Constructing Grounded Conceptual Representations via Object-Centric Active Vision

Miroslav Cibula, Kristína Malinovská, Igor Farkaš

Faculty of Mathematics, Physics and Informatics
Comenius University Bratislava
cibula25@uniba.sk, kristina.malinovska@fmph.uniba.sk, igor.farkas@fmph.uniba.sk

Abstract

Humans perform reasoning and other high-level conscious cognitive operations predominantly on symbolic representations. We tend to represent entities and events in the world using mostly discrete concepts and attributes, and we can efficiently convert these representations across different levels of abstraction. The raw perceptual input, however, is mainly sub-symbolic, and to achieve a higher level of conceptual understanding, a cascade of cognitive operations must be performed on it to translate it effectively. The present work aims to explore how visual perceptual information might be computationally transformed into symbolic representations that support concept formation and basic reasoning. We propose a preliminary active computer vision system that processes a 3D video stream to segment scenes, track objects, and infer simple object properties and affordances through interaction. Via this process, we construct object-centric, perceptually grounded conceptual representations within the framework of conceptual spaces, embedding observed object attributes and motion behaviors into a structured semantic space. This approach aims to draw on cognitively inspired and biologically plausible principles, prioritizing modularity and interpretability over state-of-the-art fully monolithic, end-to-end neural architectures. Consequently, this perceptual architecture could complement other reasoning and learning systems in the future, resulting in a full robust cognitive architecture applicable in real world.

1 Introduction and Related Work

Visual perceptual input is an essential modality for early sensorimotor discovery and learning in human infants (Gopnik et al., 1999). Especially during preverbal developmental stages, visual perception, coupled with the ability to physically interact with the environment, facilitates the construction of multimodally grounded semantic representations (Fitzpatrick et al., 2003; Lohmann et al., 2020). The semantics of concepts in this stage are not learned via language (Mandler, 1992; Carey, 2009), but rather by observing the attributes (i.e., in the case of concrete physical objects, these attributes could be related to appearance, physical properties, behavioral patterns under some physical intervention by the agent, etc.) and characterizing the concepts using the sets of such attributes (Rosch, 1978; Gärdenfors, 2000; Takáč, 2007; Baillargeon, 2004; Held and Hein, 1963).

The theory of conceptual spaces (Gärdenfors, 2000, 2014) is one of the well-established computational frameworks implementing such semantic mechanisms with a higher degree of biological plausibility. Within it, formally, observed instances of concepts, exemplars (Medin and Schaffer, 1978; Nosofsky, 1984), are aligned with other exemplars in the semantic space based on the similarity of their attributes. Consequently, the observed exemplars tend to cluster along the semantic space dimensions corresponding to their common attributes (i.e., quality dimensions), forming convex regions that represent conceptual categories, with the cluster centroids approximating prototypical instances (Rosch, 1978). As a result, conceptual spaces are formally hyperdimensional geometric spaces organized along interpretable and separable quality dimensions, supporting semantic similarity assessment by measuring spatial distances between exemplars or categories, as well as categorization and category recall, since any newly observed exemplar can be mapped to the appropriate convex region.

Hence, such a combination of a computational visual processing pipeline with conceptual spaces for semantic representations and an embodied agent leveraging physical manipulation to elicit new environmental observations may yield an approximate computational model of a bottom-up visual perception process (McMains and Kastner, 2011), specifically producing non-verbal physical-object-centric concepts grounded in perception and action (Barsalou, 1999). In the context of cognitive robotics, this process is essential for constructing high-level symbolic representations of the environment, facilitating higher-level reasoning and planning.

The integration of robotic interaction with computer vision methods to boost their performance, however, has been explored relatively extensively. Most notably, the research has focused on learning the properties of physical objects in the scene via motion tracking following the physical interaction performed by an

