Reasoning on Grasp-Action Affordances

Paola Ardón[0000−0002−3026−0706], Èric Pairet[0000−0002−3363−0426], Ron Petrick[0000−0002−3386−9568], Subramanian Ramamoorthy[0000−0002−6300−5103], and Katrin Lohan[0000−0001−9843−316X]

Edinburgh Centre for Robotics. Edinburgh, UK.
paola.ardon@ed.ac.uk

Abstract. Artificial intelligence is essential to succeed in challenging activities that involve dynamic environments, such as object manipulation tasks in indoor scenes. Most of the state-of-the-art literature explores robotic grasping methods by focusing exclusively on attributes of the target object. When it comes to human perceptual learning approaches, these physical qualities are not only inferred from the object, but also from the characteristics of the surroundings. This work proposes a method that includes environmental context to reason on an object affordance to then deduce its grasping regions. This affordance is reasoned using a ranked association of visual semantic attributes harvested in a knowledge base graph representation. The framework is assessed using standard learning evaluation metrics and the zero-shot affordance prediction scenario. The resulting grasping areas are compared with unseen labelled data to asses their accuracy matching percentage. The outcome of this evaluation suggest the autonomy capabilities of the proposed method for object interaction applications in indoor environments.

1 Introduction

One of the most significant challenges in artificial intelligence is to achieve a system that simulates human-like behaviour. Let us consider a robot in a simple task such as finding, collecting and delivering an object in home environments. Given the complexity of home settings, it is hard to provide a robot with every possible representation of the objects contained in a house. It is even harder to feed the robot with all the possible uses of those objects. Instead of learning all possible scenarios, suppose that a reasoning technique allows the system to deduce an object affordance. As a result, offering the opportunity to achieve autonomous capabilities. The term affordance refers to everything that defines the interaction with an object, from the way to grasp it to its inherited ability to perform different tasks [10]. Thus, affordance defines all possible actions depending on the target objects' physical capabilities. For instance, a glass cup looks as if it can be handed over, contain liquids, or pour liquids from it. The characteristics that define the glass cup as a container or graspable object constitute its affordance. According to different theories of human perception, the psychology of perceptual learning compounds the different qualities in the environment rather

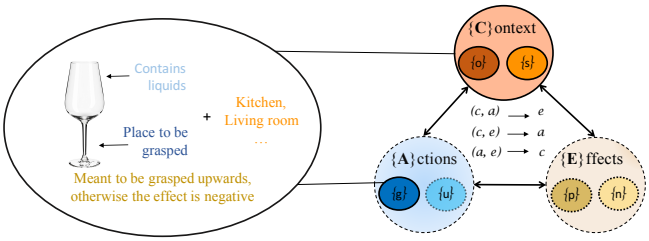


Fig. 1. Affordance map model to create a correlation between the objects properties and their environment to improve on grasp-action affordance.

than acquiring associated responses to every object [9, 3]. Thus, humans are efficient at deducing affordance for objects with different appearances and similar abilities, e.g. glasses: wine, tumbler, martini, and discern among those with similar features but different purposes, e.g. bowling pin vs water bottle. Nonetheless, in robotics, the most common approach to learn affordances is from labels [16, 14, 4]. This technique limits the number of learned objects, grasping areas and affordance groups. Moreover, the robot is unable to interact with novel objects. Further, by learning the limited set of responses, it is not possible to deduce the key features that define the objects affordance.

