Getting On Top of Things: Towards Intelligent Robotic Object Stacking through Image-Schematic Reasoning

Kaviya Dhanabalachandran[1], Maria M. Hedblom[2] and Michael Beetz[1]

[1]Institute of Artificial Intelligence, University of Bremen, Germany
[2]Jönköping Artificial Intelligence Laboratory, Jönköping University, Sweden

Abstract

In this extended abstract, we present initial work on intelligent object stacking by household robots using a symbolic approach grounded in image schema research. Image schemas represent spatiotemporal relationships that capture objects' affordances and dispositions. Therefore, they offer the first step to ground semantic information in symbolic descriptions. We hypothesise that for a robot to successfully stack objects of different dispositions, these relationships can be used to more intelligently identify both task constraints and relevant event segments.

Keywords

image schemas, object stacking, cognitive robotics, commonsense reasoning, embodied cognition

1. Introduction and problem space

Designing systems for automated spatial reasoning is one of the most challenging yet most important components for intelligent system [1]. For cognitive robots, which can be defined as intelligent agents acting in space and time, an ability for spatial reasoning is a requirement for almost any activity. Yet, spatial reasoning in uncertain environments remains a complex area to solve within cognitive robotics. To contribute to this research agenda, this extended abstract introduces our research on identifying how the 'physical rules' of objects can be tied to their successful stacking.

Stacking is an important skill for many everyday activities. For instance, rearranging a book shelf, placing groceries into a pantry and carrying objects from one place to another on a tray. These are all activities that require the knowledge of the involved objects' features and the understanding of the underlying rules of how objects with such features can be stacked.

From repeated experiences with stacking objects, humans have extracted a lot of implicit (and explicit) knowledge of how certain objects can be treated. Such rules include the understanding of object properties. For instance, that a flat, sturdy object like a tray will offer support to other objects whereas a flat, flexible object like a sheet of paper will not offer the same support. Equally important is the understanding that objects will likely slide off when placed on top of slippery,

The Sixth Image Schema Day (ISD6), April 24–25, 2022, Jönköping University, Sweden
kaviya@uni-bremen.de (K. Dhanabalachandran); maria.hedblom@ju.se (M. M. Hedblom); beetz@uni-bremen.de (M. Beetz)
0000-0002-0419-5242 (K. Dhanabalachandran); 00070-0001-8308-8906 (M. M. Hedblom); 0000–0002–7888–7444 (M. Beetz)

convex surfaces (unless, of course, executed by an expert in object balancing). Other learnt rules stem from the distinction of the objects' affordances. For instance, classic, upward-facing containers (boxes, cups and bowls) behave differently from flat supporting objects in a stacking scenario. In the latter, objects are placed on top of the flat object's surface. In the former, the stacked objects are placed inside their hull (given that they are of an appropriate size).

For any formal system dealing with stacking situations, every feature of each involved object and every waypoint for executing certain trajectory needs to be represented in perfect detail for the agent to successfully accomplish the task. Providing this to a system is not only time consuming and cost inefficient for the engineer, it also reduces the autonomy and adaptability of such artificial agents. Thus, to implement a more intelligent capacity for object stacking, new directions need to be investigated.

One such direction that often is used in cognitive robotics is based on the theory of embodied cognition (e.g. [2, 3]). Within this theoretical framework, intelligent behaviour is thought to be based on conceptual patterns of meaning that are extracted from repeated experiences, by some, called image schemas [4, 5].

Image schemas represent the underlying rules for how object properties allow certain kinds of actions and are closely connected to object affordances[1] [6]. These patterns take the form of object dispositions such as offering SUPPORT[2] and CONTAINMENT. They also capture relational properties like relative object size (SCALE) and vertical orientation (VERTICALITY), as well as the object affordances related to movement (SOURCE_PATH_GOAL). Conceptual components like these become essential in understanding how certain objects can be used and how they behave in different situations. For instance, a cup is defined as its ability to contain liquids (CONTAINMENT), a tray is defined by the ability to support objects for movement (SUPPORT + SOURCE_PATH_GOAL) but both need to be stacked with the right orientation and the correct order. Thus, an event like stacking objects can be described as a combination of vertical pick-and-place tasks (VERTICALITY + SOURCE_PATH_GOAL) with the SUPPORT and CONTAINMENT constraints of any involved objects.

