Dynamic Action Selection Using Image Schema-based Reasoning for Robots

Maria M. Hedblom[1], Mihai Pomarlan[2], Robert Porzel[3], Rainer Malaka[3] and Michael Beetz[1]

[1]Institute of Artificial Intelligence, University of Bremen, Am Fallturm 1, 28359 Bremen, Germany
[2]Applied Linguistics Department, University of Bremen, Uni-Boulevard 13, 28359 Bremen, Germany
[3]Digital Media Lab, University of Bremen, Bibliothekstr. 5, 28359 Bremen, Germany

Abstract

Dealing with robotic actions in uncertain environments has been demonstrated to be hard. Many classic planning approaches to robotic action make the closed world assumption, rendering them inefficient for everyday household activities, as they function without generalizability to other contexts or the ability to deal with unexpected changes. In contrast, humans robustly execute underspecified instructions in unfamiliar environments. In this paper, we initiate our research program where we propose the use of functional relations in the form of image-schematic micro-theories, formally represented in ISL^{FOL}, to enrich action descriptors with semantic components. It builds on the body of work in embodied cognition showing that human conceptualization of action sequences is founded on abstract patterns learned from physical experiences in the form of spatiotemporal relationships between object, agents and environments. These theories are used to inform action selection mechanisms for behavioral robotics written in EL++ and we argue how these micro-patterns can be applied in a more general way to deal with underspecified action commands and commonsense problem-solving.

Keywords

Cognitive robotics, image schemas, reasoning, uncertain environments, action descriptors

1. Introduction

Robot agents are starting to accomplish human-scale everyday manipulation tasks such as setting a table, cleaning up, and preparing (very) simple meals. Most knowledge representation approaches to reasoning about actions conceptualize the repertoire of robots as a state transition system, with actions as atomic transitions between states [1]. Representatives of this research approach are PDDL [2], situation [3] and event calculus representations [4, 5] and their variations.

This abstraction is critical from the robot agent perspective because the main reasoning task of a robot agent is to infer how it has to move its body in order to accomplish an underdetermined task such as "put the oat milk on the table" without causing unwanted side effects. However,

CAOS 2021: 5th Workshop on Cognition And OntologieS, held at JOWO 2021: Episode VII The Bolzano Summer of Knowledge, September 11-18, 2021, Bolzano, Italy
hedblom@uni-bremen.de (M. M. Hedblom); pomarlan@uni-bremen.de (M. Pomarlan); porzel@tzi.de (R. Porzel); malaka@tzi.de (R. Malaka); beetz@uni-bremen.de (M. Beetz)
0000-0001-8308-8906 (M. M. Hedblom); 0000-0002-1304-581X (M. Pomarlan); 0000-0002-7686-2921 (R. Porzel); 0000-0001-6463-4828 (R. Malaka); 0000-0002-7888-7444 (M. Beetz)

the use of abstract action models make robots inflexible as the same task can be implemented through different body motions: the robot can close a door with its hand but, analogically, it can also use its elbow if the hand is occupied by carrying an object. Further, the same motion can implement different tasks: a pushing motion can close a door but also position an object more accurately. The lack of such reasoning capabilities not only limits the manipulation tasks that robots can perform, but the accuracy by which it performs the actions as well.

There are approaches to include more detailed representations, such as the combination of task and motion planning. However, motion planning typically tries to compute collision-free paths rather than motions that achieve certain effects [6]. Approaches to axiomatizing manipulation actions have been attempted for some commonsense reasoning problems, e.g., egg cracking [7]. However, these often result in long and complicated axiomatizations that are difficult to ground into the action execution systems of robots.

In this paper, we propose to equip knowledge representation and reasoning systems for robotic agents with a generalizable layer of understanding of the conditions in the environment. This allows the robot to reason about action execution in terms of motion types when encountering unexpected changes in a given situation. For any physical state of affairs in the world, this layer of understanding can be described using a set of functional relationships between objects, agents and environments, called *image schemas*.

While in an early stage, we will argue that the contribution of our proposition is multifold:

(i) Reasoning about functional relationships: It allows robots and artificial agents to reason about the functional relationships between objects and other entities situated in complex environments in a cognitively-plausible way. *(ii) Reasoning about alternatives to a plan:* This semantic grounding of the environment offers problem-solving capabilities in that it provides means to expand the reasoning outside the specified action plan. This enables corrective actions such as avoidance of blockages. *(iii) Increase adaptability through analogy:* It also offers generalizability in that information relevant to one situation can be analogically transferred to another situation. *(iv) Improve natural language understanding:* Building on research founded in cognitive linguistics, it enables a large body of work to be leveraged towards increasing natural language understanding and for robots to follow human instructions and commands more accurately.

2. Action Semantics with Image Schemas and Affordances

Adult humans are creatures of habit. Repeated experiences with similar situations, i.e. particular states of the world, shape us into experts on a variety of situations possible in our environment. We can use generalized patterns of information from our previous experiences to reason about outcomes of uncertain situations. These expectations extend to action instructions as well. Asking someone to *set the table* or *get the milk* will likely result in a satisfactory outcome, regardless of whether that person has been in that particular kitchen or not. Humans have extensive understanding of the affordances of a kitchen, which tools are needed for eating, and an appropriate set-up of a dining table given different contexts, such as type of meal or number of participants. In comparison, robots, which are relatively proficient in well-defined tasks in constrained environments [8, 9], struggle when confronted with vague and underspecified instructions.

