, Mapping, and Exploration*, pages 181–184. Citeseer.
- [19] Jaklic, A., Leonardis, A., and Solina, F. (2013). *Segmentation and recovery of superquadrics*, volume 20. Springer Science & Business Media.
- [20] Kraft, D., Detry, R., Pugeault, N., Baseski, E., Piater, J. H., and Krüger, N. (2009). Learning objects and grasp affordances through autonomous exploration. In *ICVS*.

-
- [21] Lai, K., Bo, L., Ren, X., and Fox, D. (2012). Detection-based object labeling in 3d scenes. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 1330–1337. IEEE.
- [22] Lee, D.-T. and Schachter, B. J. (1980). Two algorithms for constructing a delaunay triangulation. *International Journal of Computer & Information Sciences*, 9(3):219–242.
- [23] Lenz, I., Lee, H., and Saxena, A. (2015). Deep learning for detecting robotic grasps. *International Journal of Robotics Research*, 34(4-5):705–724.
- [24] Madry, M., Song, D., and Kragic, D. (2012). From object categories to grasp transfer using probabilistic reasoning. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 1716–1723. IEEE.
- [25] Metta, G., Sandini, G., Vernon, D., Natale, L., and Nori, F. (2008). The icub humanoid robot: an open platform for research in embodied cognition. In *Proceedings of the 8th workshop on performance metrics for intelligent systems*, pages 50–56. ACM.
- [26] Moldovan, B., Moreno, P., van Otterlo, M., Santos-Victor, J., and De Raedt, L. (2012). Learning relational affordance models for robots in multi-object manipulation tasks. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 4373–4378. IEEE.
- [27] Montesano, L. and Lopes, M. (2009). Learning grasping affordances from local visual descriptors. In *Development and Learning, 2009. ICDL 2009. IEEE 8th International Conference on*, pages 1–6. IEEE.
- [28] Montesano, L., Lopes, M., Bernardino, A., and Santos-Victor, J. (2008). Learning object affordances: From sensory–motor coordination to imitation. *IEEE Trans. Robotics*, 24:15–26.
- [29] Moré, J. J. (1978). The levenberg-marquardt algorithm: implementation and theory. In *Numerical analysis*, pages 105–116. Springer.
- [30] Nguyen, A., Kanoulas, D., Caldwell, D. G., and Tsagarakis, N. G. (2017). Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

- [31] Oztop, E., Bradley, N. S., and Arbib, M. A. (2004). Infant grasp learning: a computational model. *Experimental brain research*, 158(4):480–503.
- [32] Pairet, È., Mistry, M., and Brox, F. (2018). Learning and generalisation of primitives skills towards robust dual-arm manipulation. In *AAAI Fall Symposium Series 2018*. AAAI Press. Forthcoming.
- [33] Pelossof, R., Miller, A., Allen, P., and Jebara, T. (2004). An svm learning approach to robotic grasping. In *Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on*, volume 4, pages 3512–3518. IEEE.
- [34] Piaget, J. and Cook, M. (1952). *The origins of intelligence in children*, volume 8. International Universities Press New York.
- [35] Quattoni, A. and Torralba, A. (2009). Recognizing indoor scenes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 413–420. IEEE.
- [36] Saxena, A., Driemeyer, J., and Ng, A. Y. (2008). Robotic grasping of novel objects using vision. *I. J. Robotics Res.*, 27:157–173.
- [37] Sridharan, M. (2017). Integrating knowledge representation, reasoning, and learning for human-robot interaction. pages 69–76. AAAI Fall Symposium. Artificial Intelligence for Human-Robot Interaction.
- [38] Stoytchev, A. (2005). Toward learning the binding affordances of objects: A behavior-grounded approach. In *Proceedings of AAAI symposium on developmental robotics*, pages 17–22.
- [39] Sung, J., Lenz, I., and Saxena, A. (2017). Deep multimodal embedding: Manipulating novel objects with point-clouds, language and trajectories. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 2794–2801. IEEE.
- [40] Vedaldi, A. and Lenc, K. (2015). Matconvnet: Convolutional neural networks for matlab. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 689–692. ACM.
- [41] Vezzani, G., Pattacini, U., and Natale, L. (2017). A grasping approach based on superquadric models. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 1579–1586. IEEE.

- [42] Wächter, A. and Biegler, L. T. (2006). On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical programming*, 106(1):25–57.
- [43] Wertsh, J. V. and Tulviste, P. (1990). Apprenticeship in thinking: Cognitive development in social context. *Science*, 249(4969):684–686.
- [44] Zech, P. and Piater, J. (2016). Active and transfer learning of grasps by sampling from demonstration.
- [45] Zhu, Y., Fathi, A., and Fei-Fei, L. (2014). Reasoning about object affordances in a knowledge base representation. In *European conference on computer vision*, pages 408–424. Springer.