Learning Grasp Affordance Reasoning through Semantic Relations

Paola Ardón*, Èric Pairet*, Ronald P. A. Petrick†, Subramanian Ramamoorthy*, and Katrin S. Lohan†

Abstract—**Reasoning about object affordances allows an autonomous agent to perform generalised manipulation tasks among object instances. While current approaches to grasp affordance estimation are effective, they are limited to a single hypothesis. We present an approach for detection and extraction of multiple grasp affordances on an object via visual input. We define semantics as a combination of multiple attributes, which yields benefits in terms of generalisation for grasp affordance prediction. We use Markov Logic Networks to build a knowledge base graph representation to obtain a probability distribution of grasp affordances for an object. To harvest the knowledge base, we collect and make available a novel dataset that relates different semantic attributes. We achieve reliable mappings of the predicted grasp affordances on the object by learning prototypical grasping patches from several examples. We show our method's generalisation capabilities on grasp affordance prediction for novel instances and compare with similar methods in the literature. Moreover, using a robotic platform, on simulated and real scenarios, we evaluate the success of the grasping task when conditioned on the grasp affordance prediction.**

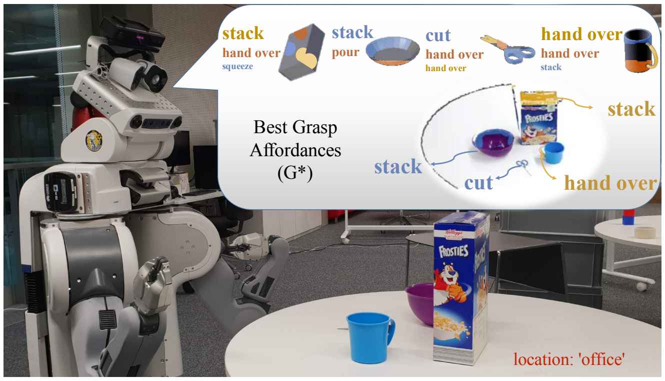


Fig. 1: PR2 reasoning about grasp affordances of objects on a tabletop office scenario. The affordances are colour coded with the corresponding grasping region on the objects. On the top three affordance labels, the larger the size, the more suitable that affordance is in the perceived context.

I. Introduction

Modern robotic platforms are capable of performing a rich set of human-scale manipulation tasks. Affordance is one of the key concepts that enables an autonomous agent to interact with a variety of objects successfully. Affordance refers to the possibility of performing different actions with an object [1]. By associating context and previous experiences, humans are very effective at creating grasp affordance relations to facilitate an intended action. For example, grasping a pair of scissors from the tip affords handing over the tool, but not a cutting task. In the context of robotics, grasp affordances have attained new relevance as agents should be able to manipulate novel objects for tasks with distinct contextualisations.

The current literature offers solutions that are successful in real-world scenarios, but typically assign a single universal grasp affordance for a given object no matter the context of the scene [2]–[5]. In reality, a single object affords different actions, and the successful accomplishment of the task is dependant on identifying the correct grasping region of the object. Nonetheless, in robotics, there is a relational gap between the interaction of different object categories associated with changing scenarios and pose-grasps. The missed connection between objects and grasp relationships has resulted in (i) a tendency only to consider a single grasp affordance per object, and (ii) a lack of datasets that take into account the relational aspects of grasp affordance. On the grounds of the limitations mentioned above, the contribution of our approach is threefold. First, we present an approach for multi-target prediction of grasp affordances on an object using Markov Logic Network (MLN) theory to build a knowledge base (KB) graph representation. Our method is able to reason about the most probable grasp affordance, among a set, by inferring the context semantics relation using Gibbs sampling. Second, to test the prediction on the grasping task, we map the obtained grasp affordances to the three-dimensional (3-D) data of the object. The system learns the object shape context and related prototypical grasping patches to create hypotheses of grasp locations. The most probable grasp affordance is then chosen to generate a reaching and grasping configuration plan. Finally, we collect and make available a new dataset for visual grasp affordance prediction[1] that promotes more robust and heterogeneous robotic grasping methods. The dataset contains different attributes from 30 different object classes. Each instance is related not only to the semantic descriptions, but also to the physical features describing visual attributes, locations, and different grasping regions for a variety of actions.

In addition, we also compare the generalisation of the grasp affordance predictions on novel objects against current state-of-the-art techniques. The reliability of the obtained grasp affordance regions is evaluated using similarity metrics. We compare these calculated hypotheses with the ground truth labels obtaining high correlation values. We analyse how feasible our approach is for a general tabletop scenario, as shown in Fig. 1, with known and novel objects in simulated and real indoor scenes.

Edinburgh Centre for Robotics. University of Edinburgh and Heriot-Watt University. Edinburgh, UK. †{r.petrick;k.lohan}@hw.ac.uk
*{paola.ardon;eric.pairet;s.ramamoorthy}@ed.ac.uk;