To aid this research agenda, we propose to utilize the underlying patterns of expectations found in humans as suggested by the theory of embodied cognition [10]. It proposes that our conceptualization and understanding of the environment comes from perceiving and interacting with it. Such information is formed into generalized patterns, image schemas. These encompass the spatiotemporal relationships between objects, agents and environments[1].

One way this could be computationally realized is to characterize the image schemas in relation to the affordances they require [11]. Concrete examples are, for instance, how a glass affords CONTAINMENT of liquids and a plate offers SUPPORT for food, but also abstract concepts and more dynamic transformations are included in this, e.g. how 'space to grow' is ENABLEMENT and SCALING. Turning affordance theory into a computationally applicable theory for commonsense reasoning of events and actions has been approached in both theoretical and applied computer science [12, 13]. Image schemas have also been proposed to provide patterns to organize hybrid reasoning involving qualitative and quantitative descriptions of scenes [14].

While affordances can be described by the interplay of respective dispositions of objects and agents [15], image schemas offer another layer of abstraction describing common manifestations of affordance-based configurations that, furthermore, constitute the meaning of many linguistic constructions [16]. Consequently, it has been argued that the meanings of linguistic units can be traced back into the generalized patterns of the sensorimotor experiences seen as *image schemas*, as well as *force-dynamic schemas* [17].

The image schemas capture relationships such as SOURCE_PATH_GOAL (SPG) - depicting movement of an object between two points, LINK - the force-dynamic relationship that connects objects with one another, and VERTICALITY - vertical movement, and relative position and symmetry on the vertical axis. Another strength of using image schemas for formal research on underspecified environments and instructions is that they have been shown to come in different levels of specificity, e.g. the difference between a tight and loose containment, or how SOURCE_PATH_GOAL ranges from simple object movement to increasingly specific notions with source and goal locations [18]. For instance, a milk-carton is a prime example of a tight container which functions as a required component of transportation (SPG) of any liquid. They have been argued to manifest as graph hierarchies of increasingly complex image-schematic relationships, and have been formalized accordingly [18].

The embodied grounding of image schemas makes them a prime subject to be learned through statistical methods and deep learning. For example, in the work by [9], subsymbolic information of robotic object manipulations are collected and transformed into symbolic Narrative-Enabled Episodic Memories (NEEMs) featuring a semantics based on the SOMA ontology [19]. These are collected into an episodic memory knowledge base and are used to learn general knowledge about particular situations – such as, that milk is usually in the fridge, or that cups can be found in cupboards. A robot that has a NEEM about a fridge containing such perishable objects as cartons of oat milk, can use this information when tasked to "put the oat milk on the table." Within these NEEMs, the image-schematic relationships exist as background knowledge evoked by the possible actions. In the next section, we provide an overview of the formal system that we employ

[1] Some cognitive scientists might oppose this rather narrow view of image schemas. However, we argue that this is a good starting point for modeling and simulating computational intelligence as this definition is more formalizable than the abstract, multi-modal notions that may be more accurate from the cognitive perspective.

for representing the abstract image schemas before moving onto a working example.

3. Action and Event Analysis using ISL^{FOL}

It has been suggested that the conceptualization of particular events and actions can be described using image schema profiles [20]. While these profiles tend to be conceptually unstructured and describe simple groups of the image-schematic relationships that represent the way we think about a particular concept or event, recent research in knowledge representation has brought forth an approach to employ *structured combinations* of image schemas to describe in conceptual detail what functional relationships take place in certain activities [21, 22]. After introducing ISL^{FOL}, we use the household action of 'fetching milk' to demonstrate how such structured combinations can look for our action selection approach.

3.1. The Image Schema Logic, ISL^{FOL}

ISL^{FOL}, the image schema logic, is an expressive multi-modal logic intended to capture the basic spatiotemporal interactions present in image-schematic events [18]. In short, it combines the Region Connection Calculus (RCC) [23], Ligozat's Cardinal Directions (CD) [24], Qualitative Trajectory Calculus (QTC) [25], with 3D Euclidean space assumed for the spatial domain, and Linear Temporal Logic over the reals (RTL). This combination of calculi allows the formal modeling of spatial relationships between objects and regions in RCC and their relative movement using a reduced version of QTC with the following syntax:

- O_1 moves towards O_2's position: $O_1 \rightsquigarrow O_2$,
- O_1 moves away from O_2's position: $O_1 \leftsquigarrow O_2$
- O_1 is at rest with respect to O_2's position: $O_1 \mid\!\circ\, O_2$.