The mission of this research endeavour (see [7, 8] for some previous work) is to use the semantic information found in image-schematic patterns when designing robotic actions descriptions to generate meaningful event segments that can be reasoned about [9].

2. Related Work on Robots Stacking Objects

Object stacking is a well-known problem in robotics that still engage researchers. The problem involves understanding how geometric shape relates to vertical stability and BALANCE and how material properties relate to sturdiness and SUPPORT. The problem has been approached with methods ranging from task planning, reinforcement learning (RL), and to vision-based learning techniques that aims to learn the naïve physics of stacking.

For instance, in [10] the stacking problem is approached by iterative incorporation of motion constraints at the task level. [11] instead use a neural network to perform a stability classification.

[1] Affordances are actions that environments and objects allow. For instance, a box offers the affordance of CONTAINMENT.

[2] Following convention, image-schemas are written in capitalised upper letters.

Their system learns geometric affordances of introduced objects and arranges the objects based on a 'stackability score.' Based on this score, they are also able to BALANCE an unstable stack by placing an object to counterbalance the composition. Another noteworthy contribution, [12] introduces a robot that can build a tower of irregular stones by employing a gradient descent based pose search algorithm to find the best pose for each added stone.

In more domestic environments, [13] studies the problem of a robot organising shelves based on user preferences. Arguably, rearranging objects possess similar characteristics to stacking but instead of balancing on a vertical axis, complexity is to find appropriate compositions on the horizontal axis. [13] approached the problem by predicting pairwise object preference of a human user by using collaborative filtering model on crowd-sourced data.

In -more recent work [14], a robot is tasked to stack objects of different colours and complex shapes such as trapezoids and parallelograms. Through an RL algorithm, the simulation is tasked to stack two of the objects. The complexity of the task is to extract how the shape of the objects affects the stack and overcome the distraction of the different colours of the objects. Other RL methods for learning object stacking use probabilistic inference for learning control (PILCO) to learn a task-dependent parameterised policy that generalises to tasks that differ only by a reward function [15]. In [16], a hierarchical RL method, integrating planning with RL is presented in which action planning is performed for high-level decision making and an RL agent is used for low-level motor control. Object stacking is formalised as a multi-step task planning problem and solved by a hierarchical RL framework in [17].

The capability to reason about object affordance property is important for a successful stacking. Research exists (e.g. [11, 18, 19]) on modelling physical properties of scenes of objects using neural network architectures to enable reasoning about the individual object behaviour and how they behave in pairwise object interactions. While efficient in extracting features, many machine learning models might not cater to providing an understanding of how these features relate to functional properties, and a semantic representation is needed to translate the information. Likewise, the results of the NetHack challenge organised as part of NeurIPS 2021 [20] showed that the robot systems based on symbolic methods outperformed machine learning agents. Arguably, this means that for intelligent reasoning about object stacking to take place, some level of semantic information needs to be introduced into the system.

In the light of this, we propose using symbolic representations of the semantic components present in the image schemas to approach object stacking. In the next section, we describe the preliminary methodology on how we intend to proceed with this problem.

3. Approaching Intelligent Stacking: Foundation and First Steps

Using a symbolic approach to robotic object stacking is a challenging task as the system needs to have access to all the (relevant) objects' properties. However, as it is based on image-schematic patterns it also provides a base to do intelligent reasoning to effectively predict which spatial arrangements will be possible and how it will affect the stability of the stack.

