# Visual Sensemaking Needs Both Vision and Semantics

## On Logic-Based Declarative Neurosymbolism for Reasoning about Space and Motion

Jakob Suchan
German Aerospace Research (DLR)
Germany
jakob.suchan@dlr.de

Mehul Bhatt
Örebro University, Sweden
CoDesign Lab EU
mehul.bhatt@oru.se

Srikrishna Varadarajan
CoDesign Lab EU
Europe
krish@codesign-lab.org

Contemporary artificial vision systems lack abilities for high-level, human-scale mental simulation, e.g., concerning perceived everyday multimodal interactions. Cognitively driven sensemaking functions such as embodied grounding for active vision, visuospatial concept formation, commonsense explanation, diagnostic introspection all remain fertile ground. We posit that developing high-level visual sensemaking capabilities requires a systematic, tight yet modular integration of (neural) visual processing techniques with high-level commonsense knowledge representation and reasoning methods pertaining to space, motion, actions, events, conceptual knowledge etc. As an exemplar of this thinking, we position recent work on deeply semantic, explainable, neurosymbolic visuospatial reasoning driven by an integration of methods in (deep learning based) *vision* and (KR based) *semantics*. The positioned work is general, but its significance is demonstrated and empirically benchmarked in the context of (active, realtime) visual sensemaking for self-driving vehicles.

## Visual Sensemaking: A Human-Centred Cognitive Perspective

*Artificial visual intelligence* [5] denotes the computational capability to semantically process, interpret, and explain diverse forms of visual stimuli (typically) emanating from sensing embodied multimodal interaction of/amongst humans and other artefacts in diverse naturalistic situations of everyday life and profession. Through semantic processing, interpretation, and explanation, alluded here are a wide-spectrum of high-level human-centred *sensemaking* capabilities. These capabilities encompass functions such as visuospatial conception formation, commonsense/qualitative generalisation, hypothetical reasoning, analogical inference, argumentation, event based episodic maintenance & retrieval for perceptual narrativisation, counterfactual reasoning and explanation etc. In essence, in scope are all high-level commonsense visuospatial sensemaking capabilities –be it mundane, analytical, or creative in nature– that humans acquire developmentally or through specialised training, and are routinely adept at performing seamlessly in their everyday life and work.

**The Need for Integrated "Vision and Semantics"**. Computational visual sensemaking requires a systematically developed general and modular integration of high-level techniques concerned with "commonsense and semantics" with low-level neural methods capable of computing primitive features of interest in visual data. Realising computational visual sensemaking —be it realtime or post-hoc— in view of human-centred AI concerns pertaining to explainability, ethics, regulation and requires a systematic (neurosymbolic) integration of Vision and Semantics, i.e., robust commonsense representation & inference about spacetime dynamics on the one hand, and powerful low-level visual computing capabilities, e.g., pertaining to object detection and tracking on the other. It is also critical that the semantics of formal representations, e.g., of space and motion, be rooted to counterparts in natural language [4]. Towards this, the positioned research [11] demonstrates the significance of semantically-driven methods

S. Costantini, E. Pontelli, A. Russo, F. Toni,
R. Calegari, A. D'Avila Garcez, C. Dodaro, F. Fabiano,
S. Gaggl, A. Mileo, (Eds.): ICLP 2023
EPTCS 385, 2023, pp. 393–395, doi:10.4204/EPTCS.385.44

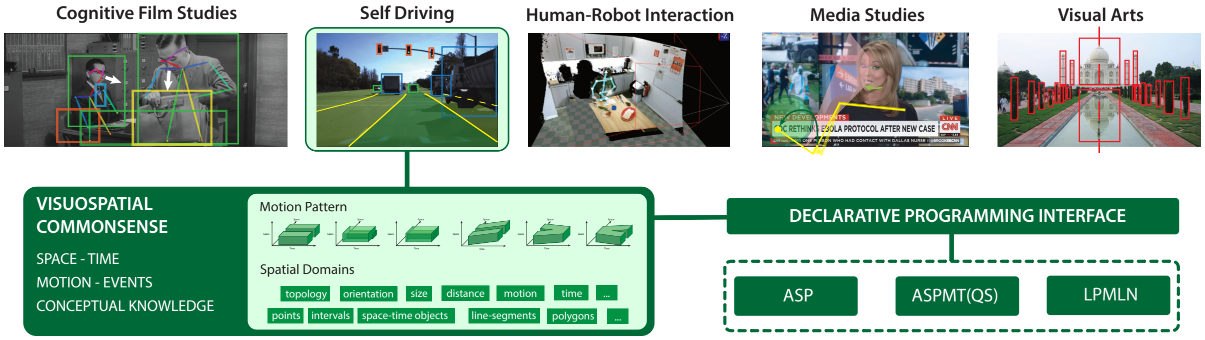

Figure 1: Application Scenarios: Autonomous Systems, Minds and Media Studies, Behavioural Research in Multimodality Interaction [5].

rooted in knowledge representation and reasoning (KR) in addressing research questions pertaining to explainability and human-centred AI particularly from the viewpoint of (perceptual) sensemaking of dynamic visual imagery. This goal is pursued within the larger agenda of *cognitive vision and perception* [5], which is an emerging line of research bringing together a novel & unique combination of methodologies from Artificial Intelligence, Vision and Machine Learning, Cognitive Science and Psychology, Visual Perception, and Spatial Cognition and Computation (also, [9, 10, 11, 12], and [7]). Research in cognitive vision and perception addresses visual, visuospatial and visuo-locomotive perception and interaction from the viewpoints of language, logic, spatial cognition and artificial intelligence [4].