Using the same analogy as the theories of human perception this paper hypothesises that using the semantic features of the object and its surroundings not only improves the affordance grasping action towards the object but it also allows a reasoning process that, in the long term, offers autonomy capabilities, a solution not yet seen in the current literature. This work summarises an architecture that addresses the previously described challenges. The focus is on affordance reasoning for calculating grasping areas, using a combination of the object and its environment features. Figure 1 shows the foundations of this proposal, which is an extended version of the affordance map presented in [16]. The proposed methodology works with the concept that an affordance relates attributes of an object and the environment to an interactive activity by an agent who has some ability, which relates back to the object causing some affordance. In other words, the attributes of the object and the environment reside in the context of the affordance, the abilities of the agent and the object in the affordance actions and the outcome of this interactive activity in the effects. This work focuses on the integration of the semantic features of the previously mentioned environment in order to obtain a good grasp affordance action, from now on referred to as grasp-action, of the object. The presented framework can reason on the object grasping areas that are strongly related to the affordance group. The reasoning process is based on a Knowledge Base (KB) graph representation. This KB is built using semantic attributes of the object and the environment. For every object explored by the framework, the KB uses weights to relate a subset of attributes. This association then leads to an affordance category which is highly correlated with a grasp-action area. The designed framework is assessed not only using standard learning evaluation metrics, but it is also tested on the zero-shot affordance prediction scenario. Moreover, the resulting grasping areas

are compared with unseen labelled data to asses their accuracy matching percentage. The results demonstrate the suitability of the method for grasp-action affordance applications, offering a generalised object interaction alternative with autonomy capabilities.

2 Related Work

Many methods extract viable grasping points on objects, independently on their affordance [14, 1]. Others focus explicitly on the task of grasp-action affordance from visual features and model parameters that are learned through reinforcement learning using biologically inspired methods [23, 4]. [23], interestingly embraces psychology theories for human development such as the ones presented in [9] to learn from exploratory behaviours the invariants to obtain the best grasps. Contrary, [2, 15] focus on the ability-action affordance of the objects. In their work, they use statistical relational learning to learn the ability affordance of different objects, which shows to cope with uncertainty. Other works go beyond the visual representation of the object and combine visual as well as textual descriptors to build a KB [25, 22]. This KB is composed of actions learned through reinforcement learning techniques with the purpose of interacting with the object. [8, 12] work on the actions and objects relations in a single interface representation to capture the needs of planning and robot control. Another extension is [5], they use these action complexes to extract the best grasping points of the objects. In literature, it is extensive the use of learning techniques such as deep Convolutional Neural Networks (CNN) to build an affordance model based on the visual objects features, resulting in a plausible generalised method given the robustness of their data [17, 6]. Unlike these works, this paper presents a methodology that combines attributes of the object and the environment to provide a denser context for object affordance interaction. Thus, allowing it to generalise the grasp-action affordance on similar objects.

3 Proposed Solution

In this paper, a grasp-action area of the object is the result of the relation between the object and its surrounding environment. Figure 1 shows a summary of the proposed affordance model. Let us consider a glass cup in an affordance map relationship. Additional to its inherited affordance action qualities, i.e. contain liquids and being graspable, there are other elements that define its opportunity of interaction. For example the way in which it is being manipulated as well as the features that describe the glass cup itself and its surrounding environment. All these elements together define the affordance of the glass cup. This does not mean they are dependant of each other but rather codefining and coherent together. Bearing this example in mind, in Figure 1, the context $\mathbf{C} = \{c_1, c_2, ..., c_n\}$ is the set of semantic attributes of the glass cup and its environment (such as kitchen and living room), ({o}bject $\cup$ {s}urrounding) $\subseteq$ {**C**}ontext. The set of available actions, $\mathbf{A} = \{a_1, a_2, ..., a_n\}$, is understood as a twofold: (i) the way in which

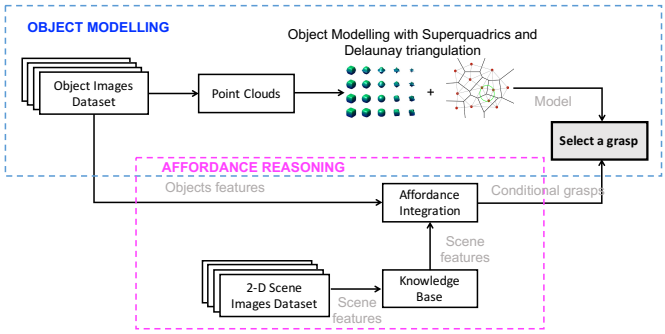


Fig. 2. Proposed framework for grasping affordance reasoning.