[1] Data and code: https://paolaardon.github.io/grasp_affordance_reasoning/

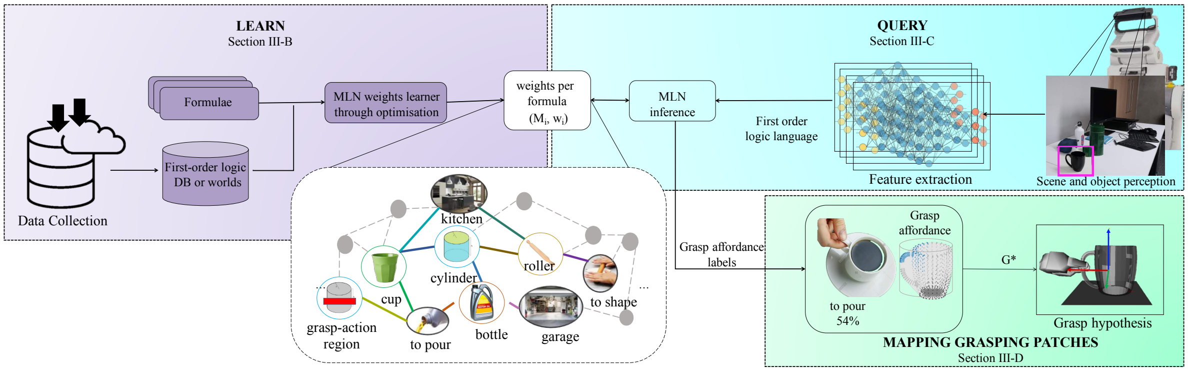


Fig. 2: Proposed framework for reasoning about object grasp affordances, composed of the learning, querying and mapping tasks. The learnt model (white box) encodes the relation between nodes with connecting coloured edges.

II. Related Work

Learning visual grasp affordances for objects has been an active area of research for robotic manipulation tasks. Ideally, an autonomous agent should be able to distinguish between objects and their utility. In real-world scenarios, a single object affords different actions, and the successful completion of the task is dependent on the correct choice of the grasping region. The literature on visually detecting robotic graspings is vast. The state-of-the-art in this area, [6], [7], offers robust methodologies for identifying candidate grasps, either with architectures based on deep learning to detect grasp areas on an object [6], or using supervised learning to obtain grasping points based on objects shape [7]. While these techniques offer robust grasp candidates, they do not differentiate among actions at those detected grasp points.

***Grasp affordances**:* Work on grasp affordances tends to focus on robust interactions between objects and the autonomous agent. However, it is typically limited to a single affordance per object. Moreover, affordance labels tend to be assigned arbitrarily instead of through data-driven techniques gathering human judgement to portray socially acceptable interactions regarding the grasps. Some works, such as [8], focus on relating abstractions of sensory-motor processes with object structures (e.g., object-action complexes (OACs)) to extract the best reaching and grasping candidate given an object affordance. Others use purely visual input to learn affordances using deep learning [5], [9] or supervised learning techniques to relate objects and actions [2]–[4], [10]–[12]. In contrast, our approach reasons about a set of affordance possibilities for an object using data-driven techniques to discern the best grasping approach to succeed at a given action task.

***Datasets**:* At present, no dataset offers an end-to-end relation between objects and grasp affordances. Some datasets relate objects with actions for affordance detection [5], [9]. Others offer a mapping of robust grasps labels for different objects without considering the actions [13]. Not having a dataset that brings together both concepts represents a problem for any methodology that aims to achieve a robust social human-robot interaction architecture for manipulation tasks. In contrast, we harvest a dataset that includes object locations, grasp labels and semantics as relational attributes, encouraging more robust and heterogeneous robotic grasping methods.

***Knowledge bases (KB)**:* A knowledge base refers to a repository of entities and rules that can be used for storing and querying objects' affordance information. MLNs [14] represent the current state-of-the-art method when it comes to reasoning about objects using knowledge bases. An MLN is a combination of Markov Random Fields (MRF) with a first-order logic (FOL) language. [15] uses MLN to learn the optimal weights that relate different object descriptions extracted with a ranking Support Vector Machine (SVM) function. On the other hand, [16] trains a battery of L1-linear regularised logistic classifiers and learn the attributes relation according to the classification score. The performance of the KBs using MLN has been shown to outperform alternatives [17], [18] given the Markovian ability to relate attributes. Using MLN in KBs is advantageous as it can incorporate the uncertainty of probabilistic graphical models. This performance depends on the quality and correlation of the data used for training.

In contrast to current approaches in the field, we collect a detailed dataset that promotes robust grasp techniques. We use this dataset to reason about an object's multiple reliable grasps, corresponding to actions resulting from the design of a KB based on MLN.

III. Proposed Method

Our primary task is to reason about feasible grasps in an object that are closely related to the success of an affordance task. The grasp affordance relationship is built using semantics as a collection of attributes (as explained in Section IV-A). Fig. 2 shows a summary of our proposed methodology as follows: (i) we learn the semantics relation between attributes, locations and grasp affordances through a unique building of grounding and combination of rules using MLN, (ii) we query an approximation of the probability distribution associated with grasp affordances using Gibbs sampling [19], and (iii) among all the possibilities, we take the one that satisfies a given affordance with the highest probability. This selected grasp affordance region is then located on the 3-D object data to calculate a grasping configuration.

A. Knowledge Base Terminology

A knowledge base can be represented as a graph, similar to the white block in Fig. 2. The nodes denote the entities and the edges the general rules that characterise their relationship. For example, a cup is a node or entity connected to other nodes depicting its visual attributes, its affordance (such as *pour*) and the corresponding grasping region. These entities are connected with edges of different colours representing the different weights. The higher the weight, the more likely that relation is to be true. We build the KB by learning these relations, i.e. the weights of the general rules. We employ an MLN [14] for knowledge representation. To construct a KB with MLN there is a pre-learning process: the first step is to collect evidence (as detailed in Section IV-A), in the form of a set of facts and assertions about the entities. The different sets of assertions create possible *worlds*. For example, scissors have metal blades and handles. These two assertions create a world where objects having these two characteristics are likely to be a pair of scissors. The second step is to create a general set of rules. Each of these rules is a *formula* M_i associated with a *weight* w_i, thus creating correlated pairs (M_i, w_i). The formulae are built by creating a relation between the entities. In an MLN, the entities are *terms* and the relation between terms are *predicates*. Table I shows examples of possible *predicates* and *formulae* in our KB. For example, the *predicate* "hasShape" is a relation between the terms "object" and "shape". All the terms, except "object", are considered *grounded terms* since we know their domain representation.

