24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

Cognitive Vision and Perception

Deep Semantics Integrating AI and Vision for Reasoning about Space, Motion, and Interaction

Mehul Bhatt[1] and **Jakob Suchan**[2]

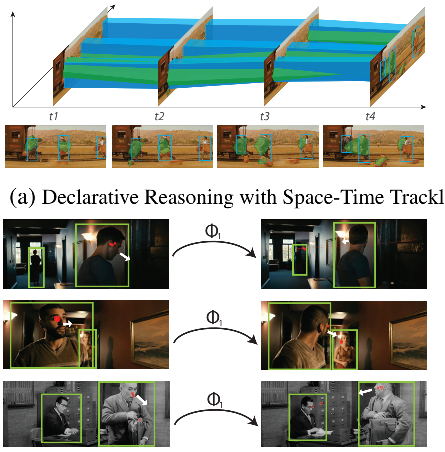
CoDesign Lab / Cognitive Vision — www.codesign-lab.org/cognitive-vision

Abstract. Semantic interpretation of dynamic visuospatial imagery calls for a general and systematic integration of methods in knowledge representation and computer vision. Towards this, we highlight research articulating & developing *deep semantics*, characterised by the existence of declarative models –e.g., pertaining *space and motion*– and corresponding formalisation and reasoning methods supporting capabilities such as semantic question-answering, relational visuospatial learning, and (non-monotonic) visuospatial explanation. We position a working model for deep semantics by highlighting select recent / closely related works from IJCAI [8, 4], AAAI [10], ILP [7], and ACS [9]. We posit that human-centred, explainable visual sensemaking necessitates both high-level semantics and low-level visual computing, with the highlighted works providing a model for systematic, modular integration of diverse multifaceted techniques developed in AI, ML, and Computer Vision.

1 Cognitive Vision and Perception

Cognitive vision 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. 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. The principal focus in cognitive vision & perception –relevant to this highlight paper– is on a systematic integration of computer vision and KR / artificial intelligence methods particularly from the viewpoints of key (computational) **perceptual sensemaking** challenges such as: **(1)**. commonsense scene understanding; **(2)**. semantic question-answering (e.g., with image, video); **(3)**. explainable visual interpretation; **(4)**. concept learning & analogical inference from multimodal stimuli; **(5)**. visuospatial representation learning; **(6)**. visual perception (e.g., with eye-tracking); and **(7)**. multimodal event perception (e.g., for embodied grounding & simulation).

In the context of **(1-7)**, we highlight research in cognitive vision and perception aiming to serve as a model and provide a roadmap for human-centred visuospatial sensemaking. This we position and demo with systematically formalised, general methods based on an integration of state of the art in (relational) AI and (deep learning based) computer vision, and their applications in multidisciplinary areas of socio-technological impact.



(a) Declarative Reasoning with Space-Time Tracklets

(b) Inducing Relational (Visuospatial) Structure

Figure 1: Vision and Perception / Video and Eye-Tracking Data

2 Deep Semantics: Integrating KR and Vision

The development of domain-independent computational models of perceptual sensemaking —e.g., capabilities in (1–7) encompassing visuospatial Q/A, learning, abduction— with multimodal human behaviour stimuli such as RGB(D), video, audio, eye-tracking requires the representational and inferential mediation of commonsense and spatio-linguistically rooted abstractions of space, motion, actions, events and interaction. We characterise Deep Semantics as:

- general methods for the processing and semantic interpretation of dynamic visuospatial imagery with an emphasis on the ability to **abstract, learn, and reason** with cognitively rooted structured characterisations of commonsense knowledge about **space, and motion**.
- the existence of declarative models –e.g., pertaining to space, space-time, motion, actions & events, spatio-linguistic conceptual knowledge (e.g., Table 1)– and corresponding formalisation supporting (domain-neutral) reasoning capabilities (e.g., visual Q/A and learning, non-monotonic visuospatial abduction)

Formal semantics and computational models of deep semantics manifest themselves in declarative AI settings such as constraint logic programming, inductive logic programming, and answer set programming. Naturally, a practical illustration of the intergated "AI and Vision" method requires a tight but modular integration of the (declarative) commonsense spatio-temporal abstraction and reasoning with robust low-level visual computing foundations (primarily) driven by state of the art visual computing techniques (e.g., for visual feature detection, tracking; see Fig. 2).

