Neurosymbolic Learning: On Generalising Relational Visuospatial and Temporal Structure

Jakob Suchan[1] and Mehul Bhatt[2]

[1] German Aerospace Center (DLR), Germany
[2] Örebro University, Sweden

CoDesign Lab – Cognition. AI. Interaction. Design.
info@codesign-lab.org – https://codesign-lab.org

Abstract. We position ongoing research aimed at developing a general framework for structured spatio-temporal learning from multimodal human behavioural stimuli. The framework and its underlying general, modular methods serve as a model for the application of integrated (neural) visuo-auditory processing and (semantic) relational learning foundations for applications (primarily) in the behavioural sciences.

Keywords: Representation learning and grounding · Relational learning · Cognitive vision and perception · Multimodality · Dynamic Visuospatial Imagery

High-level perceptual sensemaking of multimodal human behavioural stimuli is foundational to diverse cognitive assistive technologies and autonomous perception & interaction systems [3, 4]. Multimodal sensemaking, at a level of descriptive and analytical complexity that matches cognitive human performance and expectations, is also crucial for the development of next-generation AI technologies and artefacts —concerned with agency, assistance and autonomy— where human-centred considerations of personalisation, normative behaviour, explanation, empathy, trust, responsibility are at the core.

Multimodal Learning: A Neurosymbolic Foundation. In this position statement, we summarise aspects of ongoing research in 'Cognitive Vision and Perception' [4] aimed at developing a general framework for structured spatio-temporal learning from multimodal human behavioural stimuli, e.g., consisting of dynamic visuospatial and auditory features. With an emphasis on formalisations of spatio-linguistically rooted relations of *space, time and motion*, the framework is geared towards supporting high-level learning of *deep semantic* relational spatio-temporal structure —by means of inductive generalisation— from low-level stimuli (typically) emanating from embodied human interactions in everyday naturalistic settings. At the crux of the proposed framework are general and modularly developed foundational spatio-temporal learning methods intended to serve as a neurosymbolic model for with integrated (neural learning based) visuo-auditory processing and (semantics based) relational learning synergistically serving as a foundational backbone in diverse applications such

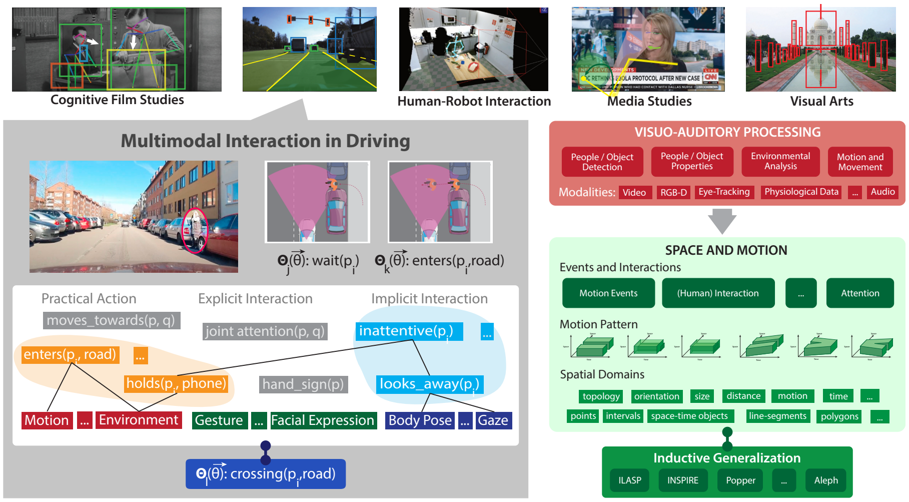

Fig. 1. Relational Spatio-Temporal Structure of Embodied Multimodal Interaction

as behavioural research in psychology (e.g., visual perception), studies in multimodal interaction, HRI and social robotics.

Relational Space-Time Generalisation: A Case for Commonsense Relational learning by inductive generalisation in the context of Aritificial Intelligence (AI) and Machine Learning (ML) is a well-established area of research. Beyond the specific context of Al and ML, the topics of knowledge discovery, explanatory reasoning, hypothesis formation, and decision making assume a far broader significance from philosophical, logical, and cognitive perspectives.

Our approach to relational generalisation from multimodal observations —in so far as this position statement is concerned— is driven by inductive learning based on a logical / knowledge representation and reasoning approach under *constraint logic* and *answer set* learning settings. At the core of the multimodal learning framework are commonsense characterisations of space and motion primitives suited for the grounding of embodied interaction specific multimodal interactional features pertaining to people, objects, gaze, body pose, visual fixation measured via eye-tracking, speech / auditory features such as those relevant to intonation (Fig. 1). For relational (space-time) learning [11], in scope are systems such as ILASP [8], Inspire [9], Popper [5], ALEPH [10]. From an applied viewpoint, the ongoing research is motivated by demonstrating the significance and value of (inductive) generalisation as a means to learn the relational spatio-temporal structure underlying multimodal data pertaining to embodied human interactions in diverse empirical research contexts where the ability to induce high-level, semantic, declaratively explainable behaviour models is of interest. In essence, the learnt behavioural models pertain to some aspect of everyday human activity and interaction.

Acknowledgments

We acknowledge partial funding by the *Federal Ministry for Economic Affairs and Climate Action* (BMWK) as part of the *SituWare* project (reference no. 19A19011C).

Bibliography

[1] Bhatt, M.: 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)

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

[3] Bhatt, M., Kersting, K.: Semantic interpretation of multi-modal human-behaviour data - making sense of events, activities, processes. Künstliche Intell. / Artificial Intelligence **31**(4), 317–320 (2017)

[4] Bhatt, M., Suchan, J.: Cognitive vision and perception. In: ECAI 2020 - 24th European Conference on Artificial Intelligence. Frontiers in Artificial Intelligence and Applications, vol. 325, pp. 2881–2882. IOS Press (2020)

[5] Cropper, A., Morel, R.: Learning programs by learning from failures. Mach. Learn. **110**(4), 801–856 (2021)

[6] Dubba, K.S.R., Cohn, A.G., Hogg, D.C., Bhatt, M., Dylla, F.: Learning relational event models from video. J. Artif. Intell. Res. **53**, 41–90 (2015)

[7] Kondyli, V., Suchan, J., Bhatt, M.: Grounding Embodied Multimodal Interaction: Towards Behaviourally Established Semantic Foundations for Human-Centered AI. In: First International Workshop on Knowledge Representation for Hybrid Intelligence (KR4HI 2022)., part of International Conference on Hybrid Human-Artificial Intelligence (HHAI 2022), Amsterdam, The Netherlands (2022)

[8] Law, M., Russo, A., Broda, K.: Inductive learning of answer set programs. In: Fermé, E., Leite, J. (eds.) Logics in Artificial Intelligence - 14th European Conference, JELIA 2014, Funchal, Madeira, Portugal, September 24-26, 2014. Proceedings. Lecture Notes in Computer Science, vol. 8761, pp. 311–325. Springer (2014)

[9] Schüller, P., Benz, M.: Best-effort inductive logic programming via fine-grained cost-based hypothesis generation - the inspire system at the inductive logic programming competition. Mach. Learn. **107**(7), 1141–1169 (2018)

[10] Srinivasan, A.: The Aleph Manual (2001), http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/

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