the glass cup can be approached, its suitable grasp-action areas, and (ii) the usages that the glass cup can achieve, its ability-action such as containing liquids, ({g}rasps ∪ {u}sage) ⊆ {**A**}ctions. The set of effects of performing those actions, $\mathbf{E} = \{e_1, e_2, ..., e_n\}$, is kept as a simple discretisation between positive or negative effects, such as holding the glass cup correctly in order not to spill the liquid, ({p}ositive ∪ {n}egative) ⊆ {**E**}ffects. The key attributes of the affordance reasoning to get those grasp-action areas are enclosed in the form of a KB. These methods are commonly used in artificial intelligence because of their advantages for harvesting data and accessing a more extensive array of queries regarding the essential features of a process, rather than just the result. KBs achieve this task by connecting a collection of attributes through a general set of rules. In this work, the attributes are the features that describe the object and the environment and are connected through a hierarchical set of decisions that result in the object affordance. This section first summarises the object modelling stage, to then reason on the object affordance that is highly correlated with the resulting grasp-action areas as schematised in Figure 2.

A KB is visualised as a graph representation, as illustrated in Figure 3 where the entities (nodes) are connected by general rules (edges). In this setup, the entities are the target object, the attributes of the object and its surrounding, and the resulting affordance groups. The general rules are the attribute to attribute relation that results from a classification process. The relation between attributes are weighted accordingly, where the higher the weight, the higher the correlation between the two entities. In order to describe objects by their attributes the best practice is to divide their features into base, semantic and discriminative [7]. In this work, the base features, such as edges and colours, are extracted using CNN. The semantic features are visual characteristics of the object as defined in Table 1. From now on, these features will be referred to as visual semantic features. They are the result of a deep CNN and are divided as (i) shape attributes, these are the set of visual attributes that describe the objects geometrical appearance; (ii) texture attributes, are categories based on visual characteristics of the objects materials; and (iii) environment attributes, which are the scenarios in which the objects are more likely to be found in. This attribute is added with

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, living room, office

Table 1. Used attributes and entities of the KB graph.

the purpose of facilitating the object affordance reasoning. The implemented KB considers different 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 are those that offer a comprehensive understanding of the semantic features. They are the result of a predictive decision tree model that uses deep CNN as nodes. The KB is composed of four different Deep Neural Networks that, through the pre-trained CNN, resnet50 [11], extract features from the perceived images. These four different deep learned CNN correspond to the four different visual semantic attributes, as described in Table 1, which result in the deduced set of entities in a graph that defines a grasp-action affordance.

3.1 Knowledge Base Predictive Model

In this paper, the KB constitutes a data library that builds a predictive model connected through a hierarchical set of decisions, such as the edges on Figure 3, from now on referred as weights. These decisions are the result of a classification task of the object semantic features, represented as the nodes in Figure 3. From each of the attributes, $\forall a \in \boldsymbol{A} : \boldsymbol{A} \in [1, ..., K]$, where K is the total number of visual semantic features as described in Table 1, a set of weights represented as a vector $\boldsymbol{\Psi_{a_k}} = [\psi_1, \psi_2, ..., \psi_n]$ is extracted, where n is the total number of entities in that attribute. These $\boldsymbol{\Psi_{a_k}}$ are hierarchically connected with the next attribute a_{k+1}. Then $\boldsymbol{\Psi_{a_k}}$ offers a way to rank on the next best entity candidate. The

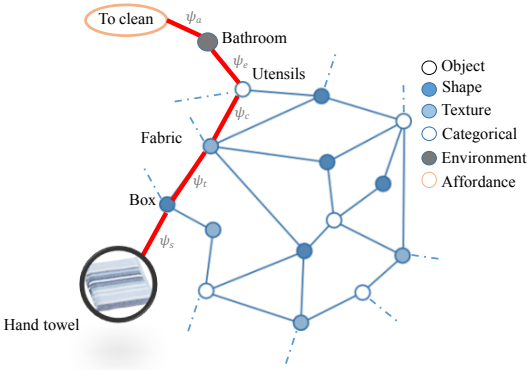


Fig. 3. Example of a cleaning object and the extracted attributes used to build the KB graph. The higher weights $\boldsymbol{\Psi}$ (red) create the reasoning to an affordance group.