B. Learning Grasp Affordance Relations

The possible *worlds* x that we collect (Section IV-A) are used for learning the formulae' weights and are translated into FOL predicates to form formula sets M. Table I shows some examples of allowable combinations of predicates used in our KB. The location, category, grasping region and visual attributes (i.e., texture, material, shape) are treated as constants and the object as a variable. Given the different sets of constants inside each term, different networks are produced. These networks are of widely varying sizes, but all grounded terms of each formula M_i have the same weight w_i. The weights are then learned generatively using the available possible *worlds* x by calculating their joint distribution as:

$$P(X=x)=\frac{1}{Z}\exp\Big(\sum_{i=1}^{n}w_i f_i(x_{\{i\}})\Big), \tag{1}$$

where Z is the normalisation constant over the potential functions ϕ_i of connected nodes given by $\sum_{x\in\mathcal{X}}\prod_i \phi_i(x_{\{i\}})$, n is the number of formulae in M, $x_{\{i\}}$ is the state of the grounded terms (i.e., the state of existing terms that appear in that world) in M_i, and the feature function $f_i(x_{\{i\}})=1$ if $M_i(x_{\{i\}})$ is true or 0 otherwise. The weights w_i indicate the likelihood of the formula being true. Using Broyden's method [19] we learn the optimal weights w^* from maximising the pseudo-log-likelihood $logP^*_w(X=x)$ of the obtained probability distribution of the available worlds. Table I shows some examples of the learned relations (M_i, w_i) in the KB. For example, a container is ten times more likely to afford a pouring task than an object categorised as electronic.

LEARNING	
Predicates	**Formula-weights** (M_i, w_i)
hS: hasShape(obj, shape)	
hT: hasTexture(obj, texture)	(hS $\wedge$ hT $\wedge$ hM $\Rightarrow$ hA, w_1)
hM: hasMaterial(obj, material)	(hS$\wedge$hT$\wedge$hM$\wedge$cF $\Rightarrow$ hA, w_2)
cF: canBeFound(obj, location)	...
hA: hasAffordance(obj, affordance)	(hC $\Rightarrow$ hA, w_{n-1})
hC: hasCategory(obj, category)	(hA $\Rightarrow$ gR, w_n)
gR: graspRegion(obj, region)	
EXAMPLES	
(hasCategory(obj, container) $\Rightarrow$ hasAffordance(obj, pour), $log(0.67)$)	
(hasCategory(obj, electronics) $\Rightarrow$ hasAffordance(obj, pour), $log(0.07)$)	

TABLE I: Knowledge base schema for the learning task. Our formulae are defined as relations between predicates. The examples give an idea of the learned relations with weights.

C. Reasoning about Grasp Affordances

To reason about an object's grasp affordances, we use the two-dimensional (2-D) image and pass it through a deep learning architecture built with pre-trained Convolutional Neural Networks (CNNs) [20] to extract the objects' attributes (i.e., shape, texture, material, category, location) as labels. These labels are translated into FOL to create possible worlds. These worlds then serve to query the most feasible grasping region for an affordance task. Table II shows an example of probable relations between affordance and grasping region labels (i.e., 1, 2 or 3) given a grasp affordance query. In the KB the more formulae a world adheres to, the more probable it is. In order to query grasp affordances from the learned weights model, we use Gibbs sampling [21]. We employ Gibbs sampling to generate posterior samples by sweeping through each grounded term while keeping the calculations tractable. We compute the expectation of a posterior distribution as:

$$E[h(s)]_{\mathcal{P}}\approx\frac{1}{n}\sum_{i=1}^{n}h(s^{(i)}), \tag{2}$$

where n is the number of grounded simulated samples from that distribution, $\mathcal{P}$ is the posterior distribution of the world of interest, $h(s)$ is the desired expectation and $h(s^{(i)})$ is the i^{th} simulated sample from $\mathcal{P}$. This inference method gives us the maximum probability for different grasping regions for every query $q\in\boldsymbol{Q}$. The corresponding labels (see example in Table II) are used to map the resulting grasping affordances on the object 3-D data. Among these grasp affordances, the most likely one is used as the optimal grasping patch label $G^* = \arg\max_n E[h(s)]$ on which we calculate a grasping configuration to be sent to the robot.

QUERY, given the attributes of a cup:
hasAffordance(obj, x) $\wedge$ graspRegion(obj, x)
(hasAffordance(obj, stack) $\wedge$ graspRegion(obj, 1), 49%)
(hasAffordance(obj, hand over) $\wedge$ graspRegion(obj, 3), 17%)
(hasAffordance(obj, pour) $\wedge$ graspRegion(obj, 2), 22%)

TABLE II: Example of a query and the top answers given an object's attributes presented as assertions that build a world.

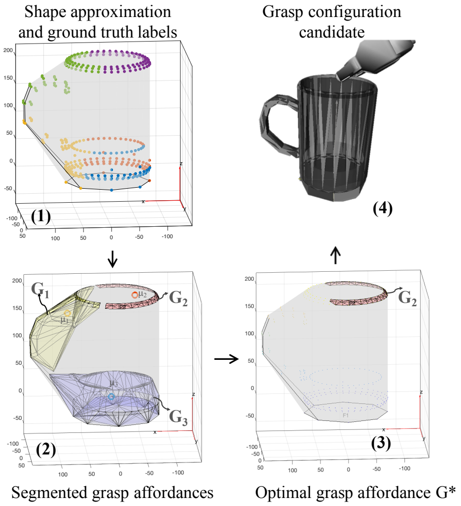

Shape approximation and ground truth labels
Grasp configuration candidate
(1)
(4)
G1
G2
G3
(2)
(3)
Segmented grasp affordances
Optimal grasp affordance G*

Fig. 3: Mapping grasp affordance patches on 3-D object data.

