ISD8 Tutorial Report: Cognitively Inspired Reasoning for Reactive Robotics - From Image Schemas to Knowledge Enrichment

Stefano De Giorgis[1,2], Mihai Pomarlan[3] and Nikolaos Tsiogkas[4]

[1] *Institute of Cognitive Sciences and Technologies - National Research Council (ISTC-CNR), Catania, Italy*
[2] *Faculty of Languages, University of Bologna, Bologna, Italy*
[3] *Applied Linguistics Department - University of Bremen, Bremen, Germany*
[4] *Department of Computer Science - KU Leuven, Leuven, Belgium*

Abstract

This report details a tutorial on image-schematic analysis and functional object detection presented at the conference. The tutorial bridges theoretical foundations with practical applications, introducing participants to a novel framework for analyzing video content through the lens of image schemas. Through hands-on demonstrations, participants engaged with an interactive system that processes video sequences to generate rich semantic representations. The system produces multi-layered output including knowledge graphs of image schematic activations, natural language descriptions, temporal event annotations, and automatically curated storyboards with generated captions. The theoretical underpinnings of image-schematic segmentation and functional parts detection were covered, but the tutorial focused on enabling participants to effectively utilize these tools for their own research and applications. The tutorial successfully demonstrated how image-schematic analysis can be practically applied to video understanding tasks, while gathering insights from participants about desired future functionalities. This approach of combining theoretical instruction with hands-on experimentation proved effective for both teaching the technology and collecting user requirements for future development.

Keywords

Image Schemas, Cognitive Robotics, Commonsense Knowledge, Neuro-symbolic AI,

1. Introduction

The tutorial provided an in-depth exploration of a cognitively inspired approach to robotic perception and reasoning based on image schemas. At its core, image schemas represent fundamental spatiotemporal patterns that emerge from embodied experiences and interactions with the physical world [1, 2, 3, 4, 5, 6, 7]. These schemas, such as LINKAGE, SUPPORT, and SOURCE-PATH-GOAL relationships, serve as the building blocks for understanding how objects and agents interact in space and time [4]. The concept originates from cognitive science research showing how humans develop these basic patterns through early sensorimotor experiences, which then scaffold more complex reasoning about physical interactions.

The approach presented in the tutorial addresses a significant challenge in contemporary artificial intelligence. While modern AI systems, particularly large language models and video generators, can produce impressively coherent outputs, they seem to fundamentally lack an understanding of physical dynamics and spatial relationships. This limitation stems from their reliance on statistical patterns learned from decontextualized data, rather than grounded, embodied knowledge of how the physical world works. The tutorial demonstrated how incorporating image schemas into artificial systems can help bridge this gap. The presenters introduced a neurosymbolic architecture that combines neural components for perception with symbolic reasoning based on image schemas. This hybrid approach allows the system to process raw sensory input while maintaining structured representations of spatial

The Eight Image Schema Day (ISD8), 25–28 November 2024, Bozen-Bolzano, Italy
stefano.degiorgis@cnr.it (S. De Giorgis)
https://stendoipanni.github.io/ (S. De Giorgis)
0000 0003 4133 3445 (S. De Giorgis)

relationships and physical interactions. Moreover, the architecture includes mechanisms for knowledge enrichment, enabling the system to learn new functional concepts through observation and interaction.

2. Image Schematic Reasoning with KHAFRE

In this section we will describe Khafre (Knowledge-driven Heideggerian AFfordance Recognition), a neurosymbolic system to perform taskable perception. For this tutorial, we focused on how to use khafre to perform image-schematic event segmentation and recognition of functional object parts from video. Khafre is available under MIT license from https://github.com/heideggerian-ai-v5/khafre.

2.1. Perception Module

The perception module forms the foundation of the system, processing raw sensor data from RGB and, optionally, depth cameras to produce qualitative descriptions of scenes. At its heart lies a sophisticated object detection system based on YOLOv8 models, which segments the visual input into meaningful object regions. This initial segmentation provides the basic vocabulary of objects that the system can reason about.

Beyond simple object detection, the perception module incorporates specialized components for analyzing interactions between objects. A dedicated contact region detector identifies areas where objects meet or touch, providing crucial information for understanding physical relationships. The system also employs optical flow analysis to track movement, enabling it to understand dynamic interactions as they unfold over time. Shape registration applied to object segmentation masks is used to estimate occluded parts of objects as well as track parts of objects that have become unoccluded, which is important to recognize events that require an object to pass through or behind another, e.g. a knife blade through an apple during cutting.