higher the ψ_n, the higher the probability that the connected two entities among attributes result in a better affordance reasoning. 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:

$$\widehat{P}(a|x) = \frac{P(x|a)P(a)}{P(x)}, \tag{1}$$

where x is an image belonging 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 3 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 path (shown in red) to the *to clean* grasping affordance. The collected information from each of the deep CNN is then used to learn a decision tree as a predictive model: $(\boldsymbol{y}, Z) = (y_1, y_2, y_3, ..., y_n, Z)$, where Z is the affordance group that the system is trying to reason, and the vector $\boldsymbol{y}$ is the set of features $\{y_1, y_2, y_3, ..., y_n\}$ used for the reasoning task. Thus, the model learns the ranking that reasons on the affordance grasping task $R(x) = \boldsymbol{\Psi}_{\boldsymbol{A}}^{\intercal}\boldsymbol{y}(x)$ where $\boldsymbol{\Psi}_{\boldsymbol{A}}$ is the transpose of the model parameters from all the attributes and $\boldsymbol{y}(x)$ is the set of visual features of a given image x.

3.2 Calculating the Grasping Points

Once the affordance is deduced, the system selects from the set of grasping points obtained in the object reconstruction stage and limits the grasps depending on the affordance reasoning obtained from the KB. In order to impose such constraints, the space of the previously obtained grasping points is discretised in the third dimension, z, so that the following decision on the grasping area can be made: (i) The grasping region should lie on those points located in the central subspaces of the discretised space for objects that are meant to contain edibles. (ii) 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, given that the affordance action-effect is not critical.

4 Evaluation

This work's goal is to achieve a system able to reason on the object grasp-action affordance, thus offering autonomy capabilities. As a result, it is of interest to evaluate the KB on (i) its attribute accuracy classification, and (ii) its reasoning efficiency with similar objects.

4.1 System Setup

The setting up of the system consists on collecting the required data for the training and the assessment of the method. This collection is built using two different datasets that are manually organised into entities of the attributes described in Table 1. After passing through the predictive model in the KB, every object in the library is expected to fall into: *to eat, to contain, to hand*

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

Table 2. Each of the attributes classification accuracies.

over, to brush, to squeeze, to clean or *to wear*. The first set of images is from the Washington-RGB dataset, which contains 300 objects providing the point clouds and the two-dimensional (2-D) images for each one of the instances [13]. The second dataset is the MIT indoor scene recognition that contains 15,620 different 2-D images of 67 different indoor scenes from which this work uses seven of those classes [20]. By unifying these two datasets, the objects are correlated to the environment in which they are more likely to be located. 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 to define good object affordances features. In order to represent the obtained grasping area of the objects, an ellipse with the iCub humanoid robot end-effector dimensions is simulated. The orientation of such ellipse is out of the scope of this work and the focus remains on the position of the grasping area.

4.2 Reasoning on the Affordance

A summary of the accuracies per deep CNN in the KB is presented in Table 2. As a reminder to the reader, the aim of the proposed methodology is not to improve the performance of the individual classifiers. Nonetheless, the illustrated accuracies match the state-of-the-art results shown in [11, 13]. To evaluate the overall performance of the KB, the accuracies before and after adding the environment features were collected. Figure 4 shows the data for both cases. A lower accuracy is obtained in the case where the environment features are not included, as illustrated in Figure 4(a) and Figure 4(b). Furthermore, Figure 4(a) not including the environment shows a slightly higher spread among different affordance classes. This misclassification is the case for affordances which objects have a