D. Mapping Grasp Affordance Patches

Our goal is to produce robust manipulation by reasoning about the best possible grasp for a given affordance. Towards this objective, we map the previously obtained grasp affordance probabilities to the object's 3-D data. We use a combination of the object shape context and hierarchical segmentation of the point cloud. To ensure the grasping regions on the object are reliable, we adopt as a starting point the pre-defined grasping regions from [13] (2-D with corresponding 3-D mapping) and use them as ground truth labels. Nonetheless, after parsing the data collected in Section IV-A, not all the ground truth labels (colour coded labels from [13] in Fig. 3, step **(1)**) afford an action or different labels could afford the same action. Hence, we redefine sets of ground truth labels into our grasp affordance patches. Using the ground truth data in [13], the points are clustered using the k-means algorithm, extracting n grasping regions (as probabilities obtained from the KB querying step). The cluster centres μ_G serve as the seeds for the representative initial patches G. A set of faces forms these patches. The faces are grouped using hierarchical mesh decomposition as proposed in [22]. The mesh decomposition produces G_n grasping patches as shown in step **(2)** in Fig. 3. We consider the features belonging to the G_n grasping patches as inputs and classify them in n grasp affordance candidate labels based on the object's context shape using an SVM classifier as done in [6], [7]. The optimal grasping patch label G^* is represented by a set $\boldsymbol{b}$ of 3-D points which have a dominant plane $\widehat{\Pi}$ with centroid $\boldsymbol{\nu_C}$ and orientation γ that serves as the position and orientation for the inverse kinematics calculation of the grasping approach (Fig. 3, step **(3)** and **(4)**).

E. End-to-end Execution

Algorithm 1 presents an outline of the framework's end-to-end execution, which aims to provide a robotic platform with a feasible grasp subject to a desired affordance. Given visual perception of the environment, the desired affordance, and the pre-trained models for label extraction (see Section III-C), region extraction (see Section III-D), and semantic relations (see Section III-B) (line 2 to 5), the end-to-end execution is as follows. First, the visual data is processed to extract the labels and map them into a binary vector, where the non-zero entries indicate the presence of an attribute (line 8) and the feature-label map (line 9) describing the object to manipulate. The extracted labels are used to define an FOL world and query the KB model, thus inferring a set of grasp affordance relations *GA_r* (line 10). *GA_r* indicates the highest affordance probability per grasping region. From this set and given the desired *affordance*, the framework calculates the most suitable *region_label**. If the *affordance* is not chosen, the framework selects the affordance corresponding to the highest probability in the set *GA_r* (line 11). The optimal *region_label** is then projected to the object 3-D data using the extracted features-labels map *3D_region* (line 12). On this optimal grasping patch G^* we calculate a grasping configuration ${}^{W}ee\_pose^*$ in world coordinates (line 13) to be sent to the robot (line 14).

Algorithm 1: end-to-end execution

```
1  Input:
2  CP: camera perception.
3  affordance: affordance choice.
4  imageToLabel: DCNN learned model.
5  regionsFromCloud: SVM learned model.
6  semanticRelation: KB learned model.
7  begin
8      2D_labels ← imageToLabels(CP.2D_image)
9      3D_region ← regionsFromCloud(CP.3D_image)
10     GA_r ← semanticRelation(2D_labels)
11     region_label* ← selectGrasp(GA_r, affordance)
12     G* ← 3D_region(region_label*)
13     Wee_pose* ← extractGraspPose(G*)
14     sendToRobot(Wee_pose*)
```

IV. Evaluation on Grasp Affordance Dataset

We evaluate our methodology on a PR2 robotic platform, in both simulated and real-world scenarios. The 2-D data is perceived with the robot's left 2-D camera and the 3-D information with a kinect mounted on its head. We use the end-to-end execution framework as presented in Algorithm 1.

A. Data Collection

This work pursues a multi-target prediction of grasp affordances on an object for which the training data needs to be diverse, accurate and consistent. However, when learning object affordances for robotic grasps, one of the greatest obstacles is the lack of datasets that offer a multi-grasp affordance relation. Therefore, we build a new dataset with highly correlated information that encourages the creation of more robust robotic grasping methods. We use this collected dataset to ensure the reliability of the obtained grasp affordance regions from the KB. We choose 30 different objects that are commonly found

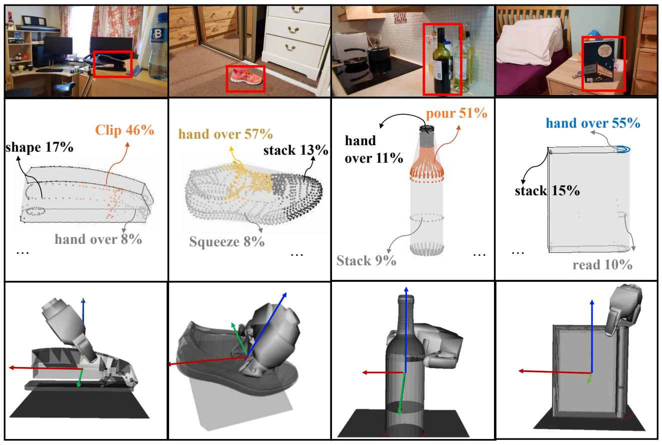


Fig. 4: The first row shows an example of an image containing an object and location. The second row is the segmented and processed object 3-D point cloud with the three highest grasp affordance predictions. The third row is one of the proposed grasp configurations on the obtained grasp affordance region.