The temporal dimension is based on linear temporal logic (RTL) over the reals [26] with future and past operators. The syntax of this logic is defined by the grammar

$$\varphi \ ::= \ p \quad \mid \top \mid \neg\varphi \mid \varphi \wedge \varphi \mid \varphi \mathbf{U} \varphi \mid \varphi \varphi$$

where $\varphi \,\mathbf{U}\, \psi$ reads as "φ holds, until ψ" and $\varphi \ \psi$ reads as "φ holds, since ψ." As it is standard in temporal logic, we can define additional temporal operators based on these, for instance, operators like: $\mathbf{F}\varphi$ (at some time in the future, φ) is defined by $\top \mathbf{U} \varphi$; and, $\mathbf{G}\varphi$ (at all times in the future, φ) is defined as $\neg \mathbf{F} \neg \varphi$,

In ISL^{FOL}, the temporal structures, often disregarded due to the increase in complexity in formal image schema modeling, constitute the primary model-theoretical object, e.g., a linear order to represent the passage of time, in which complex propositions that employ a secondary semantics are included. The atoms are topological assertions about regions in space using RCC, the relative movement of objects with respect to each other using QTC, and relative orientation, using CD. We refer the reader to [18] for more details on this language.

ISL^{FOL} axioms are based on a concept language in First Order Logic (FOL), making it an expressive tool to represent different situations and concepts. The idea is that each image schema (e.g. LINK (x,y) and SOURCE_PATH_GOAL (x,p,s,g)) is modelled using the logic's semantics and is represented using FOL.

3.2. The Underlying Logic of Image Schemas

While ISL^{FOL} is predominantly a modeling language, the image schemas are defined by internal logical rules that can be used to reason with. For instance, the CONTAINMENT relationship is transitive in that an object that is contained within another object will move if the container moves (consider how the milk will remain in the carton when the carton moves), see:

$$\forall a, b{:}Object, s, g{:}Region, p{:}Path \in X :$$
$$(\text{CONTAINED\_IN}(a, b) \wedge \text{SPG}(b, p, s, g)) \rightarrow \text{SPG}(a, p, s, g)$$

Following the same reasoning, a LINKed relationship ensures that what happens to one of the objects is transferred also to the linked object (a robot holding a carton will ensure that if the robot moves, then the carton moves as well). Likewise, a SUPPORTed object will move, if the SUPPORTing object is transferred to another location (consider a robot carrying the milk-carton on a tray).

These kinds of built-in rules for our expectations of the environment offer the possibility to make predictions for the outcomes of particular actions. These rules also offer the possibility for dealing with unexpected problems that can arise. For instance, if the robot's movement is BLOCKed by another entity getting in the way, the robot would be able to reason if that object is in movement and thus, wait for it to pass, or if it is still, circumvent the BLOCKAGE. Likewise, a robot would be able to reason about how a container needs to be opened for something to be able to exit the container. The next section will tackle this and related reasoning challenges.

3.3. Fetching and Placing Actions

One of the most common things we ask other people to do is to fetch different things for us. From a household robot's perspective, the instruction *Get the oat milk* can be described as a particular instance of a Fetch-And-Place action descriptor. In addition to the call to perform the action, the instruction includes additional important – albeit implicit – forms of knowledge. First, the noun *oat milk* evokes information about what it is, e.g. a perishable good, where it is usually stored and so on. The SOMA ontology of everyday activities [19], provides a foundational framework as well as additional dedicated modules for the household domain, expressing that `OatMilk` is_a `PerishableSubstance` that is $stored\_in$ `CoolingDevices` such as Refrigerators. In image-schematic terms, all liquid substances (LS) are further specified as requiring a container (C) for transport:

$$\forall LS, C{:}Object(Move(LS) \leftrightarrow \text{CONTAINED\_IN}(LS, C) \wedge Move(C))$$

Second, the verb *get* requires the understanding of transporting something by evoking a SOURCE_PATH_GOAL schema with the locations of the source and the goal being an integral part of the transportation expressed. The ABox for the oat milk reveals that $pos_t(OatMilk) : Fridge$, and the robot needs to understand that the source of the instruction (the person speaking) reveals the goal of the transportation – namely, close to the speaker, $pos_{goal}(OatMilk) : Speaker$.

If the robot has an episodic memory concerning oat milk, in addition to the ontological semantics contained in a NEEM, it has information that milk is a perishable liquid stored in

a container that is placed on a shelf in a fridge. Image-schematically this represents nested CONTAINMENT: the object in question is inside one container (carton) which in turn is inside another container (fridge) for different purposes. This is crucial information for knowing how to treat the object. While a robot could theoretically move the entire fridge to the person asking for oat milk, this is an inefficient way of solving the problem. Likewise, it is not very smart to take the oat milk out of the carton before attempting to moving it. One reason for this is because the purposes for CONTAINMENT are fundamentally different, the liquid needs tight CONTAINMENT for transportation, whereas the fridge's loose CONTAINMENT has nothing to do with transportation but instead for static storage and preservation.

The second image-schematic component is the movement of the oat milk from the fridge to the person asking for it. This represents the construction of SOURCE_PATH_GOAL capturing different levels of specificity of the conceptualizations of movement. The classical linguistic interpretation is that a trajector (object or agent) moves along a path from a particular SOURCE to a determined GOAL. In this case, the robotic agent needs to be able to deduce that the SOURCE is the initial location inside the fridge and that the GOAL is to reach the near vicinity of the person asking for the milk. This may sound like a trivial problem, but it includes not only the SPG schema, but also the CONTAINMENT schema as the Going_OUT schema is part of the schema's dynamic relationships and can, in isolation to the whole event, also be described as a combination of the image schemas SPG and CONTAINMENT.