The module operates on a query-based paradigm, where it actively seeks specific types of information rather than attempting to process everything at once. These queries focus on four main aspects: object identification, detection of relative movement between objects, identification of contact relationships, identification of (former) occlusions. This targeted approach allows the system to efficiently allocate computational resources to the most relevant aspects of the scene. The output of the perception module takes the form of qualitative descriptions expressed as semantic triples. For example, the system might generate assertions like ”object A contacts object B” or ”object C moves relative to object D.” These qualitative descriptions are complemented by contact mask annotations, which highlight specific regions where objects interact. This combination of symbolic descriptions and spatial annotations provides a rich foundation for higher-level reasoning.

2.2. Reasoning Module

The reasoning component builds upon the perceptual foundation by maintaining and updating beliefs about the current situation. These beliefs are primarily expressed through image schema assertions, which capture meaningful relationships between objects such as CONTACT and SUPPORT. The reasoning system employs a sophisticated rule-based inference mechanism to update these beliefs based on new perceptual information. A key feature of the reasoning module is its use of reification mechanisms. When the system identifies a relationship between objects, it creates a new entity representing that relationship. This approach allows the system to reason about relationships themselves as objects, enabling more complex forms of inference and knowledge representation. The module also incorporates graph-based querying capabilities, allowing it to analyze dependencies between entities and relationships. This feature is particularly important for understanding how different spatial relationships interact and influence each other. For example, the system can track how a SUPPORT relationship between two objects depends on specific Contact relationships being maintained.

2.3. Image Schematic Event Segmentation

By image schematic event segmentation we mean selecting frames of a video at which image schematic relations between objects change – i.e., some relation comes into effect, or ceases to hold. An example segmentation can be seen in figure 2.3. The selected frames will typically not have the same time interval between them.

Such "event frames" are usually much fewer in number than the whole frames of the video, and they represent times at which something "significant" changes. What significant means depends on the goals that the perception system was given. In the case of the segmentation displayed in figure 2.3, the tracked events were contacts, occlusions, and penetrations involving fruit and cutlery.

Khafre saves the selected event frames as pairs of files, one in JPEG format to store the image and one in Turtle format to store its semantic description, and also creates, after a video analysis is complete, an html file in which the segmentation as a whole can be viewed.

2.4. Concept Invention

One of the most innovative aspects presented in the tutorial is the system's ability to learn new functional concepts through observation. This capability addresses a fundamental limitation of traditional robotic systems, which typically operate with a fixed ontology of pre-defined object categories. Instead, this system can identify and learn about functional parts based on their roles in observed interactions. This approach is a refinement of previous approaches as in [8, 9, 10]. The process begins when the system observes objects interacting in specific ways. As an example which we used for a previous paper [11], when observing a mug being supported by a hook, the system identifies not just the objects involved, but the specific parts that enable the interaction. Through repeated observations, it builds up a concept of what makes a part functional for a particular type of interaction. The system formalizes these observations by creating new concept definitions. For instance, it might create a concept like "MugSupportedByHook" for parts of mugs that can engage in support relationships with hooks. These concepts are grounded in both geometric properties (the shape and location of the part) and functional properties (its role in supporting relationships).