[1] Örebro University, SWEDEN
[2] University of Bremen, GERMANY

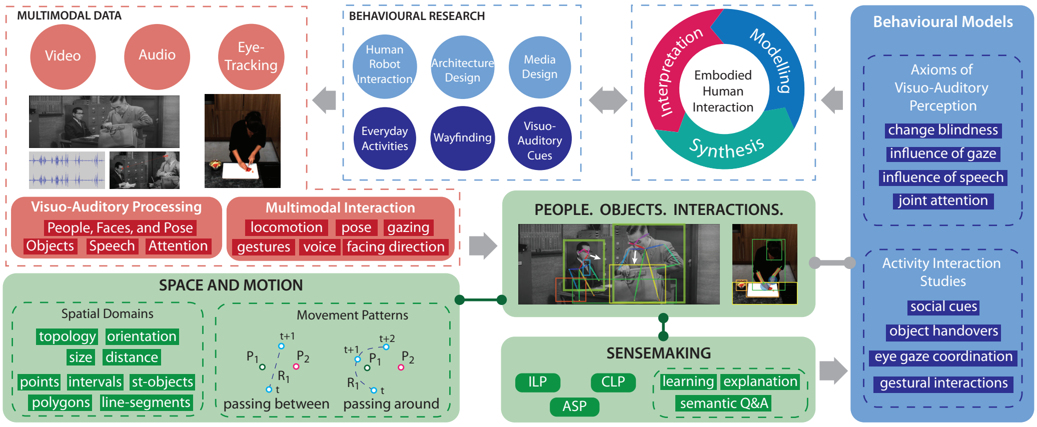


Figure 2: **Deep Semantics for Space, Action and Motion** [1] – Integrated Vision and AI foundations for human behavioural research in embodied interaction; E.g., "*multimodal visuoauditory computing in context for the case of media studies, and everyday activity analysis from the viewpoint of cognitive robotics*".

Abstraction	**Spatial, Time, Motion Relations** *(select sample)*
Mereotopology	disconnected, external contact, partial overlap, tangential proper part, non-tangential proper part, proper part, part of, discrete, overlap, contact
Orientation	left, right, collinear, front, back, on, facing towards, facing away, same direction, opposite direction
Distance, Size	adjacent, near, far, smaller, equi-sized, larger
Motion	moving: towards, away, parallel; growing / shrinking: vertically, horizontally; splitting / merging; rotation: left, right, up, down, clockwise, couter-clockwise
Time	before, after, meets, overlaps, starts, during, finishes, equals

Table 1: Commonsense Spatio-Temporal Relations for Abstracting Space and Motion in Everyday Human Interaction