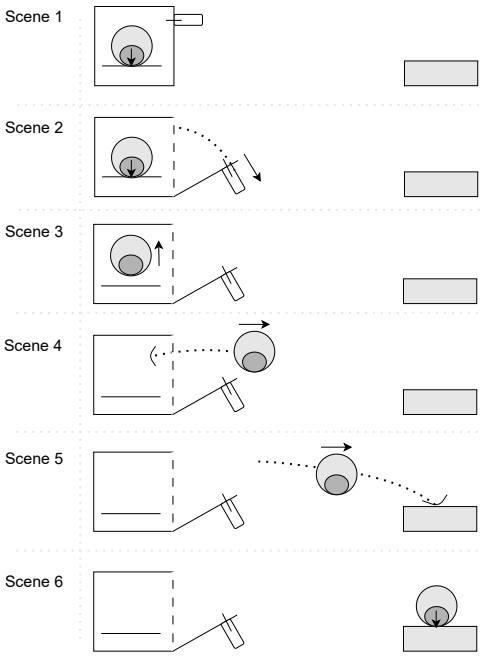
The third image-schematic relationship we cover in our working example is SUPPORT. Obviously, all objects are supported by the ground, but for a robot to masterfully be able to manipulate objects, it is not possible for it to neglect the naïve rules and structures of placing things on top of other things. With the oat milk, this offers important information to be transferred from the source to the goal state. At the source inside the fridge, the oat milk is vertically[2] SUPPORTed on a shelf. This could be seen as a required property of the oat milk-carton throughout the action and in the goal state, as it has an opening at the top.

In ISL^{FOL}, most[3] image schemas and their hierarchical graphs can be formally represented in the form of ontological patterns. Any formalization using ISL^{FOL} would use their placeholder names to access the full axiomatization. For instance, the LINK, image schema is formalized using $EC(x, y) \wedge force(x, y) \wedge force(y, x)$ describing how for two objects, x,y, to be linked, they are externally connected (asserted by the RCC8 operator EC) and there is a force from each respective object towards the other. The action event can be described as six different image-schematic states, depicted and verbalized in Figure 1.

The purpose of understanding image-schematic relationships in scenes is not only to identify the logical axiomatizations thereof in ISL^{FOL}, it is to describe the underlying patterns for understanding the meaning of events and actions. By giving an artificial agent access to this level of semantic layer, it becomes possible to reason about this particular scenario, as well as to transpose this knowledge to other similar situations with different objects and contexts. Additionally, this type of reasoning becomes vital to recover from errors and mishaps that might occur[27].

[2]Assuming that the carton is elongated on the vertical axis.

[3]The logic is not able to elegantly handle transformational relationships involved in image schemas such as Spiraling or Scaling, as they are more loosely described in terms of other objects and instead their previous states. For this the addition of mathematical functions could be a way forward.



Scene 1: Initial state. There is a LINK between the Handle and the FridgeDoor. The FridgeDoor is in CONTACT with the Fridge. The Milk is CONTAINED_INSIDE in the Carton, the Carton is VERTICAL and SUPPORTed by the Shelf which is CONTAINED_INSIDE the Fridge.

Scene 2: Opening the door. To open the door, the FridgeDoor needs to leave its state of being in CONTACT with the Fridge. This is done by SOURCE_PATH, where the Handle moves along a CircledPath away from the Fridge.

Scene 3: Lift milk-carton. Through VERTICAL oriented SOURCE_PATH_GOAL the SUPPORT from the Shelf on the Carton is removed, since the Milk is CONTAINED_INSIDE in the Carton it MOVEs as well.

Scene 4: Take the Milk out. The Carton goes OUT of the Fridge through its opening, and is no longer CONTAINED_INSIDE.

Scene 5: Milk goes to its destination. Since the Carton MOVEs towards the Table, the Milk MOVEs to the Table.

Scene 6: Goal state of the Milk resting on the Table: The Carton's movement is BLOCKed when it comes in CONTACT with the Table, turning the force from the movement into the reversed force as SUPPORT from the Table.

Figure 1: Image-schematic scene breakdown of taking the milk out of the fridge.

4. The Contributions in Practice

In this section, we will illustrate how ISL^{FOL} can contribute to robotics to address the issues listed in the introduction. We will, for each item in turn, explain the underlying problems, provide an example of reasoning in ISL^{FOL} to address such problems, and then show how the results inform the construction of axioms in a simpler formalism, EL++ [28, 29], that can be used in a quick perception-action loop of a robot. The overall approach then is to use ISL^{FOL} "off-line," in a robot's idle moments, to either imagine new possible situations or analyze past experience, and to encode knowledge thus obtained into simple rules that can be employed effectively for reflexive, fluid actions.

4.1. Reasoning about Functional Relations

One thing that we have repeatedly stressed in this paper is how the image schemas offer a cognitively plausible method for artificial agents to reason about their surroundings. A simple example is how humans, birds, and other animals understand that if you want to eat a nut you have to crack the shell before you can take it out. Likewise, humans are quite experienced with the understanding that if you want to take the milk out of the fridge, it is not possible unless you first open the fridge door.

This level of commonsense reasoning is intuitive in biological intelligence, but it has been demonstrated to be time-consuming to axiomatize the full scenarios (as with the egg cracking

problem) and inefficient to rely purely on statistical methods. This is where the image schema logic could play a vital role.