in online datasets used by grasping [6], [7], object recognition [23] and object-location association methodologies [24], [25], resulting in a prior for grasping, object category, and possible location labels. Specifically, the pre-defined grasping regions are taken from [13]. Nonetheless, these datasets do not relate to each other, and the affordance relation is still unsolved. Consequently, we design a detailed questionnaire containing these different 30 objects with their corresponding label priors alongside descriptions of visual and categorical attributes, as suggested in [16], as well as possible affordances and indoor locations. This questionnaire was presented to a total number of 1,269 subjects. The collected data led to the creation of a total of 3,280 possible worlds that were used in training and testing. These possible worlds were composed of three visual attributes, each with at least four possible values, eight possible object categories (i.e., the 30 objects organised as food, electronics and others), seven possible indoor locations (such as kitchen, office and others) and fourteen possible affordable actions closely related to at least three possible grasping regions.

B. Baselines for Grasp Affordances Evaluation

As explained in Algorithm 1, our method is able to evaluate different sets of grasps and select the one that will maximise the success of an affordance. For the following set of experiments, we query the KB not only for the grasp region but also for the affordance prediction, as the query shown in Table II, to perform a complete evaluation. First, we extract the attributes that build the KB queries for inference. We use 30% of the objects for testing. These objects are semantically similar to the 70% of objects used for training. To predict their affordances, we take ten images per object in different environments. Given a 2-D image we extract the scene location, visual and categorical attributes describing the object using a deep CNN. Second, we collect the scores

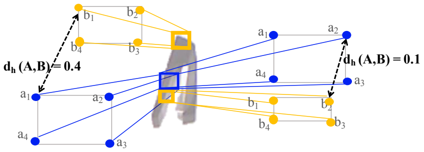


Fig. 5: Hausdorff distance example between two grasp regions on a stapler image from [13]. The grasp affordance rectangle is shown in blue while the ground truth labels are in yellow. The less similar the sets the higher the d_h value.

from the binary vector (i.e. non-zero entries). Finally, these attributes are translated into relational worlds using FOL and input in the KB where we query the grasp affordance regions. Fig. 4 shows examples of the images taken for the grasp affordance prediction. The first row contains the 2-D image of the environment with the object. The second row depicts the relevant 3-D data on the extracted shape context with the top three highest grasp affordance probabilities, and the third row is the resulting grasp configuration on G^*.

C. Metrics for Grasp Affordances Regions

To establish the reliability of our grasp affordance regions, we evaluate the obtained patches against the adopted grasping labels from [13]. In [13], the label is a rectangle that covers the grasping area with the corresponding 2-D and 3-D mappings. We choose four instances per object belonging to the 30% of the testing data from [13] and simulated to be in an office environment. On the 2-D data, we project the ground truth and our grasp affordance regions are enclosed in rectangles. Methods that have used the same database to extract the grasping labels [6], [7] have based their evaluation on measuring the Euclidean distance between rectangle centroids. As we generalise a segment of an object as a grasping patch, these metrics might overestimate the performance of the algorithm (if one ground truth rectangle is inside a large obtained grasp affordance rectangle), or underestimate it (if the obtained grasp affordance region does not intersect but is close). Thus, we use the Hausdorff distance as a metric of choice to establish the similarity between the two projected rectangles set. The Hausdorff distance is the maximum of all distances from a point in one set A to the closest point in another set B, the smaller the value, the more similar the sets are, i.e.:

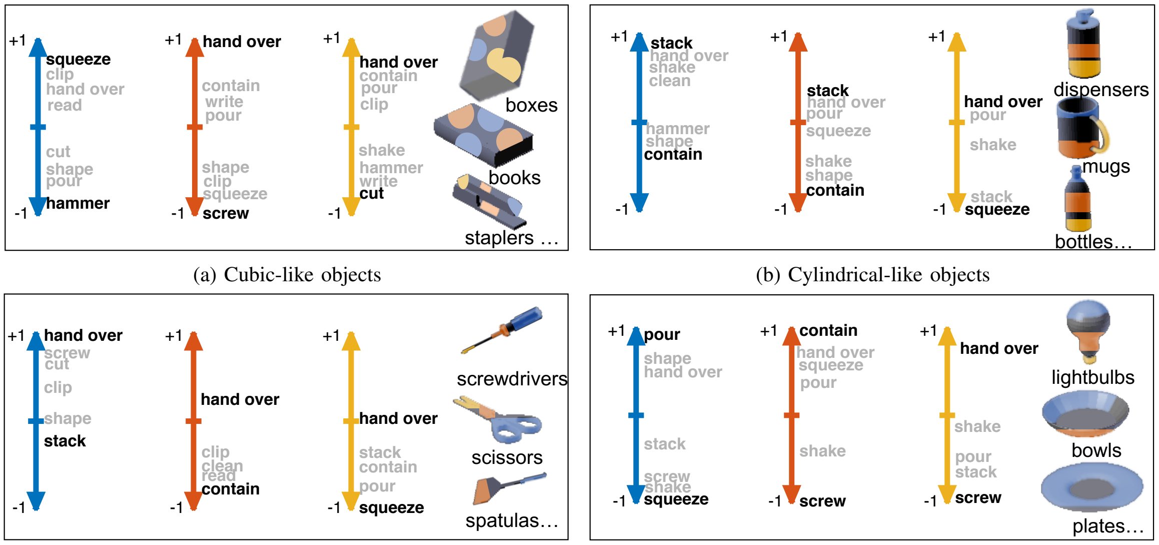
$$d_h(A,B) = \max_{a \in A}(\min_{b \in B} d(a,b)), \tag{3}$$

where a and b are points in sets A and B respectively, and $d(a,b)$ is the Euclidean distance between a and b. Fig. 5 shows how the Hausdorff distance accurately measures the similarity between rectangle patches, although they do not intersect.