To demonstrate our proposed method, take a household task of clearing the table from a set of objects; a spoon, a plate, and a cup, see Figure 1. If these are to be stacked appropriately for transport, it is required information that (in most cases) the stack will be more stable if

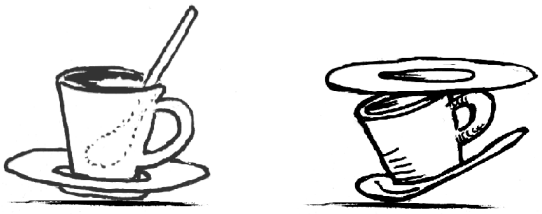

Figure 1: Stacking a cup, a saucer and a spoon in two different constellations.

small-sized objects like spoons are placed into the cup rather than balanced on the plate, or even worse, that the spoon is at the bottom. Finding the most optimal constellation is based on understanding the object affordances and dispositions.

3.1. Foundational Framework

Our framework relies on KnowRob [21], a knowledge representation and reasoning system. It has two parts. One part is an ontological knowledge base written in the Web Ontology Language (OWL). The other is a logic-based programming language, Prolog, which is used as a reasoner over the knowledge stored in the Mongo database. It provides an abstraction from sub-symbolic, high-dimensional data obtained from robot sensors by using high-level symbolic representations and thus enabling it to ground semantic information from its perceptions. Together with their image-schematic dispositions and affordance properties, objects are to be ontologically modelled in SOMA. This type of representation enables the robots to infer the functional aspect the objects through their affordance property [22]. In our framework, affordances are treated as bidirectional dispositions [23]. This means that for object having the trigger disposition of being a **container**, there must exist another object (real or hypothetical) that has the bearer disposition of **can be contained**.

3.2. Preliminary Rules for Stacking

For many stacking problems in households, placing smaller objects on the top is a sensible rule as it provides more stability. Additionally, objects of different materials and 'sturdiness' greatly impact the stackability.

In our working example of stacking tableware, we argued that the cup should ideally go on top of the saucer and the spoon inside the cup. In our framework, this means that the saucer possess the disposition property **Deposition** (SOMA's name for Support) and acts as a *bearer* for objects that can be deposited on top of it. In this case, the most suitable object is the cup, for which it affords the action description task of Stacking. Correspondingly, the cup is the *trigger* with the role of **DepositedObject** as it can be placed on top of the saucer. Similarly, the given set of objects can be checked for the disposition property of **Containment** (Containment) and can select a suitable trigger object which can be inserted into or contained within that object. Here the cup with its concave surface would have this disposition and take on the role of a container to relatively smaller objects such as the spoon. Establishing such relationships

among the objects reduces the search space for possible stack configurations and eliminates the possibility of unstable combinations such as placing the cup on top of the spoon and the saucer on top of the cup.

Based on such object properties and logical rules an algorithm for stacking order is to be implemented as part of the rule engine in KnowRob. Below are a few preliminary considerations that we estimate important for intelligent stacking.

- Objects with the **Deposition** disposition go below objects with the **Depositability** disposition.
- Objects with the **Deposition** disposition go below objects with the **Containment** disposition.
- Finally, objects with the **Containment** disposition go below objects with the **Containability** disposition.

The above considerations are not intended as complete, neither in terms of how they relate to object properties nor the relationships between them. For instance, there exist many categories in which this reasoning does not apply, regardless of their object dispositions. One example is objects classified as food, which, as a rule, should always go on top of tableware and never the other way around.

4. Discussion and future work

Stacking is a deceptively simple task for humans, with several underlying complexities when transferred to artificial and robotic agents. For robots to be able to function in household environments, intelligent reasoning skills of how object properties and affordances relate to their stackability need to be formally investigated. While still at an early stage, our system aims to help with this by enabling the robot to infer the order of objects to be stacked and the motion constraints that have to be maintained.