To give an example of how a CONTAINMENT inference can work in ISL^{FOL}, consider the following axiom about solid objects (for simplicity, we omit to include conditions about no parthood relations between the objects involved)[4]:

$$\forall O_1, O_2 : SolidObject \quad O_1 \neq O_2 \rightarrow \neg PO(O_1, O_2)$$

ISL^{FOL} allows us to define a few regions of interest around an object: its interior (which we will take here as a primitive predicate), its exterior, and openings. Let then a closed container be an object with an interior but no openings, and an enclosed object be one contained inside a closed container.

$$\begin{aligned} exterior(E, O) &:= interior(I, O) \wedge E = Com(O \cup I) \\ opening(op, O) &:= interior(I, O) \wedge exterior(E, O) \\ &\wedge PP(op, E) \wedge EC(op, I) \\ closed(O) &:= \neg\exists op \;\; opening(op, O) \\ enclosed(O) &:= \exists C \;\; closed(C) \wedge \text{CONTAINED\_IN}(O, C) \end{aligned}$$

Reasoning inside the RCC8 fragment of ISL^{FOL} then establishes that, for solid objects, to touch an enclosed object without also being enclosed in the same container is impossible:

$$\begin{aligned} &\forall R, O, C : SolidObject \\ Enclosed(O, C) \wedge \neg Enclosed(R, C) &\rightarrow \neg EC(R, O) \end{aligned}$$

In upcoming work, we intend to use EL++ ontologies to encode image-schematic knowledge for action selection. The previous result from ISL^{FOL} may be approximated in EL++ as:

$$\begin{aligned} Enclosed &\equiv \exists isContainedIn.Obj \\ Free \sqcap Enclosed &\sqsubseteq \bot \\ EC \sqcap \exists(hasParticipant.Enclosed) & \\ \sqcap (\exists hasParticipant.Free) &\sqsubseteq \bot \end{aligned}$$

Meaning, a relationship in which one participant is a free object and another is enclosed cannot be of type EC. This has consequences for action selection because, e.g., an EC relation between a gripper and an item to grasp is necessary. If such a relation is impossible, then reaching for an item should be delayed. Further axioms could pinpoint possible alternative actions, such as to manipulate handles on the container to open it.

4.2. Reasoning about Alternatives to a Plan

Another vital contribution that the image schemas bring to reasoning about action execution is dealing with unexpected situations. In a kitchen, it is not unlikely that multiple humans are

[4] As defined in RCC8: PO - Partial Overlap; EC - Externally Connected. PP - Proper Part, Com - set theoretical complement in $\mathbb{R}^3$.

present at the same time and may continuously change the state of the environment. If a person at a table asks the robot to fetch them the milk, the robot needs to be able to reason about any changes that might have taken place. Perhaps the originally intended path is no longer possible to take because another person put something in front of the robot or is actively crossing the path in that instance. Not only it is expected that the robot should stop and not run over the other human, but it should also be able to redirect its movement if the object on the path is unlikely to move anytime soon, i.e. if it does not have a movement state. If the human blocking its movement is simply walking past as part of an SOURCE_PATH_GOAL of its own right, then the robot can simply take a break before continuing along the same path. However, if the person, or item, remains on the path, the robot needs to be able to reroute to still be able to successfully reach the goal of the instructions. It would do this, by generating a new SOURCE_PATH_GOAL construct in which the source is no longer based at the fridge, but at the location of the BLOCKAGE, and the path is no longer the fastest route (or what previously had been suggested) but a route that bypasses the blocking object.

The above discussion can be formally represented in ISL^{FOL} as follows with omitting the assumption that the source and goal (S,G) are TPP, tangential proper parts, of the path. Assuming again the axiom of solid objects, and the following axiom about SPG:

$$\forall A, X{:}SolidObject, \forall S, G{:}Region, \forall P{:}Path$$
$$\mathrm{SPG}(A, P, S, G) \wedge \mathbf{G}(PO(X, P)) \rightarrow \mathbf{F}(PO(A, X))$$

That is, if a trajector A follows a path P, it will pass through any region along the path. Then, one can show the following:

$$\forall A, B{:}SolidObject, \forall S, G{:}Region, \forall P{:}Path$$
$$A \neq B \wedge \mathbf{G}(PO(B, P)) \rightarrow \mathbf{G}(\neg \mathrm{SPG}(A, P, S, G))$$

In other words, if a path is forever blocked, it is never possible to use it to go from S to G.

As before, we wish to provide a robotic agent with some simple rules to select, or filter out, actions. The previous result from ISL^{FOL} could be approximated in EL++ thusly, assuming appropriate recognition procedures for entities such as paths and stationary objects, and appropriate controllers for actions such as moving the robot base:

$$BlockedPath \equiv Path \sqcap \exists overlappedBy.StaticObj$$
$$BaseMovement \sqcap \exists uses.BlockedPath \sqsubseteq \bot$$

That is, no action that involves moving the base should use a blocked path. If such a path is the one currently used by the robot, it should search for a different one. A similar set of simple rules might encode for the robot that it may be worth waiting if a path is blocked by a moving object:

$$BusyPath \equiv Path \sqcap \exists overlappedBy.MovingObj$$
$$BaseMovement \sqcap \exists uses.BusyPath \sqsubseteq DelayedAct$$

Such a set of action rules would be justified if one believed that objects moving away from a path eventually do not overlap it:

$$\forall B{:}SolidObject, \forall P{:}Path:$$
$$\mathbf{G}(B \leftarrow P) \rightarrow \mathbf{F}(\neg PO(B, P))$$

4.3. Increase Adaptability through Analogy

Bypassing things like BLOCKED_MOVEMENT by redirecting routes requires the robot to rethink its actions based on new image-schematic states of the world. This is useful, but it is also possible to use this in a more general way by abstracting away from the actual parts of the world.

