Neurosymbolic Visual Commonsense

On Integrated Reasoning and Learning about Space and Motion in Embodied Multimodal Interaction

Mehul Bhatt

School of Science and Technology, Örebro University – Sweden

CoDesign Lab EU (Artificial and Human Intelligence).,
Cognitive Vision and Perception » https://codesign-lab.org/cognitive-vision

Abstract

We present recent and emerging advances in **computational cognitive vision** addressing artificial visual and spatial intelligence at the interface of (spatial) language, (spatial) logic and (spatial) cognition research. With a primary focus on explainable sensemaking of dynamic visuospatial imagery, we highlight the (systematic and modular) integration of methods from knowledge representation and reasoning, computer vision, spatial informatics, and computational cognitive modelling. A key emphasis here is on generalised (declarative) neurosymbolic reasoning & learning about space, motion, actions, and events relevant to embodied multimodal interaction under ecologically valid naturalistic settings in everyday life. Practically, this translates to general-purpose mechanisms for computational visual commonsense encompassing capabilities such as (neurosymbolic) semantic question-answering, relational spatio-temporal learning, visual abduction etc.

The presented work is motivated by and demonstrated in the applied backdrop of areas as diverse as autonomous driving, cognitive robotics, design of digital visuoauditory media, and behavioural visual perception research in cognitive psychology and neuroscience. More broadly, our emerging work is driven by an interdisciplinary research mindset addressing human-centred responsible AI through a methodological confluence of AI, Vision, Psychology, and (human-factors centred) Interaction Design.

Keywords

Cognitive vision, Knowlede representation and reasoning (KR), Machine Learning, Integration of reasoning & learning, Commonsense reasoning, Declarative spatial reasoning, Relational Learning, Computational cognitive modelling, Human-Centred AI, Responsible AI

1. Motivation

Multimodality in embodied interaction is an inherent aspect of human activity, be it in social, professional, or everyday mundane contexts. Next-generation human-centred AI technologies, operating in such contextualised everyday settings, will require an inherent foundational capacity to "**make sense**" of —e.g., perceive, understand, explain, anticipate— everyday, naturalistic interactional multimodality. This would be essential towards successfully achieving technology mediated ("human-in-the-loop") collaborative assistance, as well as ensuring compliance with emerging human-centred ethical and legal requirements, performance benchmarks, and inclusive usability expectations. It is therefore crucial that the foundational building blocks of such next-generation systems be semantically aligned with the descriptive, analytical, and explanatory characteristics and complexity of human task conceptualisation, performance benchmarks, and usability expectations. Against this backdrop, we define **artificial visual intelligence** [1] as:

> » **The computational capability to semantically process and interpret diverse forms of visual stimuli (typically, but not necessarily) emanating from sensing embodied multimodal interactions of / amongst humans and other artefacts in diverse naturalistic situations of everyday life and work.**

Within the scope of artificial visual intelligence are a wide-spectrum of high-level human-centred sensemaking capabilities. These capabilities encompass operational functions such as:

- Visuospatial conception formation, commonsense/qualitative generalisation, analogical inference;
- Hypothetical reasoning, argumentation, explanation, counterfactual reasoning;
- Event based episodic maintenance & retrieval for perceptual narrativisation.

The afore enumeration is by no means exhaustive: in essence, in scope of artificial visual intelligence are diverse high-level **cognitive visuospatial sensemaking** capabilities —be it mundane, analytical, or creative— that humans acquire developmentally or through specialised training, and are routinely adept at performing seamlessly in their everyday life and work (e.g., driving a vehicle, tracking moving objects, navigating a crowded urban environment, engaging in sports, interpreting subtle cues in everyday people-communication from visual / gestural and auditory signals).