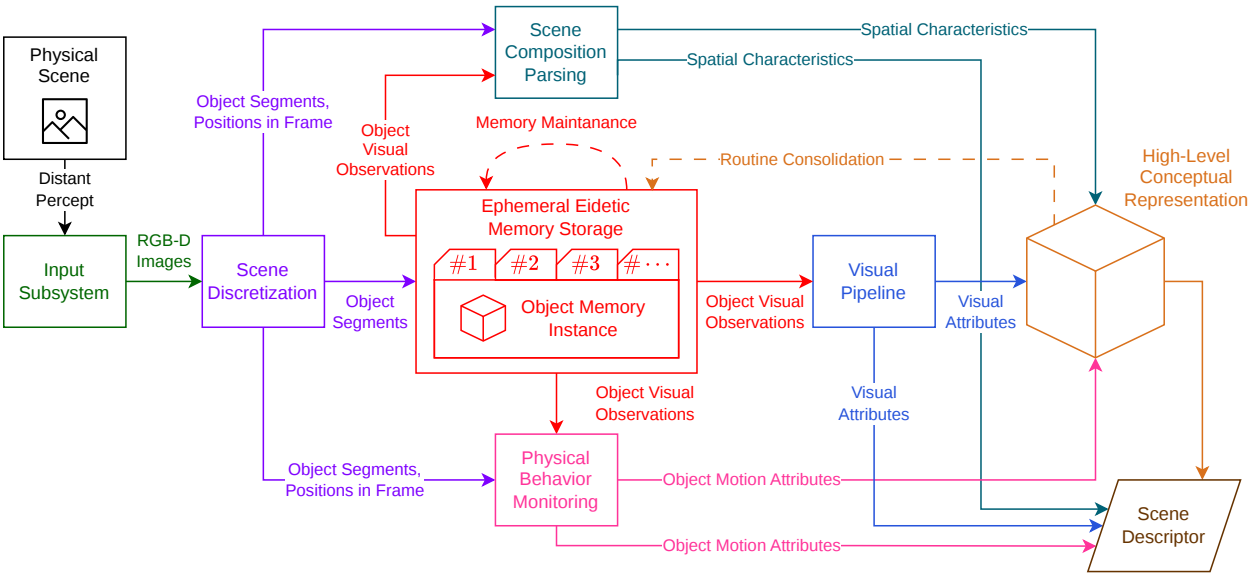

Fig. 1: High-level schematic of the proposed perceptual architecture with the essential subsystems facilitating the conversion of a low-level visual percept from a physical scene to a high-level scene description while learning high-level conceptual representations modeled as a conceptual space.

embodied agent (Fitzpatrick et al., 2003; Fitzpatrick, 2003; Katz et al., 2013; Nematollahi et al., 2020; Kruzliak et al., 2024). Additionally, there has been a focus on learning manipulation strategies to explore the environment more efficiently (Lohmann et al., 2020; Gadre et al., 2021), or on linking robotic actions to object perception to construct affordance representations (Gupta and Davis, 2007; Gadre et al., 2021).

The present paper proposes an artificial perceptual architecture (Section 2) that leverages active computer vision, enabling robotic agents to study and learn the attributes and behaviors of physical objects and to generalize across multiple observations to form representations of discrete conceptual symbols. In our design, we rely on standard computer vision and machine learning methods to facilitate this kind of visual processing and to assemble them into a modular system that functionally resembles pertinent parts of the human visual system. Thus, the proposed architecture is transparent, explainable, and biologically plausible. Subsequently, we suggest how the architecture and the representations it yields can give rise to more complex cognitive functions, such as language processing and primitive social cognition (Section 3). Our design methodology is hence in accordance with the developmental perspective of human cognition: a child first learns foundational, object-centric representations through sensorimotor exploration of the physical world (Mandler, 2007), subsequently grounds early linguistic labels in these pre-existing perceptual concepts (Tomasello, 2003; Bloom, 2000), and eventually leverages this acquired linguistic system as a cognitive tool to bootstrap the formation of increasingly abstract and complex representations (Vygotskij, 1978; Lupyan, 2012).

2 Proposed Architecture

The minimal computational high-level formulation of the proposed perceptual architecture could be defined as a stateful ϑ-parametrized function $\Pi_\vartheta : \boldsymbol{v}(t) \mapsto \boldsymbol{o}(t)$, where $\boldsymbol{v}(t) \in [0.0; 1.0]^{H \times W \times 4 \times c}$ denotes a visual percept/experience at the discrete environmental time $t \in \mathbb{N}$, in our context defined as a sequence of $1 \leq c \leq t$ normalized RGB-D images, and $\boldsymbol{o}(t)$ being an abstract set of high-level disentangled state variables [akin to the ones in our previous work (Cibula et al., 2024)] extracted from the visual percept $\boldsymbol{v}(t)$.