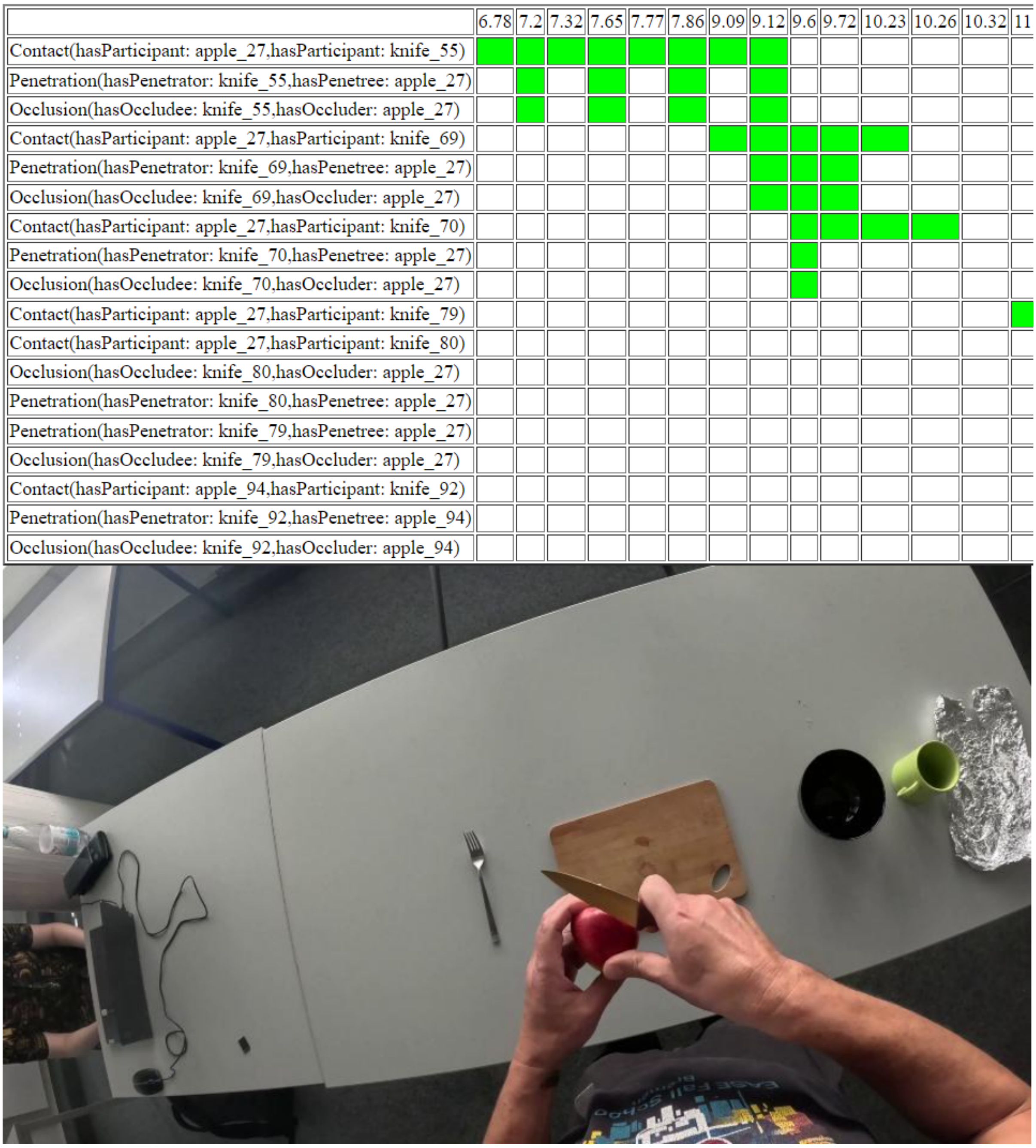
Another example is of functional parts involved in penetrating objects that are the patients of cutting tasks. In this case, the functional part disappears from view – is occluded – while it performs the task, and the system has to guess the occluded shape, as well as recognize when the occluded part re-emerges into view. This is done via shape registration: the shapes of the masks for the same object at consecutive frames are matched to each other, with areas where no overlap occurs being then subject to further processing.

The tutorial demonstrated how the system collects training examples of these functional parts through automated annotation of observed interactions. This collected data is then used to retrain the perception models, enabling them to recognize these functional parts in new situations. The learned concepts could significantly improve the system's ability to plan and execute interactions. Rather than treating objects as atomic wholes, it can reason about specific functional parts and their roles.

3. Knowledge Enrichment

The knowledge enrichment component of our framework operates through a two-stage process that leverages image schemas to derive deeper semantic understanding from visual and spatial information. The framework described here is a refinement based on [12]. This section details how the system progressively builds richer knowledge representations from initial perceptual data.