## Visuospatial Commonsense, Driven by Answer Set Programming

Answer Set Programming (ASP) is now widely used as a foundational declarative language and robust methodology for a range of (non-monotonic) knowledge representation and reasoning tasks [6, 8]. With ASP as a foundation, and motivated by requirements of semantics, commonsense and explainability, our work bridges the gap between high-level formalisms for logical visual reasoning (e.g., by abduction) and low-level visual processing[1] by tightly integrating semantic abstractions of *space and motion* with their underlying numerical representations within the declarative framework of ASP. Furthermore, directly within ASP, we also address several challenges concerning *epistemological* and *phenomenological* aspects relevant to a wide range of *dynamic spatial systems* [1, 2, 3]: projection and interpolation, object identity maintenance at a semantic level, ability to make default assumptions, maintaining consistent beliefs etc.

Although the work positioned herein [10, 11] selectively focusses on the needs and challenges of active / online sensemaking in autonomous driving, the generality and modularly of the developed ASP-based neurosymbolic framework ensures foundational applicability in diverse applied contexts requiring perceptual interpretation, interaction and control functions. Of at least equal importance are the modularity and elaboration tolerance of the framework, enabling seamless integration and experimentation with advances in (fast evolving) computer vision methods, as well as experimenting with different forms of formal methods for (declarative) *reasoning about space, actions, and change* [1].

[1] Visual processing encompasses capabilities including but not limited to motion analysis, object detection, pose estimation, semantic segmentation, image classification. A detailed review is available in [5].

## References

[1] Mehul Bhatt (2012): *Reasoning about Space, Actions and Change: A Paradigm for Applications of Spatial Reasoning*. In: *Qualitative Spatial Representation and Reasoning: Trends and Future Directions*, IGI Global, USA, doi:10.4018/978-1-4666-4607-0.ch016.

[2] Mehul Bhatt, Hans W. Guesgen, Stefan Wölfl & Shyamanta M. Hazarika (2011): *Qualitative Spatial and Temporal Reasoning: Emerging Applications, Trends, and Directions*. *Spatial Cognition & Computation* 11(1), pp. 1–14, doi:10.1080/13875868.2010.548568.

[3] Mehul Bhatt & Seng W. Loke (2008): *Modelling Dynamic Spatial Systems in the Situation Calculus*. *Spatial Cognition & Computation* 8(1-2), pp. 86–130, doi:10.1080/13875860801926884.

[4] Mehul Bhatt, Carl Schultz & Christian Freksa (2012): *The 'Space' in Spatial Assistance Systems: Conception, Formalisation and Computation*. In Thora Tenbrink, Jan Wiener & Christophe Claramunt, editors: *Representing space in cognition: Interrelations of behavior, language, and formal models. Series: Explorations in Language and Space*, Oxford University Press, doi:10.1093/acprof:oso/9780199679911.003.0009.

[5] Mehul Bhatt & Jakob Suchan (2023): *Artificial Visual Intelligence: Perceptual Commonsense for Human-Centred Cognitive Technologies*, pp. 216–242. Springer International Publishing, Cham, doi:10.1007/978-3-031-24349-3_12.

[6] Gerhard Brewka, Thomas Eiter & Miroslaw Truszczyński (2011): *Answer Set Programming at a Glance*. *Commun. ACM* 54(12), pp. 92–103, doi:10.1145/2043174.2043195.

[7] David Marr (1982): *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., New York, NY, USA.
(Republished 2010, MIT Press, doi: 10.7551/mitpress/9780262514620.001.0001.)

[8] Torsten Schaub & Stefan Woltran (2018): *Special Issue on Answer Set Programming*. *KI* 32(2-3), pp. 101–103, doi:10.1007/s13218-018-0554-8.

[9] Jakob Suchan & Mehul Bhatt (2016): *Semantic Question-Answering with Video and Eye-Tracking Data: AI Foundations for Human Visual Perception Driven Cognitive Film Studies*. In Subbarao Kambhampati, editor: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, IJCAI/AAAI Press, pp. 2633–2639. Available at `http://www.ijcai.org/Abstract/16/374`.

[10] Jakob Suchan, Mehul Bhatt & Srikrishna Varadarajan (2019): *Out of Sight But Not Out of Mind: An Answer Set Programming Based Online Abduction Framework for Visual Sensemaking in Autonomous Driving*. In Sarit Kraus, editor: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, ijcai.org, pp. 1879–1885, doi:10.24963/ijcai.2019/260.

[11] Jakob Suchan, Mehul Bhatt & Srikrishna Varadarajan (2021): *Commonsense visual sensemaking for autonomous driving - On generalised neurosymbolic online abduction integrating vision and semantics*. *Artif. Intell.* 299, p. 103522, doi:10.1016/j.artint.2021.103522.

[12] Jakob Suchan, Mehul Bhatt, Przemyslaw Andrzej Walega & Carl Schultz (2018): *Visual Explanation by High-Level Abduction: On Answer-Set Programming Driven Reasoning About Moving Objects*. In Sheila A. McIlraith & Kilian Q. Weinberger, editors: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, AAAI Press, pp. 1965–1972, doi:10.1609/aaai.v32i1.11569.