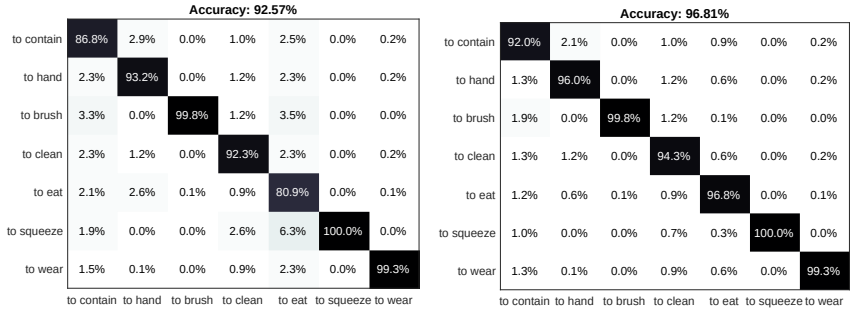


Fig. 4. Affordance category classification performance: (a) before adding environment features, showing an average diagonal accuracy of 92.57%; (b) after including the environment, showing a diagonal average accuracy of 96.81%.

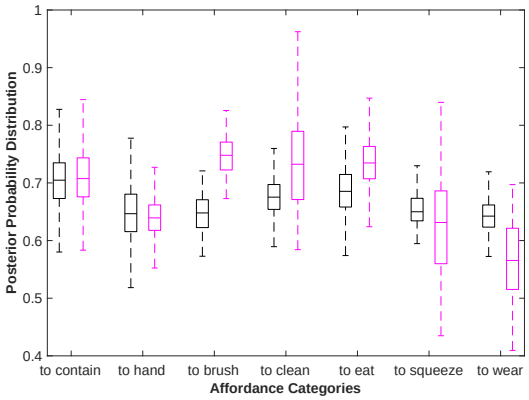
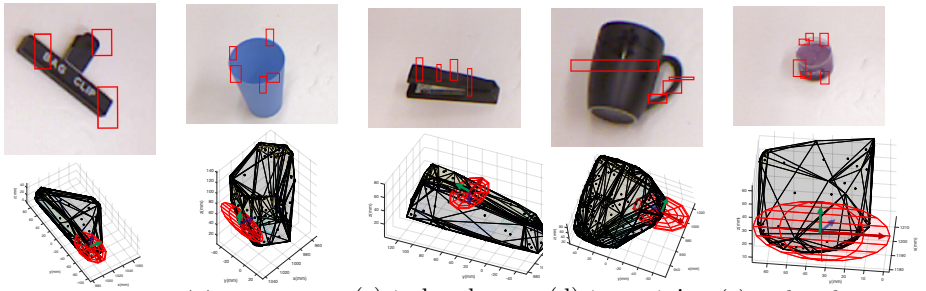


Fig. 5. Distributional posterior probabilities per class of the KB before (shown in black) and after (shown in magenta) the environment features.

general semantic categorical attribute such as "miscellaneous" or "container". Thus, a percentage of objects are misclassified among the *to contain, to brush, to eat,* and *to squeeze* categories. Regarding grasping, this miscue represents a significant adverse effect, especially for objects which real affordance is *to contain*, and its misclassification 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 Figure 4(b), especially in categories such as *to contain, to hand over* and *to eat*. The posterior probability distribution of the affordances categories is also evaluated. Figure 5 shows that while there is a decrement in the distribution for some categories such as *to hand*, there is an increment for others such as *to clean*. This change in the distribution is accredited to the variation in environments where these objects are found.

4.3 Zero-shot Affordance



(a) to hand over (b) to contain (c) to hand over (d) to contain (e) to hand over

Fig. 6. Zero-shot affordance prediction on semantically similar objects. The original images contain the labels (rectangles) for the preferred grasping regions from [14, 24].

Considering the changing nature of indoor scenes, it is useful to measure the method's affordance prediction on new objects. 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-shot affordance, in this case, refers to the affordance prediction of a familiar but previously unseen object. For this part of the experiments, a set of semantically similar objects has been chosen from a third dataset, Cornell [24]. This dataset is used to learn how to grasp objects in other works such as [14, 24]. These works exploit the fact that the Cornell dataset contains the three-dimensional (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 reasoning. In order to deduce the affordance of an unknown object, the same hierarchical procedure previously explained is followed. The set of weights $\boldsymbol{\Psi}_A$ has ranked a connection of attributes that results in an affordance, depending on the perceived semantics. Furthermore, this hierarchical connection has been learned in a predictive model to result in the grasping areas of the object. Figure 6 shows a sample of the familiar objects tested using the KB with their affordance group and deduced grasping area (shown with the red ellipse). Out of this subset, the most critical case is shown by the ones which affordance is to contain edibles, the cup and the mug in Figures 6(b) and 6(d), for which the grasping areas are correctly calculated.

