Deep Semantics for Explainable Visuospatial Intelligence

Perspectives on Integrating Commonsense Spatial Abstractions and Low-Level Neural Features

Jakob Suchan[1] , **Mehul Bhatt**[2] and **Srikrishna Varadarajan**

[1]University of Bremen, Germany – [2]Örebro University, Sweden

CoDesign Lab EU / Cognitive Vision – www.cognitive-vision.org

Abstract. High-level semantic interpretation of (dynamic) visual imagery calls for general and systematic methods integrating techniques in knowledge representation and computer vision. Towards this, we position *deep semantics*, denoting the existence of declarative models –e.g., pertaining *space and motion*– and corresponding formalisation and methods supporting (domain-independent) explainability capabilities such as semantic question-answering, relational (and relationally-driven) visuospatial learning, and (non-monotonic) visuospatial abduction. Rooted in recent work, we summarise and report the status quo on deep visuospatial semantics —and our approach to neurosymbolic integration and explainable visuo-spatial computing in that context— with developed methods and tools diverse settings such as behavioural research in psychology, art & social sciences, and autonomous driving.

Visuospatial Intelligence: Cognitive Vision and Perception

Cognitive vision and perception research addresses (embodied) visual, visuospatial and visuo-locomotive perception and interaction from the viewpoints of language, logic, spatial cognition and artificial intelligence. The principal focus is on a systematic integration of vision and artificial intelligence methods particularly from the viewpoint of (computational) visuospatial intelligence encompassing capabilities such as: commonsense scene understanding; semantic question-answering (e.g., with image, video); explainable visual interpretation; analogical inference with visual imagery; relational concept learning; visuospatial representation learning; visual perception (e.g., with eye-tracking); multimodal event perception (e.g., for embodied grounding & simulation)

Cognitive vision presents 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.

Deep Semantics: Integrating AI and Vision

The development of domain-independent computational models of visuospatial intelligence with multimodal human behavioural stimuli (such as RGB-D, audio, eye-tracking) requires the representational and inferential mediation of commonsense and spatio-linguistically rooted abstractions of space, motion, actions, events and interaction. Driven by this, and particularly in the backdrop of *perceptual sensemaking* capabilities such as visuospatial question-answering, (relational) visuospatial concept learning, (non-monotonic) visuospatial abduction, we characterise *deep visuospatial semantics* by:

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, Motion, Spatio-Temporal Structure in Everyday Human Interaction

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

Formal semantics and computational models of deep (visuospatial) semantics manifest themselves as systematic, general, and domain-neutral methods developed by modular but tight neurosymbolic integration consisting of (Fig. 1):

(D1). **Commonsense / Space, Events, Actions, and Change.**[1]
The ability to (declaratively) specify and solve foundational problems related to (mixed) geometric and qualitative visuospatial representation and reasoning pertaining to temporal, spatial, and spatio-temporal *things*, be it abstract regions of

[1]Commonsense spatio-temporal relations and patterns (Table 1; e.g. left-of, touching, part-of, during, approaching) offer a human-centered and cognitively adequate formalism for grounding and logic-based automated reasoning about embodied spatio-temporal interactions, e.g., such as those involved in everyday activities involving object manipulation and control, physical locomotion, interpersonal interaction, visuo-spatial thinking.

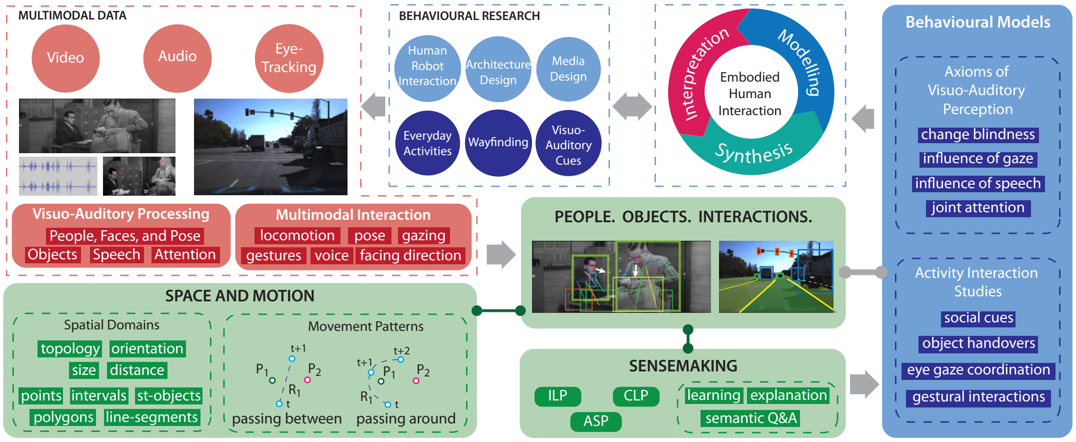

Figure 1: **Deep Semantics Integrating Knowledge Representation and Visual Computing** – "*multi-modal visuo-auditory computing in context*".

space, time, space-time, geometric entities and physical objects, or spatial artefacts without any real physical manifestation (e.g., shadows of objects, areas of visual attention). The declarative programming paradigms being alluded to here are *constraint logic programming* (CLP) [Jaffar and Maher, 1994], *inductive logic programming* (ILP) [Muggleton and Raedt, 1994], and *answer set programming* (modulo theories) (ASP, ASPMT) [Brewka *et al.*, 2011; Lee and Meng, 2013].