First Stage: Spatial and Sensorimotor Knowledge Extraction The initial enrichment stage focuses on identifying and formalizing implicit spatial and sensorimotor relationships present in the scene. The system analyzes visual input alongside an existing RDF knowledge graph to identify instances of fundamental image schemas, including Movement, Source-Path-Goal, Contact, Link, Containment,



	6.78	7.2	7.32	7.65	7.77	7.86	9.09	9.12	9.6	9.72	10.23	10.26	10.32	11
Contact(hasParticipant: apple_27,hasParticipant: knife_55)	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩						
Penetration(hasPenetrator: knife_55,hasPenetree: apple_27)		🟩		🟩		🟩		🟩						
Occlusion(hasOccludee: knife_55,hasOccluder: apple_27)		🟩		🟩		🟩		🟩						
Contact(hasParticipant: apple_27,hasParticipant: knife_69)							🟩	🟩	🟩	🟩	🟩			
Penetration(hasPenetrator: knife_69,hasPenetree: apple_27)								🟩	🟩	🟩				
Occlusion(hasOccludee: knife_69,hasOccluder: apple_27)								🟩	🟩	🟩				
Contact(hasParticipant: apple_27,hasParticipant: knife_70)									🟩	🟩	🟩	🟩		
Penetration(hasPenetrator: knife_70,hasPenetree: apple_27)									🟩					
Occlusion(hasOccludee: knife_70,hasOccluder: apple_27)									🟩					
Contact(hasParticipant: apple_27,hasParticipant: knife_79)														🟩
Contact(hasParticipant: apple_27,hasParticipant: knife_80)														
Occlusion(hasOccludee: knife_80,hasOccluder: apple_27)														
Penetration(hasPenetrator: knife_80,hasPenetree: apple_27)														
Penetration(hasPenetrator: knife_79,hasPenetree: apple_27)														
Occlusion(hasOccludee: knife_79,hasOccluder: apple_27)														
Contact(hasParticipant: apple_94,hasParticipant: knife_92)														
Penetration(hasPenetrator: knife_92,hasPenetree: apple_94)														
Occlusion(hasOccludee: knife_92,hasOccluder: apple_94)														

Figure 1: An image schematic timeline of a video. Rows correspond to schematic relations, columns to times. Cells mark whether a relation is active at that time.

Balance, Center-Periphery, and Blockage. Operating on both image data and the base knowledge graph serialized in Turtle syntax, the system identifies spatial and dynamic relationships that may not be explicitly represented in the initial perception. For instance, when observing an object being lifted, the system not only recognizes the immediate Movement schema but also identifies the implicit Balance relationships that must be maintained and any potential Blockage that might affect the motion. This enrichment process extends the original knowledge graph by adding new triples that capture these image schematic relationships. The system anchors these new assertions to existing entities in the knowledge base, ensuring that the enriched knowledge remains grounded in the concrete objects and situations observed in the environment.

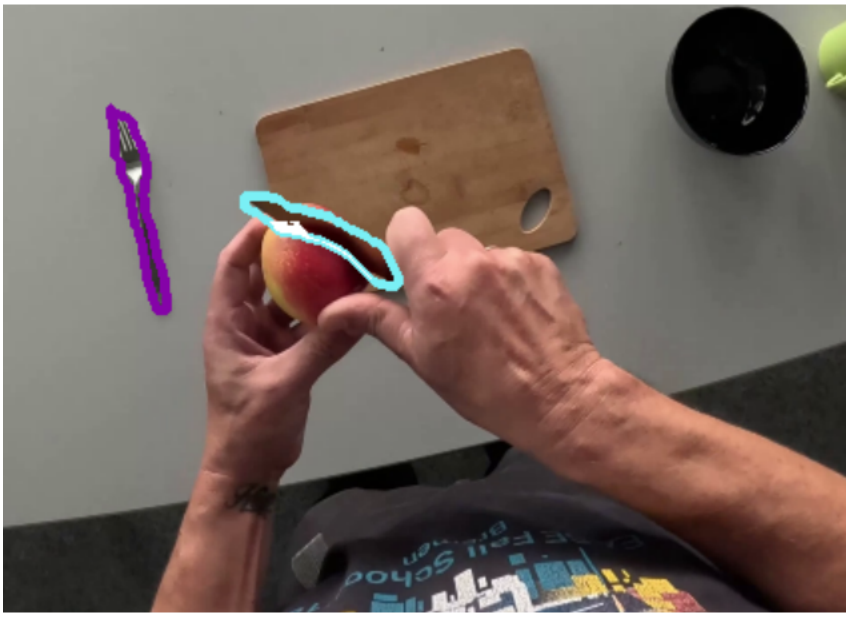


Figure 2: An area of a cutlery item (highlighted in cyan) that was previously occluded by a fruit while in contact with the fruit is now visible (highlighted in white). This is taken to indicate that the highlighted area penetrated the fruit and is functionally important for cutting tasks.