V. Experiments and Discussion

The goal of this work is to reason about the feasible grasps for an object given an affordance. Thus it is important to (i) test the accuracy of grasp affordance prediction, and (ii) test the reliability of the grasps in changing scenarios[2].

[2] More experiments can be found in https://youtu.be/aaA3NA-S5KY.



(a) Cubic-like objects

(b) Cylindrical-like objects

(c) Irregular-like objects

(d) Spherical-like objects

Fig. 6: Visualisation of the normalised grasp affordance likelihood learned in Section III-B subject to objects' shape context ((a)-(d)) and grasp regions (colour coded arrows with corresponding regions on the objects). The more positive the weight, the more likely that region offers a feasible grasp for the indicated affordance. We include the likelihoods close to the extremes.

A. Zero-Shot Grasp Affordance

Our first evaluation tests the performance of the KB on unseen scenarios. In contrast to [15], our feature extraction approach is a deep CNN architecture that extracts the objects attributes. We use eight objects semantically similar to the ones used in training, as explained in Section IV-B. We evaluate the performance of our proposed KB against three state-of-the-art reasoning methodologies: (i) a KB built with a series of L1-regularised logistic classifiers [16], (ii) our KB based on decision trees [18] and, (iii) the KB based on MLN proposed in [15] with their SVM ranking function for feature extraction. Table III shows the mean area under the curve (AUC) under the Receiver Operating Characteristics (ROC) curve over all the possible grasp affordances. The results show that our method has the best performance of all methods tested since we train our KB over a combination of highly correlated object predicates and relevant constant terms. Additionally, we note the improved performance of KBs that use MLN over the ones trained with a battery of classifiers. By using MLN, the attributes build relationships regarding object grasp affordances that the classifiers fail to incorporate.

B. Grasp Affordances Relation

Our primary contribution is to associate a set of grasp affordances with an object. Fig. 6 portrays possible grasp affordances for objects in different shape contexts across different indoor scenes. The three arrows represent the three grasping regions across different objects with more affordance possibilities. The regions are colour coded on the corresponding area of the different objects. The affordances are sorted by the normalised weights between -1 to +1 per grasping region, where the higher the weight, the more likely that affordance is to be successful when grasping the object using the colour coded grasping region. We group the objects by shape context for a clearer grasp affordance representation. Out of the 14 possible affordances, we extracted those with higher and lower relational weight. Among the different grasp affordances, one of the most probable ones across shape contexts is object *hand over*. Specifically, this is the case for objects that are used as tools and are recognised with an irregular shape (Fig. 6c). Because the KB has learned from data collected from humans, it reflects the likelihood of that grasp affordance region to be more or less "acceptable" than others for a particular action. Also in Fig. 6c, the likelihood of success at *handing over* the objects from the grasp region indicated in blue (i.e., object handles) is higher than the other two. The same case is shown in Fig. 6a and Fig. 6b for *stack*. Moreover, Fig. 7 illustrates some of the grasp affordance feature patches of the *hand over*, *stack* and *pour* affordances learned with our method. These patches (specifically red areas) correspond to graspable regions of objects such as handles or other raised regions.

C. Grasp Affordance Selection Reliability

We demonstrate the reliability of our grasp affordance hypothesis by (i) evaluating the accuracy of the affordance detection, and (ii) checking if the obtained grasping regions correspond to stable grasps using Hausdorff distance. We

Approach	**Performances per attribute ($\overline{\text{AUC}}$)**		
	visual	**categorical**	**all+location**
L1-LR [16]	0.70	0.74	0.77
SVM KB [15]	0.73	0.77	0.82
our previous KB [18]	0.72	0.75	0.79
our KB	0.75	0.79	**0.84**

TABLE III: Performance of Zero-Shot grasp-affordance prediction for different attributes and their combination.

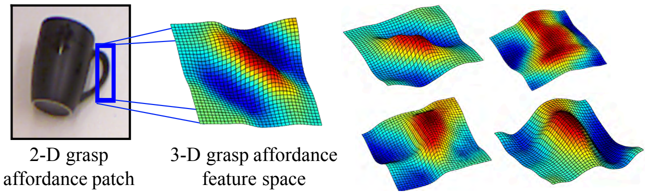

2-D grasp affordance patch
3-D grasp affordance feature space

Fig. 7: Examples of learned grasp affordance patch features of affordances such as *hand over*, *stack* and *pour*.

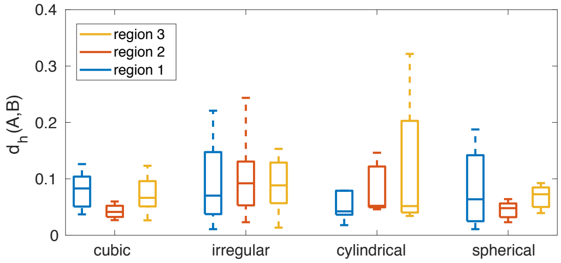


Fig. 8: Hausdorff distance to measure the similarity between ground truth regions and our grasp affordances.

train two state-of-the-art methods, using deep learning CNN methodologies, with our dataset and evaluate separately (i) and (ii) given that the current literature does not treat the grasp affordance as a unified task.