3 Semantic Interpretation of Multimodal Stimuli

Cognitive vision research is driven by (but not limited to) applications where, for instance, the processing and semantic interpretation of (potentially large volumes of) highly dynamic visuo-spatial imagery is central: autonomous systems, human-machine interaction in cognitive robotics, visuoauditory media technology, and psychology & behavioural research domains where data-centred analytical methods are gaining momentum. For this highlight paper, we position three recent (and mutually related) works that are representative of the crux of *deep semantics*, namely: human-centred representation & relational explainability, declarative reasoning enabled by systematic formalisation, and domain-independence vis-a-vis commonsense spatio-linguistic abstractions supported for modelling space, events, actions, motion, and (inter)action. We position three representative works respectively encompassing question-answering [4, 5], abduction [10, 11], and integration of learning and reasoning [7, 9]: **(1)**. *Semantic Question-Answering with Video & Eye-Tracking*: In [4], a computational framework for semantic-question answering with video and eye-tracking data founded in constraint logic programming is developed; we also demonstrate an application in cognitive film & media studies, where human perception of films vis-a-via cinematographic devices [5] is of interest; **(2)**. *Relational Visual Learning and Neurosymbolic Integration*: In [7, 9], we develop 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 visuospatial symmetry in art); and **(3)**. *Visual Abduction and Moving Objects*: In [8, 10], we develop 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 TRACKLETS as native entities, and founded (functional) answer set programming based spatial reasoning [11]. Other works closely related to the development and application of the deep semantics methodology address relational inductive-abductive inference with video [2], and modelling and interpretation of embodied multimodal interaction in cognitive robotics [3, 6].

REFERENCES

[1] Mehul Bhatt, 'Reasoning about space, actions and change: A paradigm for applications of spatial reasoning', in *Qualitative Spatial Representation and Reasoning: Trends & Future Directions*. IGI Global, 2012.

[2] Krishna Sandeep Reddy Dubba, Anthony G. Cohn, David C. Hogg, Mehul Bhatt, and Frank Dylla, 'Learning Relational Event Models from Video', *JAIR*, **53**, 41–90, (2015).

[3] Michael Spranger, Jakob Suchan, and Mehul Bhatt, 'Robust Natural Language Processing - Combining Reasoning, Cognitive Semantics and Construction Grammar for Spatial Language', in *IJCAI 2016*. AAAI Press, (July 2016).

[4] Jakob Suchan and Mehul Bhatt, 'Semantic question-answering with video and eye-tracking data: AI foundations for human visual perception driven cognitive film studies', in *IJCAI 2016, New York, USA*, ed., S. Kambhampati, pp. 2633–2639. IJCAI/AAAI Press, (2016).

[5] Jakob Suchan and Mehul Bhatt, 'The Geometry of a Scene: On Deep Semantics for Visual Perception Driven Cognitive Film Studies', in *2016 IEEE WACV 2016, NY, USA*, pp. 1–9, (2016).

[6] Jakob Suchan and Mehul Bhatt, 'Deep Semantic Abstractions of Everyday Human Activities: On Commonsense Representations of Human Interactions', in *ROBOT 2017: 3rd Iberian Robotics Conference, Advances in Intelligent Systems and Computing 693*, (2017).

[7] Jakob Suchan, Mehul Bhatt, and Carl P. L. Schultz, 'Deeply semantic inductive spatio-temporal learning', in *ILP 2016*, volume 1865, pp. 73–80. CEUR-WS.org, (2016).

[8] Jakob Suchan, Mehul Bhatt, and Srikrishna Varadarajan, 'Out of sight but not out of mind: An answer set programming based online abduction framework for visual sensemaking in autonomous driving', in *IJCAI 2019*., pp. 1879–1885. ijcai.org, (2019).

[9] Jakob Suchan, Mehul Bhatt, Srikrishna Vardarajan, Seyed Ali Amirshahi, and Stella Yu, 'Semantic Analysis of (Reflectional) Visual Symmetry: A Human-Centred Computational Model for Declarative Explainability', *Advances in Cognitive Systems*, **6**, 65–84, (2018).

[10] Jakob Suchan, Mehul Bhatt, Przemyslaw Andrzej Walega, and Carl P. L. Schultz, 'Visual explanation by high-level abduction: On answer-set programming driven reasoning about moving objects', in *AAAI 2018*, pp. 1965–1972. AAAI Press, (2018).

[11] Przemyslaw Andrzej Walega, Mehul Bhatt, and Carl P. L. Schultz, 'ASPMT(QS): non-monotonic spatial reasoning with answer set programming modulo theories', in *LPNMR 2015*, vol. 9345 of *LNCS*, pp. 488–501. Springer, (2015).