Future work includes integrating the formal representations of stackable objects and their dispositions such as **Containment** and **Deposition**. We also intend to develop the algorithmic rules for determining the order on how to stack particular objects based on their dispositions. One important aspect to consider is that only a limited number of rules can be defined in KnowRob as they need to be handcrafted. In an ideal case, the system should instead be able to learn these rules by itself. tThere exist many methods for this that can be considered, but we envision a combination of observation data (human demonstration data) to model the relation between the physical attributes of the object and the stacking stability and allow the system to extract rules from a curiosity-driven exploration in simulation [24] and [25].

Another step following this work is to establish a method to infer the motion constraints for the task to be executed by the robot as it has to handle objects with varying physical properties. These constraints are intended to be passed onto Giskard [26], a constraint-based robot controller. This will enable Giskard to take the predicate goal as input and convert them to robot control commands. This means that defining image-schematic goals such as CONTACT *(table, cup)* as part of the task description is sufficient for the robot to infer that the cup should

be placed *OnTopOf* the table and correctly execute the sequence of actions necessary to reach the goal.

Acknowledgements

The research reported in this paper has been partially supported by the German Research Foundation DFG, as part of Collaborative Research Center (Sonderforschungsbereich) 1320 "EASE - Everyday Activity Science and Engineering", University of Bremen (http://www.ease-crc.org/).

References

[1] B. Bennett, A. Cohn, Automated common-sense spatial reasoning: still a huge challenge, in: S. Muggleton, N. Chater (Eds.), Human-Like Machine Intelligence, Oxford University Press, 2021.

[2] D. Batra, A. X. Chang, S. Chernova, A. J. Davison, J. Deng, V. Koltun, S. Levine, J. Malik, I. Mordatch, R. Mottaghi, et al., Rearrangement: A challenge for embodied ai, arXiv preprint arXiv:2011.01975 (2020).

[3] L. Smith, M. Gasser, The development of embodied cognition: Six lessons from babies, Artif. Life 11 (2005) 13–30. URL: https://doi.org/10.1162/1064546053278973. doi:Path10.1162/1064546053278973.

[4] M. Johnson, The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason, The University of Chicago Press, Chicago and London, 1987.

[5] M. M. Hedblom, Image Schemas and Concept Invention: Cognitive, Logical, and Linguistic Investigations, Cognitive Technologies, Springer Computer Science, 2020.

[6] J. J. Gibson, The theory of affordances, in: R. Shaw, J. Bransford (Eds.), Perceiving, Acting, and Knowing: Toward an Ecological Psychology, NJ: Lawrence Erlbaum, Hillsdale, 1977, pp. 67–82.

[7] M. M. Hedblom, M. Pomarlan, R. Porzel, R. Malaka, M. Beetz, Dynamic action selection using image schema-based reasoning for robots, in: Proceedings of the 7th Joint Ontology Workshops, 2021.

[8] K. Dhanabalachandran, V. Hassouna, M. M. Hedblom, M. Küempel, N. Leusmann, M. Beetz, Cutting events: Towards autonomous plan adaption by robotic agents through image-schematic event segmentation, in: Proceedings of the 11th on Knowledge Capture Conference, K-CAP '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 25–32. URL: https://doi.org/10.1145/3460210.3493585. doi:Path10.1145/3460210.3493585.

[9] M. M. Hedblom, O. Kutz, R. Peñaloza, G. Guizzardi, Image schema combinations and complex events, KI-Künstliche Intelligenz 33 (2019) 279–291.

[10] N. T. Dantam, Z. K. Kingston, S. Chaudhuri, L. E. Kavraki, Incremental task and motion planning: A constraint-based approach, in: Robotics: Science and systems, volume 12, Ann Arbor, MI, USA, 2016.