The parametrization ϑ affects the system's subprocesses by configuring the hyperparameters of various modules within the architecture, with specifics depending on the system implementation. In our conceptualization, ϑ enables top-down influence from the rest of the architecture, thereby facilitating bio-plausible top-down/bottom-up bidirectionality in the perceptual process (Dijkstra et al., 2017; McMains and Kastner, 2011). More specifically, ϑ can theoretically serve two purposes: facilitating longer-term perceptual learning and adjusting the perceptual system based on feedback from a potential reasoning component, and providing short-term influence, for example, in attentional filtering (Deutsch and Deutsch, 1963; Cowan, 1998).

In designing the internal mechanisms of Π_ϑ (Figure 1), we explicitly avoid monolithic, end-to-end neural architectures. Such models tend to be opaque and resistant to integration into larger, modular cognitive architectures in which specific perceptual domains must be isolated and manipulated. Instead, the proposed framework leverages a decoupled, multi-branch pipeline utilizing specialized computer vision methods.

This design functionally resembles the human visual system, partitioning continuous visual perception into discrete object representations, kinematic tracking, and spatial topology extraction.

2.1 Scene Discretization and Visual Memory

As raw perceptual input is predominantly sub-symbolic, a bottom-up quantization mechanism is required to convert the continuous RGB-D signal into discrete entities. In human infants, this segmentation relies heavily on motion and physical cohesion (Spelke, 1990). Computationally, we propose to operationalize this using standard prompt-free instance segmentation models coupled with multi-object tracking.

To maintain object permanence during occlusions and manipulation, the system employs Ephemeral Eidetic Memory Storage (EEMS). As objects are tracked, the EEMS captures geometric and visual crops of the objects from various viewpoints, provided the observations meet heuristic quality thresholds, such as low visual similarity to already recorded observations or high image sharpness. EEMS serves as a functional analog to short- to mid-term eidetic visual memory, accumulating a view-invariant 3D understanding of the physical entities in the scene.

2.2 Feature Extraction

Subsequently to the scene discretization into tracked entities, the architecture processes the EEMS data through three parallel branches, extracting representations that map to fundamental domains of human object cognition: intrinsic visual appearance (the "what"), spatial topology (the "where"), and dynamic behavior (the "how"). In this categorization, we take inspiration from the two-stream processing hypothesis (Goodale and Milner, 1992), and the theory of affordances (Gibson, 2014). By isolating these modalities, the system generates independent representations ready for conceptual embedding.

Visual Domain To capture the "what" attributes of an object (e.g., color, shape, texture), the system should avoid language-biased models in favor of unsupervised representation learning. Relatively recently, a new class of slot-attention-based models (Locatello et al., 2020) has emerged that can disentangle raw pixel inputs into distinct, continuous latent factors (Stammer et al., 2024). When applied to sets of visual observations for each object stored in the EEMS, the architecture mirrors human visual decomposition, isolating object visual properties into independent feature dimensions without requiring linguistic supervision.

Spatial Domain To represent the spatial organization of the scene, spatial relationships between objects should be formalized using purely geometric metrics, echoing developmental evidence that preverbal infants understand primitives like spatial containment and support before acquiring spatial vocabulary (Mandler, 1992; Baillargeon, 2004). For each object, an egocentric spatial signature could be constructed by evaluating relative 3D directional vectors (i.e., azimuth, elevation, distance) and volumetric overlaps with surrounding objects. This approach would embed variable-sized scene graphs into fixed-length relational vectors, forming the embeddable spatial domain.

Kinematic Domain Finally, to characterize how objects behave and interact dynamically, the system should track 6-degree-of-freedom (6-DoF) trajectories of their movement. This could be achieved by any standard 6D object-pose estimator tracking each object during the recording session. Such a tracking process yields a 6D time-series trajectory that should be converted into discrete motion symbols. Acting as a signal, each trajectory can be filtered and transformed into the frequency domain (Mao et al., 2020). By retaining only the low-frequency coefficients, we can generate a compact, fixed-length vector that captures the prototypical "shape" of the motion (e.g., tumbling, sliding) and is invariant to the event's specific speed or duration.

2.3 Semantic Representation Construction

The objective of the perceptual architecture would be to construct a structured semantic representation utilizing the framework of conceptual spaces (Gärdenfors, 2000).

The vectors extracted from the three processing branches (Section 2.2) would act as entities within their respective, separable quality dimensions. Each observation of an object could function as an exemplar. As the embodied agent interacts with the environment, exemplars sharing common attributes would naturally cluster together. The centroids of these clusters would emerge as the prototypical instances of concepts.