Our central focus is on the development of **general**, **domain-independent methods** that may be seamlessly integrated as part of hybrid computational cognitive system, or even within computational cognitive models / cognitive architectures [2]. We also contextualise and demonstrate in the backdrop of applications in autonomous driving, cognitive robotics, visuoauditory media design, and cognitive psychology (e.g. [3, 4, 5, 6], [7, 8]). Through applied case-studies, we provide a systematic model and general methodology showcasing the integration of diverse, multi-faceted AI methods pertaining Knowledge Representation and Reasoning, Computer Vision, Machine Learning, and Visual

International Joint Conference on Artificial Intelligence (IJCAI)., STRL 24: Third International Workshop on Spatio-Temporal Reasoning and Learning (STRL), IJCAI 2024 – 5 August 2024, Jeju, South Korea
mehul.bhatt@oru.se (M. Bhatt)
https://mehulbhatt.org (M. Bhatt)

Perception towards realising practical, human-centred, computational visual intelligence.

2. Neurosymbolic Visual Commonsense: Integrated Reasoning and Learning about Space, Motion, and Inter(A)ction

In the present status quo, our research in (computational) neurosymbolic visual commonsense categorically addresses three key questions:

I. What kind of (relational) abstraction mechanisms are needed to computationally "make-sense" of embodied multimodal interaction ?

II. How can (and why should) abstraction mechanisms (such as in **I**) be founded on behaviourally established cognitive human- factors emanating from naturalistic empirical observation in real-world applied contexts?

III. How to articulate behaviourally established abstraction mechanisms, preferences (etc) as formal declarative models suited for computational modelling aimed at operational"sensemaking" (encompassing capabilities such as abduction, relational learning, counterfactual inference) ?

Present work is particularly aimed at developing general methods for the semantic interpretation of (multimodal) dynamic visuospatial imagery with an emphasis on the ability to neurosymbolically perform abstraction, reasoning, and learning with cognitively rooted structured characterisations of commonsense knowledge pertaining to space and motion. Here, we specifically emphasise:

- General foundational commonsense abstractions of space, time, and motion needed for representation mediated (grounded) reasoning and learning with dynamic visuospatial stimuli (e.g., emanating from multimodal human behavioural signals in modalities such as RGB(D), video, audio, eye-tracking and possibly even bio signals [9]);
- Deep (visuospatial) semantics, entailing systematically formalised declarative (neurosymbolic) reasoning and learning with aspects pertaining to space, space-time, motion, actions & events, spatiolinguistic conceptual knowledge. Here, it is of the essence that an expressive ontology consisting of, for instance, space, time, space-time motion primitives as first-class 'neurosymbolic' objects is accessible within the (declarative) programming paradigm under consideration; and
- Explainable models of computational visuospatial commonsense based on a systematic integration of symbolic/relational methods on the one hand, and neural techniques aimed at low level quantitative (e.g., visual) data processing on the other;

At a higher level of abstraction, **deep (visualspatial) semantics** (or deep semantics for short) entails inherent support for tackling a range of challenges concerning epistemological and phenomenological aspects relevant to dynamic spatial systems [10] where integrated reasoning about action and change [11, 12] is involved:

- **interpolation and projection** of missing information, e.g., what could be hypothesised about missing information (e.g., moments of occlusion [13]); how can this hypothesis support planning an immediate next step?
- object **identity maintenance** at a semantic level, e.g., in the presence of occlusions, missing and noisy quantitative data, error in detection and tracking
- ability to make **default assumptions**, e.g., pertaining to persistence objects and/or object attributes
- maintaining **consistent beliefs** respecting (domain-neutral) commonsense criteria, e.g., related to compositionality & indirect effects, space-time continuity, positional changes resulting from motion
- inferring / computing **counterfactuals** [14], in a manner akin to human cognitive ability to perform mental simulation for purposes of introspection about the past or anticipation of the future, or performing "what-if" reasoning tasks etc

We particularly emphasise the abilities to **abstract, learn, and reason** with cognitively rooted structured characterisations of commonsense knowledge about **space and motion**, encompassing visuospatial question-answering, abduction, and relational learning:

I. Visuospatial Question-Answering. Focus is on a computational framework for semantic-question answering with video and eye-tracking data founded in constraint logic programming; we also demonstrate an application in cognitive film & media studies, where human perception of films vis-a-via cinematographic devices is of interest.
» [4, 6, 7, 8]

II. Visuospatial Abduction. Focus is on a hybrid architecture for systematically computing robust visual explanation(s) encompassing hypothesis formation, belief revision, and default reasoning with video data (for active vision for autonomous driving, as well as for offline processing). The architecture supports visual abduction with space-time histories as native entities, and founded in (functional) answer set programming based spatial reasoning.
» [3, 13, 15][16, 17]

III. Relational Visuospatial Learning. Focus is on a general framework and pipeline for: relational spatio-temporal (inductive) learning with an elaborate ontology supporting a range of space-time features; and generating semantic, (declaratively) explainable interpretation models in a neurosymbolic pipeline demonstrated for the case of analysing visuospatial symmetry in visual art.
» [18][5][19]

Formal semantics and computational models of deep semantics manifest themselves as neurosymbolic spatio-temporal extensions of established declarative AI frameworks such as Constraint Logic Programming (CLP) [20], Inductive Logic Programming (ILP) [21], and Answer Set Programming (ASP) [22]. The more foundational aspects pertaining

declarative spatial reasoning (built on top of CLP, ILP, ASP) independent of its relationship to cognitive vision research may be consulted in [23], [16, 24], [18].

3. Discussion

The vision that drives our scientific methodology is:

> » To shape the nature and character of (machine-based) artificial visual intelligence with respect to human-centred cognitive considerations, demonstrating an exemplar for developing, applying, and disseminating such methods in socio-technologically relevant application areas where:
>
> (**a**) embodied (multimodal) human interaction is inherent;
>
> (**b**) human-in-the-loop collaborative work is of the essence; and
>
> (**c**) normative ethico-legal compliance based on regulatory requirement and human-factors driven inclusive or universal design criteria is to be ensured.

Towards realising this vision, we adopt an interdisciplinary approach –at the confluence of Cognition, AI, Interaction, and Design– which we deem necessary to better appreciate the complexity and spectrum of varied human-centred challenges for the design and (usable) implementation of (explainable) artificial visual intelligence solutions in diverse human-system interaction contexts.

One of the key technical driving forces in our work is that of "**representation mediated multimodal sensemaking**". In essence, we consider (neurosymbolic) representation mediated grounding as being significant in semiotic construction, e.g., enabling high-level meaning-making. This view stems from the long-established value of "grounding" in Artificial Intelligence and related disciplines [25]. Our research advances the theoretical, methodological, and applied understanding of "grounded representation" mediated multimodal sensemaking of embodied human interaction at the interface of spatial language, spatial logic, and spatial cognition. In our view, the significance of this form of (neurosymbolic) grounding must now be reiterated, re-asserted even, in view of recent advances in neural machine learning and the well-recognised "explainability" and "interpretability" requirements from the viewpoint of human-centred AI [26, 27, 28]. We believe that research in knowledge representation and reasoning (KR) has, since its inception, concerned itself with the "hard" problem of semantics, emphasising explainability, formal verification and diagnosis, elaboration tolerance amongst other things. Research in KR, and more broadly in symbolic AI and semantics, and their role and contribution towards large-scale hybrid "human-in-the-loop" intelligence is of even greater significance now than ever before given the tremendous synergistic opportunities afforded by the widely demonstrated power of deep learning driven techniques in computer vision (and beyond). The onus now, we posit, is on KR research to drive itself towards developing methods that can seamlessly integrate (and be "usable") with other kinds of AI methods, be data-centric neural learning techniques, or otherwise.

In this invited position statement, we have attempted to summarise our mindset and ongoing work in the CoDesign Lab towards:

> » Establishing a human-centric foundation and roadmap for the development of neurosymbolically grounded inference about embodied multimodal interaction as identifiable in a range of real-world application contexts.