1) Affordance Detection: We compare our affordance detection with [5] (**Detection (%)** in Table IV). We use a total of 64 images of 16 object classes in different environments, among which the eight used for zero-shot prediction are included. Our method shows a higher performance, by using MLN, since we build a series of relationships around an object (*worlds*) that traditional machine learning methods fail to connect. Namely, [5] does not consider the task given a context. Instead, it learns object part labels and categories to assign an affordance. Fig. 9a shows an example of a knife grasp affordance detection. The knife affords equally two tasks when using [5], while the affordances are correct (*cut and hand over*), when grasping, determining one task is essential.

2) Similarity between patches: First, we check the similarity between our regions and the original ones from [13]. We use the subset of 32 images (Section IV-C) to project the ground truth and our obtained areas. Fig. 8 shows the Hausdorff distance, d_h between the obtained hypotheses, set A, and the ground truth labels, set B, grouped by shape context and grasping regions. The d_h mean per grasp affordance region are below 0.1 for all the objects. Specifically, a low d_h is obtained for rectangles that are nearby ($d_h \leq 0.1$) while larger values might be obtained by far apart sets ($d_h \geq 0.4$) (Section IV-C). Second, we compare our method with [6] which finds multiple reliable grasping regions on the data in [13]. We use the subset of 64 images and compare the Hausdorff distance. Table IV shows that both methods achieve considerably small and similar d_h, thus learning stable grasping patches. Examples of grasping patches obtained with both methods, ours and [6], are shown in Fig. 9a.

D. Grasp Affordance on a Robotic Platform

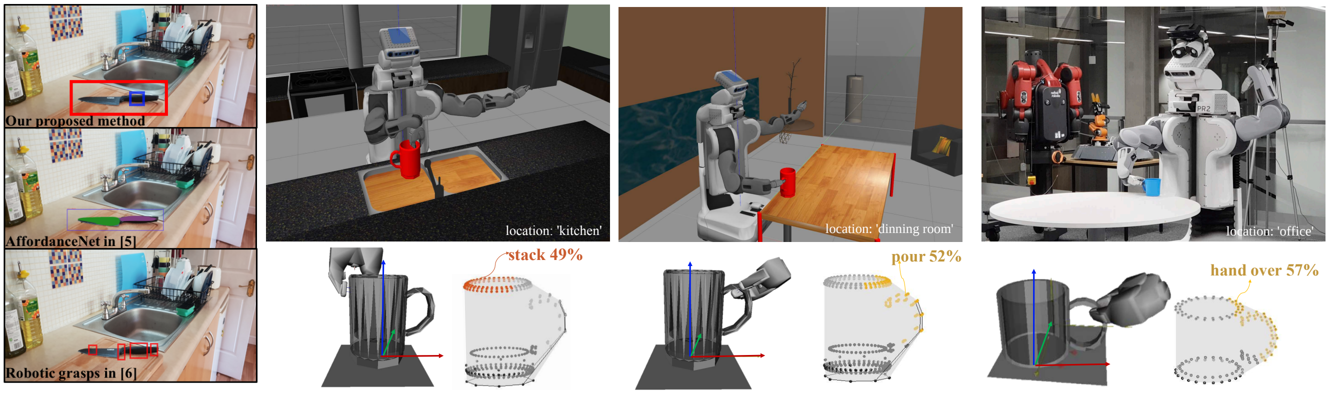
To explore the generalisability and effect of the robot on the success rate of our method, we ran an extensive number of experiments on both a simulated and real PR2 robotic platform. We tested our method in three different indoor scenarios: a kitchen and dining room setting in simulation and a real office environment. For this experiment, we use the previously selected 16 different object classes, among which we assess robustness by variating instances, and try the affordance detection and grasping task 25 times per object class for a total of 400 evaluations. The performance of the grasp affordance task is detailed in Table IV (**On the robot (%)**). We average the grasp affordance prediction with the actual grasping action success on the correct detection cases (**avg** in Table IV). The grasp is considered successful if: (i) the gripper approaches the grasp affordance region of the object below a Hausdorff distance threshold ($d_h < 0.2$), and (ii) if the object is successfully grasped. The lowest performance is obtained by objects with irregular shape context where the most significant setback is the success of the grasping action given that, in general, the set of objects were too small for the PR2 gripper; thus they frequently slip. Fig. 9b-9d illustrate the experimental set-up for the three different scenarios, focusing on the grasp affordances of a mug. Interestingly, the affordances detected on the object in the three locations are different. In two out of the three scenarios, dining room and office, the grasping region coincides even though the affordances are different. The method relates the location with object semantics to decide on a grasp affordance that will potentially ensure the accomplishment of an action. Given that the KB learned from our collected data, it reflects what is socially "acceptable" in an environment. Although the grasping region for the mug is the same in the dining room and the office, in the office it is more likely *hand over* the mug than to *pour* liquids from it.

VI. Conclusions and Future Work

We presented a new method for reasoning about the different grasp affordances of an object. In contrast to state-of-the-art techniques, instead of hand-defining the grasp affordance labels on the objects, we collected data from 1,269 different participants to obtain their input on the relation of object attributes, locations and grasp affordance labels. Using this collected data, our approach not only learns grasp affordances but also learns to characterise socially acceptable grasp behaviours on different objects in various scenarios. The information included in this dataset opens doors in the research community towards more robust and heterogeneous robotic

Objects shape	Detection (%)		d_h		On the robot (%)	
	[5]	ours	[6]	ours	detection/grasp	avg
Cubic	82.7	88.5	0.05	0.01	90.3 / 100	95.2
Cylindrical	79.6	87.4	0.01	0.03	87.1 / 96.1	91.6
Irregular	77.8	88.6	0.09	0.12	87.9 / 77.3	82.6
Spherical	83.2	90.5	0.02	0.03	91.6 / 100	95.8

TABLE IV: Comparison with state-of-the-art methods and grasp affordance performance of the robotic platform.