5 Discussion

The proposed methodology is not only able to (i) reason on the object affordance of known and semantically similar objects, but also (ii) to extract a suitable grasp-action region of the target depending on the interpreted affordance. Given these features, this section discusses the performance of the KB on discerning the affordance of semantically similar objects, followed by a comparison of the obtained grasp-action regions with other methods' ground truth data.

5.1 Similar Shape, Different Affordance

One of the most significant arguments for building this framework is to help a robot generalise on object affordances. 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 objects in the library, it is of interest to evaluate the different affordance and grasping regions obtained for objects with similar shape but different affordance thus different preferred grasping regions. Figure 6(b) and Figure 6(e) are examples of two different everyday objects (a cup and a candle respectively) with considerably different affordance, where the located grasping regions differ according to the deduced affordance of the object.

5.2 Quality on the Calculated Grasping Area

Different works have been done in the field of affordance detection and grasping. However, they commonly 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 paper deduces the grasping region without any *a-priori* information about the grasping points. 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. There are works that use deep learning techniques to learn the grasping points of the objects mapped in the Cornell dataset images [14, 24, 21]. It is worth mentioning that these works do not account for affordance learning but for object classification. They simulate the end-effector with a rectangle, allowing it to account for its 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 as a success if it is within some distance from at least one of the ground truth rectangles. Contrary to this work, their labelled grasping areas are based on their end-effector control, and kinematic constraints and not on object affordance. 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 labelled rectangles of the Cornell dataset and the ellipses 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. 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 Figure 6. By measuring the Euclidean distance, 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, or entirely in a new area given that it has followed the constraints of the grasps depending on the object affordance, as it is the case of the cup in Figure 6(b).

6 Conclusions and Future Work

Contrary to the available methods, the framework presented in this paper is able to (i) reason on the affordance grasp-action of known and familiar objects without previously acknowledging the grasping areas, thus (ii) offering a reasoning process for object interaction with autonomy capabilities. The results of the evaluation performed on the framework support the hypothesis presented at the beginning of this work: that the grasp-action affordance does not depend solely on the object semantic features but on their combination with the features that describe the environment. The results show that without any *a-priori* awareness

on the grasping regions, the designed KB can reason on the object's affordance grasping points. The presented framework has room for improvement. The performance of the KB can be increased by adding more attributes to the base, as well as modifying the predictive model to classify more than one affordance 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 robot and the environment are considered out of the scope of the presented work. Nonetheless, [18, 19] offers a learning-based framework that comprises relative and absolute robotic skills for dual-arm manipulation suitable for dynamic environments, that together with a dense context representation of the scenario semantics offers a complete solution for an interactive object platform.

7 ACKNOWLEDGEMENTS

Thanks to the support of the EPSRC IAA 455791 along with ORCA Hub EPSRC (EP/R026173/1, 2017-2021) and consortium partners.