[11] O. Groth, F. B. Fuchs, I. Posner, A. Vedaldi, Shapestacks: Learning vision-based physical

intuition for generalised object stacking, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 702–717.
[12] F. Furrer, M. Wermelinger, H. Yoshida, F. Gramazio, M. Kohler, R. Siegwart, M. Hutter, Autonomous robotic stone stacking with online next best object target pose planning, in: 2017 IEEE international conference on robotics and automation (ICRA), IEEE, 2017, pp. 2350–2356.
[13] N. Abdo, C. Stachniss, L. Spinello, W. Burgard, Robot, organize my shelves! tidying up objects by predicting user preferences, in: 2015 IEEE international conference on robotics and automation (ICRA), IEEE, 2015, pp. 1557–1564.
[14] A. X. Lee, C. Devin, Y. Zhou, T. Lampe, K. Bousmalis, J. T. Springenberg, A. Byravan, A. Abdolmaleki, N. Gileadi, D. Khosid, C. Fantacci, J. E. Chen, A. Raju, R. Jeong, M. Neunert, A. Laurens, S. Saliceti, F. Casarini, M. Riedmiller, R. Hadsell, F. Nori, Beyond pick-and-place: Tackling robotic stacking of diverse shapes, 2021. arXiv:2110.06192.
[15] M. P. Deisenroth, P. Englert, J. Peters, D. Fox, Multi-task policy search for robotics, in: 2014 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2014, pp. 3876–3881.
[16] M. Eppe, P. D. Nguyen, S. Wermter, From semantics to execution: Integrating action planning with reinforcement learning for robotic causal problem-solving, Frontiers in Robotics and AI 6 (2019).
[17] X. Yang, Z. Ji, J. Wu, Y.-K. Lai, C. Wei, G. Liu, R. Setchi, Hierarchical reinforcement learning with universal policies for multistep robotic manipulation, IEEE Transactions on Neural Networks and Learning Systems (2021) 1–15. doi:Path10.1109/TNNLS.2021.3059912.
[18] P. W. Battaglia, R. Pascanu, M. Lai, D. Rezende, K. Kavukcuoglu, Interaction networks for learning about objects, relations and physics, 2016. arXiv:1612.00222.
[19] D. Raposo, A. Santoro, D. Barrett, R. Pascanu, T. Lillicrap, P. Battaglia, Discovering objects and their relations from entangled scene representations, arXiv preprint arXiv:1702.05068 (2017).
[20] E. Hambro, E. Grefenstette, H. Küttler, T. Rocktäscel, The nethack challenge: Dungeons, dragons, and tourists, https://nethackchallenge.com/report.html, 2021. Accessed on 14.12.2021.
[21] M. Beetz, D. Beßler, A. Haidu, M. Pomarlan, A. K. Bozcuoglu, G. Bartels, KnowRob 2.0 – A 2nd Generation Knowledge Processing Framework for Cognition-Enabled Robotic Agents, in: 2018 IEEE Int. Conf. on Robotics and Automation, ICRA 2018, Brisbane, Australia, May 21-25, 2018, 2018, pp. 512–519. doi:Path10.1109/ICRA.2018.8460964.
[22] D. Beßler, R. Porzel, M. Pomarlan, M. Beetz, R. Malaka, J. Bateman, A formal model of affordances for flexible robotic task execution, in: ECAI 2020, IOS Press, 2020, pp. 2425–2432.
[23] M. Turvey, Ecological foundations of cognition: Invariants of perception and action. (1992).
[24] M. Gasse, D. Grasset, G. Gaudron, P.-Y. Oudeyer, Causal reinforcement learning using observational and interventional data, 2021. arXiv:2106.14421.
[25] M. M. Hedblom, M. Pomarlan, R. Porzel, Panta Rhei: curiosity-driven exploration to learn the image-schematic affordances of pouring liquids, in: Proceedings of the 29th Irish Conference on Artificial Intelligence and Cognitive Science, Dublin, Ireland, 2021.
[26] Z. Fang, G. Bartels, M. Beetz, Learning models for constraint-based motion parameteriza-

tion from interactive physics-based simulation, in: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2016, pp. 4005–4012.