(a) Method comparison (b) Kitchen (c) Dinning room (d) Office

Fig. 9: (a) Comparison of our method with state-of-the-art alternatives [5], [6], and (b)-(d) our method running in PR2 to grasp an object in different simulated and real-world scenarios while checking the variations on the detected grasp affordances.

grasping methods. The proposed method also outperforms alternative grasp affordance recognition techniques. We attribute this performance to our structures for grounding and relating data. Our method is able to (i) reason about the most probable grasp affordance, among a set, by inferring the contextual semantics relation, and (ii) map the optimal grasp affordance to the 3-D data of the object and proceed with the grasp using a robotic manipulator. Moreover, this work encourages interesting future studies such as the prediction of action probabilities to be executed by associating objects in the scene, the evaluation of the generalisability of our method with different manipulators, and the assessment of end-state comfort-effect for grasping in human-robot collaboration tasks.

References

[1] J. Gibson, "The theory of affordances," in *Perceiving, Acting, and Knowing: Toward and Ecological Psychology* (R. Shaw and J. Bransford, eds.), pp. 62–82, Hillsdale, NJ: Erlbaum, 1977.

[2] L. Montesano and M. Lopes, "Learning grasping affordances from local visual descriptors," in *Development and Learning, 2009. ICDL 2009. IEEE 8th International Conference on*, pp. 1–6, IEEE, 2009.

[3] J. Bonaiuto and M. A. Arbib, "Learning to grasp and extract affordances: the Integrated Learning of Grasps and Affordances (ILGA) model," *Biological cybernetics*, vol. 109, no. 6, pp. 639–669, 2015.

[4] S. Hart, P. Dinh, and K. A. Hambuchen, "The affordance template ros package for robot task programming," *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6227–6234, 2015.

[5] T.-T. Do, A. Nguyen, and I. Reid, "Affordancenet: An end-to-end deep learning approach for object affordance detection," in *International Conference on Robotics and Automation (ICRA)*, 2018.

[6] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 705–724, 2015.

[7] J. Bohg and D. Kragic, "Learning grasping points with shape context," *Robotics and Autonomous Systems*, vol. 58, no. 4, pp. 362–377, 2010.

[8] N. Krüger, C. Geib, J. Piater, R. Petrick, M. Steedman, F. Wörgötter, A. Ude, T. Asfour, D. Kraft, D. Omrčen, *et al.*, "Object–action complexes: Grounded abstractions of sensory–motor processes," *Robotics and Autonomous Systems*, vol. 59, no. 10, pp. 740–757, 2011.

[9] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis, "Object-based affordances detection with convolutional neural networks and dense conditional random fields," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017.

[10] D. Song, K. Huebner, V. Kyrki, and D. Kragic, "Learning task constraints for robot grasping using graphical models," in *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pp. 1579–1585, IEEE, 2010.

[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*, pp. 4373–4378, IEEE, 2012.

[12] R. Detry, E. Baseski, M. Popovic, Y. Touati, N. Kruger, O. Kroemer, J. Peters, and J. Piater, "Learning object-specific grasp affordance densities," in *Development and Learning, 2009. ICDL 2009. IEEE 8th International Conference on*, pp. 1–7, IEEE, 2009.

[13] Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from rgbd images: Learning using a new rectangle representation," in *2011 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3304–3311, IEEE, 2011.

[14] M. Richardson and P. Domingos, "Markov logic networks," *Machine learning*, vol. 62, no. 1-2, pp. 107–136, 2006.

[15] Y. Zhu, A. Fathi, and L. Fei-Fei, "Reasoning about object affordances in a knowledge base representation," in *European conference on computer vision*, pp. 408–424, Springer, 2014.

[16] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 1778–1785, IEEE, 2009.

[17] L. Fadiga, L. Fogassi, V. Gallese, and G. Rizzolatti, "Visuomotor neurons: Ambiguity of the discharge or motorperception?," *International journal of psychophysiology*, vol. 35, no. 2-3, pp. 165–177, 2000.

[18] P. Ardón, È. Pairet, S. Ramamoorthy, and K. S. Lohan, "Towards robust grasps: Using the environment semantics for robotic object affordances," in *Proceedings on AAAI FS on Reasoning and Learning in Real-World Systems for Long-Term Autonomy*, pp. 5–12, AAAI Press, 2018.

[19] D. D. Johnson, "Modified broydens method for accelerating convergence in self-consistent calculations," *Physical Review B*, vol. 38, no. 18, p. 12807, 1988.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[21] C.-J. Kim, C. R. Nelson, *et al.*, "State-space models with regime switching: classical and gibbs-sampling approaches with applications," *MIT Press Books*, vol. 1, 1999.

[22] S. Katz and A. Tal, *Hierarchical mesh decomposition using fuzzy clustering and cuts*, vol. 22. ACM, 2003.

[23] W. Wohlkinger, A. Aldoma, R. B. Rusu, and M. Vincze, "3dnet: Large-scale object class recognition from cad models," in *2012 IEEE International Conference on Robotics and Automation*, pp. 5384–5391, IEEE, 2012.

[24] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 413–420, IEEE, 2009.

[25] E. Kolve, R. Mottaghi, D. Gordon, Y. Zhu, A. Gupta, and A. Farhadi, "Ai2-thor: An interactive 3d environment for visual ai," *arXiv preprint arXiv:1712.05474*, 2017.