This builds on the idea of analogical reasoning, that there are underlying patterns that can be transferred from an information rich source domain to an underspecified target domain. For robotic actions, this offers the possibility to reuse previously learned relationships. One crucial component for successful analogical transfer is that the source and target share the same structure. In the settings of functional relations as the foundation to guide robotic action selection, these patterns are also useful as a basis for generalization. For instance, if a robot has access to the image-schematic information that for something to be taken out of a closed container as for instance a fridge, it can use this information to reason about similar CONTAINMENT situations. It can use this generalized information to take the lasagna out of the oven (take note of how Figure 1 would be exactly the same with this example), a letter out of a envelope and join the masses of biological species whose evolutionary predecessors learned long ago that the nut needs to come out of its shell.

A less complex analogy, as only one object needs to be exchanged, is how it is possible to close a door with another body part than a hand, should it be preoccupied holding milk-cartons or lasagnas. In this case, what we want is to describe, image-schematically, what it means for a solid object to push another, via some third solid object:

$$push(R, A, L) :=$$
$$(R = L \vee \textsc{Link}(R, L)) \wedge (L \rightsquigarrow A) \wedge EC(L, A)$$

A robot using EL++ rules to select and parameterize actions might then be interested in classifying what objects it can push with:

$$Pushing \sqsubseteq \exists uses.AffordsPushing$$
$$OwnBodyPart \sqsubseteq AffordsPushing$$
$$\exists linkedTo.OwnBodyPart \sqsubseteq AffordsPushing$$

4.4. Improve Natural Language Understanding

Another important requirement of a successful household robot system is to be able to understand human instructions. Natural language instructions are usually underspecified and contain vast amounts of ambiguity, polysemy, and implicit information that need to resolved and explicated in order to execute the corresponding actions appropriately. Another problem is that instructions often omit vital semantic components such as determiners, quantities and even the object themselves [30]. For instance, in the example above *get the oat milk*, neither the source nor the target locations are made explicit next to the omission of the addressee. Yet any adult human would be able to successfully reach the correct goal state based on this instruction. While we have not dived deeper into the linguistic aspects of image schemas for this particular paper, the theory, analysis and application of image schemas stem from research in cognitive linguistics.

To improve natural language understanding in robotics, with a special focus on instructions, we employ an efficient construction-based parser [31] that produces semantic specifications as connected RDF triples that represent as much of the meaning of the instructions as contained in the textual commands. All terms in these semantic specifications are aligned to the SOMA-SAY module [32] that is part of the larger SOMA framework [19] that rests on the DUL+D&S foundational ontology [33]. Image-schematic theories are part of the descriptive branch of SOMA and constitute the central anchoring point of the semantic representations of the instructions given to the robotic agents. While these OWL-DL based representations only afford limited reasoning as compared to ISL^{FOL}, we ensure a seamless usage of the ensuing semantic representations by using the terms provided in SOMA as a *lingua franca* throughout the system. Additional mechanisms that are part of our deep language understanding pipeline are needed for further explicating the implicit information to arrive at executable robotic action plans. This concerns, for example, the learning of tool selections via human computation approaches [34] or the setting of action parameters and execution variations by means of physics-based simulations that satisfy the expectation constraints provided by the given ISL^{FOL} models [14].

5. Discussion on Past and Future Work

Using commonsense reasoning to improve robotic action selection is not a novel idea, it has been a fundamental component since the beginning of formal research on intelligent systems (a comprehensive overview is given in [35]). Many researchers (e.g. [36, 8, 37]) have worked on providing robotic systems with human-like commonsense knowledge so that the agents more efficiently can plan and execute their actions.

Similar to the ideas in this paper, is the work by [38]. They consider activity knowledge as a means to fill the gaps in abstract instructions, but treat these gaps in a much more general point of view than the specifics found in image schemas. A more general approach to activity modeling for robotic agents is presented by the IEEE-RAS working group ORA [39]. The group has the goal of defining a standard ontology for various sub-domains of robotics, including a model for object manipulation tasks. It has defined a core ORA ontology [40], as well as additional modules for industrial tasks such as kitting [41]. In terms of methodology, we differ in the foundational assumptions we assert, with important consequences on the structure of our ontology, modeling workflow, and inferential power. In the case of ORA, the SUMO upper-level ontology is used as foundational layer. Compared to SUMO, we use a richer axiomatization of entities on the foundational layer, and put particular emphasis on the distinction between physical and social activity context.