Second Stage: Temporal and Causal Relationship Inference The second stage of knowledge enrichment moves beyond immediate spatial relationships to understand deeper temporal and causal patterns. This stage processes the enriched knowledge graph from the first stage to identify four key types of relationships:

Causal Dependencies: The system identifies when one event directly influences or causes another. For example, when a pushing action causes an object's movement, this causal relationship is explicitly encoded in the knowledge graph.

Event Sequences: Temporal ordering between events is captured, allowing the system to understand not just what happens but the sequence in which events unfold. This temporal knowledge is crucial for understanding process flows and action planning.

Implied Future Events: Based on current observations and known patterns, the system can infer potential future states or events. These predictions are encoded as probabilistic relationships in the knowledge graph.

Prevented Events: The system recognizes and represents potential events that are prevented from occurring due to current conditions or other events. This understanding of "what could have happened" enriches the system's comprehension of situational dynamics.

Integration with the Knowledge Base Both enrichment stages operate within the constraints of RDF and Turtle syntax, ensuring that all derived knowledge integrates seamlessly with the existing knowledge base. Each new relationship is expressed through well-formed RDF triples, maintaining the semantic structure of the knowledge representation. This enriched knowledge serves multiple purposes within the broader system:

1. It enables more sophisticated reasoning about spatial relationships and physical interactions. By making implicit relationships explicit, the system can better understand the consequences of actions and the constraints of physical situations.
2. The temporal and causal knowledge supports better prediction and planning. Understanding event sequences and their causal relationships allows the system to anticipate potential outcomes and plan more effectively.
3. The enriched knowledge base provides a foundation for learning new concepts. By identifying

patterns in how objects participate in various image schemas and causal relationships, the system can develop new functional categories and relationships.

The two-stage enrichment process demonstrates how combining image schematic understanding with temporal and causal reasoning can create a rich semantic representation of physical situations. This enhanced knowledge representation supports more sophisticated reasoning about physical interactions and enables more adaptive behavior in complex environments.

4. Resources

Khafre is available under MIT license at https://github.com/heideggerian-ai-v5/khafre.
The tutorial colab for knowledge enrichment is available at https://colab.research.google.com/drive/1s8BtRvLCNWL2GpHK7y50w2s2WaK6db6g?usp=sharing.

Acknowledgments

This work was supported by the Future Artificial Intelligence Research (FAIR) project, code PE00000013 CUP 53C22003630006, the German Research Foundation DFG, as part of Collaborative Research Center (Sonderforschungsbereich) 1320 Project-ID 329551904 “EASE - Everyday Activity Science and Engineering”, subproject “P01 – Embodied semantics for the language of action and change: Combining analysis, reasoning and simulation”, and by the FET-Open Project #951846 “MUHAI – Meaning and Understanding for Human-centric AI” by the EU Pathfinder and Horizon 2020 Program.

References

[1] M. Johnson, The Body in the Mind Metaphors, University of Chicago Press, 1987.

[2] G. Lakoff, M. Johnson, Metaphors we live by, University of Chicago press, 1980.

[3] G. Lakoff, M. Johnson, Metaphors we Live by, University of Chicago Press, Chicago, 1980.

[4] J. M. Mandler, How to build a baby: Ii. conceptual primitives., Psychological review 99 (1992) 587.

[5] J. M. Mandler, The Foundations of Mind: Origins of Conceptual Thought: Origins of Conceptual Though, Oxford University Press, New York, 2004.

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

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

[8] M. M. Hedblom, M. Pomarlan, R. Porzel, R. Malaka, M. Beetz, Dynamic action selection using image schema-based reasoning for robots, in: The 7th Joint Ontology Workshops (JOWO), Bolzano, Italy, 2021.

[9] M. Pomarlan, M. M. Hedblom, 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.

[10] M. Pomarlan, M. M. Hedblom, R. Porzel, Curiously exploring affordance spaces of a pouring task, Expert Systems 40 (2023) e13213.

[11] M. Pomarlan, S. De Giorgis, R. Ringe, M. M. Hedblom, N. Tsiogkas, Hanging around: Cognitive inspired reasoning for reactive robotics, in: Formal Ontology in Information Systems, IOS Press, 2024, pp. 2–15.

[12] S. De Giorgis, A. Gangemi, A. Russo, Neurosymbolic graph enrichment for grounded world models, arXiv preprint arXiv:2411.12671 (2024).