By measuring semantic similarity via spatial distances, the system could dynamically categorize novel objects by mapping them to the nearest established convex region. This architecture would effectively ground abstract symbolic meaning in multimodal perception and physical action (Barsalou, 1999; Harnad, 1990), providing a foundation for higher-level reasoning independent of prior linguistic knowledge.

Concurrently, this categorization process yields the immediate high-level scene descriptor $\boldsymbol{o}(t)$. At any discrete time step t, $\boldsymbol{o}(t)$ serves as an instantaneous, disentangled state representation of the environment, encapsulating the tracked entities, their current 6-DoF kinematics, spatial relationships, and their mapped conceptual categories. Consequently, the perceptual process yields a dual product: it progressively builds a long-term, generalized semantic space while simultaneously producing a structured, real-time descriptor ready to be leveraged by downstream action-selection or reasoning modules.

3 Cognitive Corollaries

The proposed architecture natively affords extensions into higher-order cognitive capabilities without requiring fundamental structural changes. By producing a geometrically grounded semantic space, we hypothesize that the system provides a direct substrate for language acquisition and primitive social cognition.

Language Emergence Following the framework of conceptual spaces, the proposed system establishes meaning geometrically prior to any linguistic involvement. Within this paradigm, language can emerge simply as a system of pointers that guide attention to specific convex regions or centroids in conceptual space (Gärdenfors, 2024). This allows for a natural mapping between mathematical clusters and grammatical categories: visual attribute clusters may correspond to adjectives (e.g., "yellow", "cubic"), kinematic clusters may correspond to verbs (e.g., "slides", "falls"), and spatial clusters may correspond to prepositions (e.g., "behind", "on"). Consequently, attaching linguistic labels to these autonomously formed categories would yield a compositional language inherently grounded in sensorimotor perception, thereby addressing the symbol grounding problem (Harnad, 1990; Vogt, 2002).

Social Cognition Furthermore, the object-centric nature of the architecture offers an implicit pathway toward primitive social cognition. By treating other agents as physical systems, the discretization module could segment a person into constituent body parts, effectively tracking each part as an independent "object". Consequently, the system would implicitly perform human body pose estimation, extracting the behavioral kinematics and spatial topology of the observed agent. If the robotic agent has a model of its own body, it could measure semantic similarities between the tracked trajectories of the observed agent's limbs and its own end-effectors. This mapping between the self and the other forms the foundational prerequisite for observational imitation learning and intent recognition.

4 Conclusion

Here we propose a theoretical framework for a modular active computer vision architecture designed to transform continuous sub-symbolic video streams into discrete, high-level semantic representations. By explicitly avoiding opaque, end-to-end monolithic models, we outline a transparent, bio-plausible pipeline that partitions visual processing into visual, spatial, and kinematic domains, mimicking human visual perception. By embedding these extracted features, we demonstrate how the system could dynamically construct a conceptual space that reconciles exemplar-based observations with prototype-based category formation through geometric proximity, without linguistic influence.

Ultimately, this perceptual architecture is designed to serve as a non-verbal sensory module for embodied agents. By outputting real-time structural scene descriptors and progressively learned conceptual representations, the proposed system provides the necessary grounding for downstream cognitive architectures, facilitating sensorimotor learning and inference, and eventually, enabling language acquisition in complex real-world environments.

Acknowledgement

The authors thank Markus Vincze (Automation and Control Institute, TU Wien) for his advice and feedback. This research was supported by project APVV-21-0105. Research results were in part obtained using the computational resources of the supercomputer PERUN at the Supercomputing Center at TU Košice, with support from the EU, funds of the Recovery and Resilience Plan of the Slovak Republic, project 17I03-04-P03-00001.

References

Baillargeon, R. (2004). Infants' physical world. *Current Directions in Psychological Science*, 13(3):89–94.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4):577–660.

Bloom, P. (2000). *How Children Learn the Meanings of Words*. Learning, Development, and Conceptual Change. The MIT Press, Cambridge, MA, US.

Carey, S. (2009). *The Origin of Concepts*. Oxford University Press.

Cibula, M., Kerzel, M., and Farkaš, I. (2024). Learning low-level causal relations using a simulated robotic arm. In *Artificial Neural Networks and Machine Learning*, pages 285–298. Springer Nature.

Cowan, N. (1998). Attentional filtering and orienting. In *Attention and Memory: An Integrated Framework*, pages 137–166. Oxford University Press.

Deutsch, J. A. and Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological Review*, 70(1):80–90.