Unlike most previous methods, that often build action descriptors for particular actions and scenarios, we suggest relying on the generalized information learned from the sensorimotor experiences, encoded as functional relationships based on image schemas. While there exists research on how to formalize image schemas [42] and to use them for simulation-based reasoning [43], the role they play in active applications is not quite as thoroughly investigated. Another novel approach is to construct hybrid reasoning pipelines that connect simulation-based reasoning with qualitative reasoning about functional relations [14], but this needs further investigation.

At this stage, the contributions of the paper remain purely theoretical. However, the novelty of

the approach and our conviction of the ideas underlying the core concepts and their contributions motivates future work.

The next steps of this research program is to further strengthen the applicability of this work by providing a more feasible connection between ISL^{FOL} and simple formal languages commonly used in robotics, such as EL++. Additionally, we intend to develop an image schema parser that can identify and extract image-schematic relationships from the subsymbolic data of robotic simulations and visually recorded human activity to provide automation to the system. Thirdly, we aim to connect the formal part and the identification parser to the body of work in cognitive linguistics to improve the robotic agents' understanding of instructions in natural language.

Acknowledgements

The authors thank John Bateman and Fabian Neuhaus for valuable insights and constructive feedback on the paper. The research reported in this paper has been supported by FET-Open Project #951846 "MUHAI - Meaning and UNderstanding for Human-centric AI" funded by the EU Program H2020 and 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/). The research was conducted in sub-projects "P01 Embodied Semantics for the Language of Action and Change" and "R01 CRAM 2.0 - a 2nd Generation Cognitive Robot Architecture for Accomplishing Everyday Manipulation Tasks".

References

[1] M. Ghallab, D. Nau, P. Traverso, Automated Planning: theory and practice, Elsevier, 2004.

[2] Z. Kootbally, C. Schlenoff, C. Lawler, T. Kramer, S. Gupta, Towards robust assembly with knowledge representation for the planning domain definition language (pddl), Robot. Comput.-Integr. Manuf. 33 (2015) 42–55.

[3] R. Reiter, A logic for default reasoning, Artificial intelligence 13 (1980) 81–132.

[4] M. Thielscher, Introduction to the fluent calculus, Electronic Transactions on Artificial Intelligence 2 (1998) 179–192.

[5] M. Shanahan, The event calculus explained, in: Artificial intelligence today, Springer, 1999, pp. 409–430.

[6] J.-C. Latombe, Robot motion planning, volume 124, Springer Science & Business Media, 2012.

[7] L. Morgenstern, Mid-Sized Axiomatizations of Commonsense Problems: A Case Study in Egg Cracking, Studia Logica 67 (2001) 333–3384.

[8] L. Kunze, M. Tenorth, M. Beetz, Putting people's common sense into knowledge bases of household robots, in: Annual Conf. on Artificial Intelligence, Springer, 2010, pp. 151–159.

[9] M. Beetz, D. Beßler, A. Haidu, M. Pomarlan, A. K. Bozcuoglu, G. Bartels, Know Rob 2.0 - A 2nd Generation Knowledge Processing Framework for Cognition-Enabled Robotic Agents, Proc. - IEEE Int. Conf. on Robotics and Automation (2018) 512–519.

[10] L. Shapiro, Embodied Cognition, New problems of philosophy, Routledge, London, 2011.

[11] A. Galton, The Formalities of Affordance, in: M. Bhatt, H. Guesgen, S. Hazarika (Eds.), Proc. of workshop Spatio-Temporal Dynamics, 2010, pp. 1–6.
[12] M. Raubal, M. Worboys, A formal model of the process of wayfinding in built environments, in: Int. Conf. on Spatial Information Theory, Springer, 1999, pp. 381–399.
[13] D. Beßler, R. Porzel, M. Pomarlan, M. Beetz, R. Malaka, J. Bateman, A formal model of affordances for flexible robotic task execution, in: Proc. of the 24th European Conf. on Artificial Intelligence, 2020.
[14] M. Pomarlan, J. Bateman, Embodied functional relations: a formal account combining abstract logical theory with grounding in simulation, in: 11th Int. Conf. on Formal Ontology in Information Systems (FOIS), 2020.
[15] M. T. Turvey, Affordances and prospective control: An outline of the ontology, Ecological psychology 4 (1992) 173–187.
[16] N. Chang, J. Feldman, R. Porzel, K. Sanders, Scaling cognitive linguistics: Formalisms for language understanding, in: Proc. of the First International Workshop On Scalable Natural Language Understanding, 2002.
[17] L. Talmy, Toward a Cognitive Semantics. Volume 2: Typology and Process in Concept Structuring, Language, Speech, and Communication, MIT Press, Cambridge, MA, 2000.
[18] M. M. Hedblom, Image Schemas and Concept Invention: Cognitive, Logical, and Linguistic Investigations, Cognitive Technologies, Springer Computer Science, 2020.
[19] D. Bessler, R. Porzel, M. Pomarlan, S. Hoefner, J. Batemann, R. Malaka, M. Beetz, Foundations of the Socio-physical Model of Activities (SOMA) for Autonomous Robotic Agents, in: Proc. of the Int. Conf. on Formal Ontology in Information Systems, 2021.
[20] T. Oakley, Image Schema, in: D. Geeraerts, H. Cuyckens (Eds.), The Oxford Handbook of Cognitive Linguistics, Oxford University Press, 2010, pp. 214–235.
[21] M. M. Hedblom, O. Kutz, R. Peñaloza, G. Guizzardi, Image schema combinations and complex events, KI-Künstliche Intelligenz 33 (2019) 279–291.
[22] R. St. Amant, C. T. Morrison, Y.-H. Chang, P. R. Cohen, C. Beal, An image schema language, in: Int. Conf. on Cognitive Modeling (ICCM), 2006, pp. 292–297.
[23] D. A. Randell, Z. Cui, A. G. Cohn, A spatial logic based on regions and connection, in: Proc. of the 3rd Int. Conf. on Knowledge Representation and Reasoning (KR-92), 1992.
[24] G. Ligozat, Reasoning about cardinal directions, J. Vis. Lang. Comput. 9 (1998) 23–44.
[25] N. V. D. Weghe, A. G. Cohn, G. D. Tré, P. D. Maeyer, A qualitative trajectory calculus as a basis for representing moving objects in geographical information systems, Control and cybernetics 35 (2006) 97–119.
[26] M. Reynolds, The complexity of temporal logic over the reals, Annals of Pure and Applied Logic 161 (2010) 1063–1096.
[27] M. Diab, M. Pomarlan, D. Beßler, A. Abkari, J. Rossel, J. Bateman, M. Beetz, An ontology for failure interpretation in automated planning and execution, in: Fourth Iberian Robotics Conference, ROBOT '19, Porto, Portugal, 2019.
[28] F. Baader, S. Brandt, C. Lutz, Pushing the el envelope., 2005, pp. 364–369.
[29] F. Baader, C. Lutz, S. Brandt, Pushing the el envelope further, in: OWLED, 2008.
[30] R. Porzel, V. S. Cangalovic, J. A. Bateman, Filling constructions: Applying construction grammar in the kitchen, in: Proc. of the 11th Int. Conf. on Construction Grammar, Antwerp, Belgium, 2021.