References

1. Ardón, P., Dragone, M., Erden, M.S.: Reaching and grasping behaviours by humanoid robots through visual servoing. In: Haptics: Science, Technology and Applications, Springer International Publishing AG. pp. 353–365. Springer Nature (2018)
2. Ardón, P., Pairet, È., Ramamoorthy, S., Lohan, K.S.: Towards robust grasps: Using the environment semantics for robotic object affordances. In: Proceedings of the AAAI Fall Symposium on Reasoning and Learning in Real-World Systems for Long-Term Autonomy. pp. 5–12. AAAI Press (2018)
3. de Beeck, H.P.O., Torfs, K., Wagemans, J.: Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway. Journal of Neuroscience **28**(40), 10111–10123 (2008)
4. Bonaiuto, J., Arbib, M.A.: Learning to grasp and extract affordances: the Integrated Learning of Grasps and Affordances (ILGA) model. Biological cybernetics **109**(6), 639–669 (2015)
5. Detry, R., Baseski, E., Popovic, M., Touati, Y., Kruger, N., Kroemer, O., Peters, J., Piater, J.: Learning object-specific grasp affordance densities. In: Development and Learning, 2009. ICDL 2009. IEEE 8th International Conference on. pp. 1–7. IEEE (2009)
6. Do, T.T., Nguyen, A., Reid, I.: Affordancenet: An end-to-end deep learning approach for object affordance detection. In: International Conference on Robotics and Automation (ICRA) (2018)
7. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. pp. 1778–1785. IEEE (2009)
8. Geib, C., Mourao, K., Petrick, R., Pugeault, N., Steedman, M., Krüeger, N., Wörgötter, F.: Object action complexes as an interface for planning and robot control. In: IEEE RAS International Conference on Humanoid Robots (2006)
9. Gibson, J.J.: The ecological approach to visual perception: classic edition. Psychology Press (2014)

10. Gibson, J.: The theory of affordances. In: Shaw, R., Bransford, J. (eds.) Perceiving, Acting, and Knowing: Toward and Ecological Psychology, pp. 62–82. Erlbaum, Hillsdale, NJ (1977)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Krüger, N., Geib, C., Piater, J., Petrick, R., Steedman, M., Wörgötter, F., Ude, A., Asfour, T., Kraft, D., Omrčen, D., et al.: Object–action complexes: Grounded abstractions of sensory–motor processes. Robotics and Autonomous Systems **59**(10), 740–757 (2011)
13. Lai, K., Bo, L., Ren, X., Fox, D.: Detection-based object labeling in 3d scenes. In: Robotics and Automation (ICRA), 2012 IEEE International Conference on. pp. 1330–1337. IEEE (2012)
14. Lenz, I., Lee, H., Saxena, A.: Deep learning for detecting robotic grasps. International Journal of Robotics Research **34**(4-5), 705–724 (2015)
15. Moldovan, B., Moreno, P., van Otterlo, M., Santos-Victor, J., De Raedt, L.: 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)
16. Montesano, L., Lopes, M., Bernardino, A., Santos-Victor, J.: Learning object affordances: From sensory–motor coordination to imitation. IEEE Trans. Robotics **24**, 15–26 (2008)
17. Nguyen, A., Kanoulas, D., Caldwell, D.G., Tsagarakis, N.G.: 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)
18. Pairet, È., Ardón, P., Broz, F., Mistry, M., Petillot, Y.: Learning and generalisation of primitives skills towards robust dual-arm manipulation. In: Proceedings of the AAAI Fall Symposium on Reasoning and Learning in Real-World Systems for Long-Term Autonomy. pp. 62–69. AAAI Press (2018)
19. Pairet, È., Ardón, P., Mistry, M., Petillot, Y.: Learning and composing primitive skills for dual-arm manipulation. In: Conference Towards Autonomous Robotic Systems. Springer (2019)
20. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. pp. 413–420. IEEE (2009)
21. Saxena, A., Driemeyer, J., Ng, A.Y.: Robotic grasping of novel objects using vision. I. J. Robotics Res. **27**, 157–173 (2008)
22. Sridharan, M.: Integrating knowledge representation, reasoning, and learning for human-robot interaction. In: AAAI Fall Symposium. Artificial Intelligence for Human-Robot Interaction. pp. 69–76. AAAI Press (2017)
23. Stoytchev, A.: Toward learning the binding affordances of objects: A behavior-grounded approach. In: Proceedings of AAAI symposium on developmental robotics. pp. 17–22 (2005)
24. Sung, J., Lenz, I., Saxena, A.: Deep multimodal embedding: Manipulating novel objects with point-clouds, language and trajectories. In: Robotics and Automation (ICRA), 2017 IEEE International Conference on. pp. 2794–2801. IEEE (2017)
25. Zhu, Y., Fathi, A., Fei-Fei, L.: Reasoning about object affordances in a knowledge base representation. In: European conference on computer vision. pp. 408–424. Springer (2014)