This summary is not meant to be a comprehensive literature review; this may be obtained through the cited works. For key technical details and to obtain a summary of open directions, we direct interested readers to select publications as follows: a compact starting point may be obtained via the comprehensive summary in [1], or through the shorter/-focussed components in [15, 5, 4, 13, 3]. Longer summaries in the form of (recent) doctoral dissertations are available in [29] and [30, 31].

Acknowledgments

We acknowledge funding by the Swedish Research Council (**VR** - Vetenskapsrådet) - https://www.vr.se, and the Swedish Foundation for Strategic Research (**SSF** – Stiftelsens för Strategisk Forskning) - https://strategiska.se. Previously, this research has been supported by the German Research Foundation (**DFG** – Deutsche Forschungsgemeinschaft) - https://www.dfg.de.

References

[1] M. Bhatt, J. Suchan, Artificial visual intelligence: Perceptual commonsense for human-centred cognitive technologies, in: Human-Centered Artificial Intelligence: Advanced Lectures, Springer-Verlag, Berlin, Heidelberg, 2023, p. 216–242. URL: https://doi.org/10.1007/978-3-031-24349-3_12. doi:10.1007/978-3-031-24349-3_12.

[2] S. Jones, J. Laird, Anticipatory thinking in cognitive architectures with event cognition mechanisms, in: A. Amos-Binks, D. Dannenhauer, R. E. Cardona-Rivera, G. A. Brewer (Eds.), Short Paper Proc. of Workshop on Cognitive Systems for Anticipatory Thinking (COGSAT 2019), AAAI Fall Symp., volume 2558, 2019. URL: http://ceur-ws.org/Vol-2558/short1.pdf.

[3] J. Suchan, M. Bhatt, S. Varadarajan, Commonsense visual sensemaking for autonomous driving - on generalised neurosymbolic online abduction integrating vision and semantics, Artif. Intell. 299 (2021) 103522. URL: https://doi.org/10.1016/j.artint.2021.103522. doi:10.1016/j.artint.2021.103522.

[4] J. Suchan, M. Bhatt, Semantic question-answering with video and eye-tracking data: AI foundations for human visual perception driven cognitive film studies, in: S. Kambhampati (Ed.), Proceedings of the Twenty-Fifth International Joint Conference on Ar-

tificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016, IJCAI/AAAI Press, 2016, pp. 2633–2639. URL: http://www.ijcai.org/Abstract/16/374.

[5] J. Suchan, M. Bhatt, S. Vardarajan, S. A. Amirshahi, S. Yu, Semantic Analysis of (Reflectional) Visual Symmetry: A Human-Centred Computational Model for Declarative Explainability, Advances in Cognitive Systems 6 (2018) 65–84. URL: http://www.cogsys.org/journal.

[6] J. Suchan, M. Bhatt, The geometry of a scene: On deep semantics for visual perception driven cognitive film, studies, in: 2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016, Lake Placid, NY, USA, March 7-10, 2016, IEEE Computer Society, 2016, pp. 1–9. URL: https://doi.org/10.1109/WACV.2016.7477712. doi:10.1109/WACV.2016.7477712.

[7] J. Suchan, M. Bhatt, Deep Semantic Abstractions of Everyday Human Activities: On Commonsense Representations of Human Interactions, in: ROBOT 2017: Third Iberian Robotics Conference, Advances in Intelligent Systems and Computing 693, 2017.

[8] M. Spranger, J. Suchan, M. Bhatt, Robust Natural Language Processing - Combining Reasoning, Cognitive Semantics and Construction Grammar for Spatial Language, in: IJCAI 2016: 25th International Joint Conference on Artificial Intelligence, AAAI Press, 2016.

[9] M. Bhatt, K. Kersting, Semantic interpretation of multimodal human-behaviour data - making sense of events, activities, processes, Künstliche Intell. 31 (2017) 317–320. URL: https://doi.org/10.1007/s13218-017-0511-y. doi:10.1007/S13218-017-0511-Y.

[10] M. Bhatt, S. W. Loke, Modelling dynamic spatial systems in the situation calculus, Spatial Cognition & Computation 8 (2008) 86–130. URL: https://doi.org/10.1080/13875860801926884. doi:10.1080/13875860801926884.