[31] V. S. Cangalovic, R. Porzel, J. A. Bateman, Streamlining formal construction grammar, in: Proc. of the ICCG11 Workshop on Constructional approaches in formal grammar, Antwerp, Belgium, 2021.
[32] R. Porzel, V. Cangalovic, What say you: An ontological representation of imperative meaning for human-robot interaction, in: Proc. of the Joint Ontology Workshops, Virtual, Bozen-Bolzano, Italy, 2020.
[33] A. Gangemi, P. Mika, Understanding the semantic web through descriptions and situations, in: Proc. of the ODBASE Conference, Springer, 2003.
[34] J. Pfau, R. Malaka, We asked 100 people: How would you train our robot?, in: Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play, CHI PLAY '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 335–339.
[35] A. Olivares-Alarcos, D. Beßler, A. Khamis, P. Gonçalves, M. Habib, J. Bermejo, M. Barreto, M. Diab, J. Rosell, J. Quintas, J. Olszewska, H. Nakawala, E. Pignaton de Freitas, A. Gyrard, S. Borgo, G. Alenyà, M. Beetz, H. Li, A review and comparison of ontology-based approaches to robot autonomy, The Knowledge Engineering Review 34 (2019).
[36] N. J. Nilsson, Shakey the robot, Technical Report, SRI, Menlo Park, CA, 1984.
[37] R. D. Nielsen, R. Voyles, D. Bolanos, M. H. Mahoor, W. D. Pace, K. A. Siek, W. H. Ward, A platform for human-robot dialog systems research, in: 2010 AAAI Fall Symposium Series, 2010.
[38] M. Tenorth, M. Beetz, Representations for robot knowledge in the knowrob framework, Artificial Intelligence 247 (2017) 151–169.
[39] C. Schlenoff, E. Prestes, R. Madhavan, P. Goncalves, H. Li, S. Balakirsky, T. Kramer, E. Miguelanez, An IEEE standard ontology for robotics and automation, in: IEEE Int. Conf. on Intelligent Robots and Systems (IROS), 2012, pp. 1337–1342.
[40] E. Prestes, J. L. Carbonera, S. R. Fiorini, V. A. M. Jorge, M. Abel, R. Madhavan, A. Locoro, P. Goncalves, M. E. Barreto, M. Habib, A. Chibani, S. Gérard, Y. Amirat, C. Schlenoff, Towards a core ontology for robotics and automation, Robotics and Autonomous Systems 61 (2013) 1193 – 1204. Ubiquitous Robotics.
[41] S. R. Fiorini, J. L. Carbonera, P. Gonçalves, V. A. Jorge, V. F. Rey, T. Haidegger, M. Abel, S. A. Redfield, S. Balakirsky, V. Ragavan, H. Li, C. Schlenoff, E. Prestes, Extensions to the core ontology for robotics and automation, Robot. Comput.-Integr. Manuf. 33 (2015) 3–11.
[42] A. U. Frank, M. Raubal, Formal specification of image schemata – a step towards interoperability in geographic information systems, Spatial Cognition and Computation 1 (1999) 67–101.
[43] S. Nayak, A. Mukerjee, Concretizing the image schema: How semantics guides the bootstrapping of syntax, in: 2012 IEEE Int. Conf. on Development and Learning and Epigenetic Robotics, ICDL 2012, 2012.