Dijkstra, N., Zeidman, P., Ondobaka, S., van Gerven, M. A. J., and Friston, K. (2017). Distinct top-down and bottom-up brain connectivity during visual perception and imagery. *Scientific Reports*, 7(1).

Fitzpatrick, P. (2003). First contact: An active vision approach to segmentation. In *Int. Conf. on Intelligent Robots and Systems*, pages 2161–2166. IEEE.

Fitzpatrick, P., Metta, G., Natale, L., Rao, S., and Sandini, G. (2003). Learning about objects through action - initial steps towards artificial cognition. In

Int. Conf. on Robotics and Automation, pages 3140–3145. IEEE.

Gadre, S. Y., Ehsani, K., and Song, S. (2021). Act the part: Learning interaction strategies for articulated object part discovery. In *Int. Conf. on Computer Vision*, pages 15732–15741. IEEE.

Gibson, J. J. (2014). *The Ecological Approach to Visual Perception: Classic Edition*. Psychology Press, NY.

Goodale, M. A. and Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25.

Gopnik, A., Meltzoff, A., and Kuhl, P. (1999). *The Scientist in the Crib*. William Morrow and Company, Inc., New York, first Perennial edition.

Gupta, A. and Davis, L. S. (2007). Objects in action: An approach for combining action understanding and object perception. In *Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.

Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. MIT Press, Cambridge, Mass.

Gärdenfors, P. (2014). *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. MIT Press.

Gärdenfors, P. (2024). Event structure, force dynamics and verb semantics. *Language Sciences*, 102:101610.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42:335–346.

Held, R. and Hein, A. (1963). Movement-produced stimulation in the development of visually guided behavior. *Journal of Comparative and Physiological Psychology*, 56(5):872–876.

Katz, D., Kazemi, M., Andrew Bagnell, J., and Stentz, A. (2013). Interactive segmentation, tracking, and kinematic modeling of unknown 3D articulated objects. In *Int. Conf. on Robotics and Automation*, pages 5003–5010. IEEE.

Kruzliak, A. et al. (2024). Interactive learning of physical object properties through robot manipulation and database of object measurements. In *Int. Conf. on Intelligent Robots and Systems*, pages 7596–7603. IEEE.

Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., and Kipf, T. (2020). Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, volume 33, pages 11525–11538.

Lohmann, M., Salvador, J., Kembhavi, A., and Mottaghi, R. (2020). Learning about objects by learning to interact with them. In *Advances in Neural Information Processing Systems*, pages 3930–3941.

Lupyan, G. (2012). What do words do? Toward a theory of language-augmented thought. In Ross, B. H., editor, *The Psychology of Learning and Motivation*, volume 57, pages 255–297. Academic Press.

Mandler, J. M. (1992). How to build a baby: II. Conceptual primitives. *Psychological Review*, 99(4):587–604.

Mandler, J. M. (2007). *Foundations of Mind: Origins of Conceptual Thought*. Oxford University Press.

Mao, W., Liu, M., and Salzmann, M. (2020). History repeats itself: Human motion prediction via motion attention. In *Computer Vision – ECCV 2020*, pages 474–489, Cham. Springer International Publishing.

McMains, S. and Kastner, S. (2011). Interactions of top-down and bottom-up mechanisms in human visual cortex. *Journal of Neuroscience*, 31(2):587–597.

Medin, D. L. and Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85(3):207–238.

Nematollahi, I., Mees, O., Hermann, L., and Burgard, W. (2020). Hindsight for foresight: Unsupervised structured dynamics models from physical interaction. In *Int. Conf. on Intelligent Robots and Systems*, pages 5319–5326. IEEE.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(1):104–114.

Rosch, E. (1978). Principles of categorization. In Rosch, E. and Lloyd, B. B., editors, *Cognition and Categorization*, pages 27–48. Lawrence Erlbaum, Hillsdale, N.J.

Spelke, E. S. (1990). Principles of object perception. *Cognitive Science*, 14(1):29–56.

Stammer, W., Wüst, A., Steinmann, D., and Kersting, K. (2024). Neural concept binder. In *Advances in Neural Information Processing Systems*, volume 37, pages 71792–71830. Curran Associates, Inc.

Takáč, M. (2007). *Construction of Meanings in Living and Artificial Agents*. PhD thesis, Comenius University Bratislava.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press, Cambridge, MA.

Vogt, P. (2002). The physical symbol grounding problem. *Cognitive Systems Research*, 3(3):429–457.

Vygotskij, L. S. (1978). *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press, Cambridge, Mass.