[11] M. Bhatt, H. W. Guesgen, S. Wölfl, S. M. Hazarika, Qualitative spatial and temporal reasoning: Emerging applications, trends, and directions, Spatial Cognition & Computation 11 (2011) 1–14. URL: https://doi.org/10.1080/13875868.2010.548568. doi:10.1080/13875868.2010.548568.

[12] M. Bhatt, 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, 2012.

[13] J. Suchan, M. Bhatt, S. Varadarajan, Out of sight but not out of mind: An answer set programming based online abduction framework for visual sensemaking in autonomous driving, in: S. Kraus (Ed.), Proc. of 25th Intnl. Joint Conference on Artificial Intelligence, IJCAI 2019, 2019, pp. 1879–1885. doi:10.24963/ijcai.2019/260.

[14] R. Byrne, Counterfactual thought, Annual Review of Psychology 67 (2016) 135–157. URL: https://doi.org/10.1146/annurev-psych-122414-033249. doi:10.1146/annurev-psych-122414-033249. pMID: 26393873.

[15] J. Suchan, M. Bhatt, P. A. Walega, C. P. L. Schultz, Visual explanation by high-level abduction: On answer-set programming driven reasoning about moving objects, in: 32nd AAAI Conference on Artificial Intelligence (AAAI-18), USA, AAAI Press, 2018, pp. 1965–1972.

[16] P. A. Walega, M. Bhatt, C. P. L. Schultz, ASPMT(QS): non-monotonic spatial reasoning with answer set programming modulo theories, in: F. Calimeri, G. Ianni, M. Truszczynski (Eds.), Logic Programming and Nonmonotonic Reasoning - 13th International Conference, LPNMR 2015, Lexington, KY, USA, September 27-30, 2015. Proceedings, volume 9345 of *Lecture Notes in Computer Science*, Springer, 2015, pp. 488–501. URL: https://doi.org/10.1007/978-3-319-23264-5_41. doi:10.1007/978-3-319-23264-5_41.

[17] P. A. Walega, C. P. L. Schultz, M. Bhatt, Non-monotonic spatial reasoning with answer set programming modulo theories, Theory Pract. Log. Program. 17 (2017) 205–225. URL: https://doi.org/10.1017/S1471068416000193. doi:10.1017/S1471068416000193.

[18] J. Suchan, M. Bhatt, C. P. L. Schultz, Deeply semantic inductive spatio-temporal learning, in: J. Cussens, A. Russo (Eds.), Proceedings of the 26th International Conference on Inductive Logic Programming (Short papers), London, UK, 2016, volume 1865, CEUR-WS.org, 2016, pp. 73–80.

[19] K. S. R. Dubba, A. G. Cohn, D. C. Hogg, M. Bhatt, F. Dylla, Learning Relational Event Models from Video, J. Artif. Intell. Res. (JAIR) 53 (2015) 41–90. URL: http://dx.doi.org/10.1613/jair.4395. doi:10.1613/jair.4395.

[20] J. Jaffar, M. J. Maher, Constraint logic programming: A survey, The journal of logic programming 19 (1994) 503–581.

[21] S. Muggleton, L. D. Raedt, Inductive logic programming: Theory and methods, Journal of Logic Programming 19 (1994) 629–679.

[22] G. Brewka, T. Eiter, M. Truszczyński, Answer set programming at a glance, Commun. ACM 54 (2011) 92–103. doi:10.1145/2043174.2043195.

[23] M. Bhatt, J. H. Lee, C. P. L. Schultz, CLP(QS): A declarative spatial reasoning framework, in: M. J. Egenhofer, N. A. Giudice, R. Moratz, M. F. Worboys (Eds.), Spatial Information Theory - 10th International Conference, COSIT 2011, Belfast, ME, USA, September 12-16, 2011. Proceedings, volume 6899 of *Lecture Notes in Computer Science*, Springer, 2011, pp. 210–230. URL: https://doi.org/10.1007/978-3-642-23196-4_12. doi:10.1007/978-3-642-23196-4_12.

[24] C. P. L. Schultz, M. Bhatt, J. Suchan, P. A. Walega, Answer Set Programming Modulo Space-Time, in: C. Benzmüller, F. Ricca, X. Parent, D. Roman (Eds.), Rules and Reasoning - Second International Joint Conference, RuleML+RR 2018, Luxembourg, September 18-21, 2018, Proceedings, volume 11092 of *Lecture Notes in Computer Science*, Springer, 2018, pp. 318–326. URL: https://doi.org/10.1007/978-3-319-99906-7_24. doi:10.

1007/978-3-319-99906-7_24.

[25] S. Harnad, The symbol grounding problem, Physica D 42 (1990) 335–346.

[26] AI HLEG, High-level expert group on artificial intelligence: Ethical guidelines for trustworthy ai, 2019. URL: https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf.

[27] EU Commission, Communication: Building trust in human centric artificial intelligence, 2019.

[28] EU Commission, Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

[29] J. Suchan, Declarative Reasoning about Space and Motion in Visual Imagery - Theoretical Foundations and Applications, Ph.D. thesis, Universität Bremen, 2022. URL: https://elib.dlr.de/188919/.

[30] V. Kondyli, Behavioural Principles for the Design of Human-Centred Cognitive Technologies : The Case of Visuo-Locomotive Experience, Ph.D. thesis, Örebro University, School of Science and Technology, 2023.

[31] V. Nair, The Observer Lens: Characterizing Visuospatial Features in Multimodal Interactions, Ph.D. thesis, , School of Informatics, Informatics Research Environment, 2024.

[32] M. Bhatt, J. Suchan, Cognitive vision and perception, in: G. D. Giacomo, A. Catalá, B. Dilkina, M. Milano, S. Barro, A. Bugarín, J. Lang (Eds.), ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020), volume 325 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2020, pp. 2881–2882. URL: https://doi.org/10.3233/FAIA200434. doi:10.3233/FAIA200434.

[33] V. Kondyli, M. Bhatt, D. Levin, J. Suchan, How do drivers mitigate the effects of naturalistic visual complexity? on attentional strategies and their implications under a change blindness protocol, Cognitive Research: Principles and Implications 8 (2023). doi:10.1186/s41235-023-00501-1.

[34] K. S. R. Dubba, M. Bhatt, F. Dylla, D. C. Hogg, A. G. Cohn, Interleaved inductive-abductive reasoning for learning complex event models, in: S. H. Muggleton, A. Tamaddoni-Nezhad, F. A. Lisi (Eds.), Inductive Logic Programming - 21st International Conference, ILP 2011, Windsor Great Park, UK, July 31 - August 3, 2011, Revised Selected Papers, volume 7207 of *Lecture Notes in Computer Science*, Springer, 2011, pp. 113–129. URL: https://doi.org/10.1007/978-3-642-31951-8_14. doi:10.1007/978-3-642-31951-8_14.

[35] M. Bhatt, J. Suchan, C. Schultz, Cognitive interpretation of everyday activities - toward perceptual narrative based visuo-spatial scene interpretation, in: M. A. Finlayson, B. Fisseni, B. Löwe, J. C. Meister (Eds.), 2013 Workshop on Computational Models of Narrative, CMN 2013, August 4-6, 2013, Hamburg, Germany, volume 32 of *OASIcs*, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2013, pp. 24–29. URL: https://doi.org/10.4230/OASIcs.CMN.2013.24. doi:10.4230/OASICS.CMN.2013.24.

[36] C. Lewis, Representation, Inclusion, and Innovation: Multidisciplinary Explorations, Synthesis Lectures on Human-Centered Informatics, Morgan & Claypool Publishers, 2017. URL: https://doi.org/10.2200/S00812ED1V01Y201710HCI038. doi:10.2200/S00812ED1V01Y201710HCI038.