(D2). Visual Computing / Object Detection, Tracking, Etc.

Robust low-level visual computing foundations (primarily) driven by state of the art *deep learning* techniques (for visual feature detection, tracking etc). In particular: object / people detection and tracking (Faster RCNN [Ren *et al.*, 2015; Bewley *et al.*, 2016], YOLOv3 [Redmon and Farhadi, 2016]), faces (TinyFaces [Hu and Ramanan, 2016]), body-structure (OpenPose [Cao *et al.*, 2018]).

With the aim to position neurosymbolic integration(s) (with **D1** and **D2**), this report summarises recent results and ongoing work from the viewpoints of **(I–II)** : **(I)**. development of knowledge representation and reasoning methods (in the CLP, ILP and ASP family) that enable handling space and motion as first-class objects within the declarative programming settings afforded by the respective methods under consideration; in particular, CLP(QS), ILP(QS), and ASPMT(QS) [Bhatt *et al.*, 2011; Schultz *et al.*, 2018; Suchan *et al.*, 2016a; Walega *et al.*, 2015]; and **(II)**. practical manifestations of deep semantic reasoning and learning capabilities (e.g., Q/A, relational learning, visuaospatial abduction) in diverse application domains [Suchan *et al.*, 2019; Suchan *et al.*, 2018a; Suchan *et al.*, 2018b; Suchan and Bhatt, 2017a; Suchan and Bhatt, 2017b; Suchan, 2017; Suchan *et al.*, 2016b; Spranger *et al.*, 2016; Suchan and Bhatt, 2016a; Suchan and Bhatt, 2016b; Dubba *et al.*, 2011; Dubba *et al.*, 2015].

Semantic Interpretation of Multimodal Stimuli

Cognitive vision research is driven by application areas where, for instance, the processing and semantic interpretation of (potentially large volumes of) highly dynamic visuo-spatial imagery is central: autonomous systems, cognitive robotics, self-driving vehicles, visuo-auditory media design,

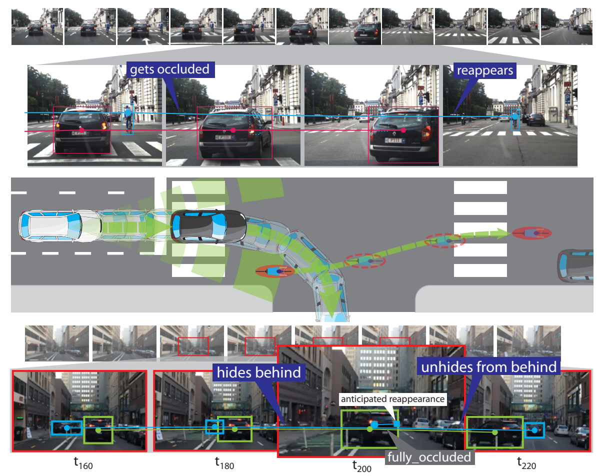

Figure 2: **Out of sight but not out of mind**; Abducing occlusion events to anticipate reappearance; the case of hidden entities: e.g., occluded cyclist (top) / vehicle (bottom).

and psychology & behavioural research domains where data-centred analytical methods are gaining momentum. We summarise select cases in these contexts based on recently published and emerging lines of research, in particular [Suchan *et al.*, 2019; Suchan *et al.*, 2018a; Suchan and Bhatt, 2016a; Suchan *et al.*, 2016a]:

CASE I. Human-Centred Semantic Explainability Considerations in Autonomous Driving.

Autonomous driving research has developed (and been driven by) advances in *deep learning* based computer vision research. Although deep learning based vision & control has (arguably) been very successful for self-driving vehicles, we posit that there is a clear need and tremendous potential for hybrid visual sensemaking solutions, e.g., integrating *vision and semantics*, towards fulfilling essential legal and ethical responsibilities involving explainability, human-centred AI, and industrial standardisation (e.g, pertaining to representation, realisation of rules and norms) [Suchan *et al.*, 2019].

▶ *Standardisation & Regulation, Diagnostics etc* Current autonomous driving research is primarily focussed on two basic considerations: *how fast to drive, and which way and*

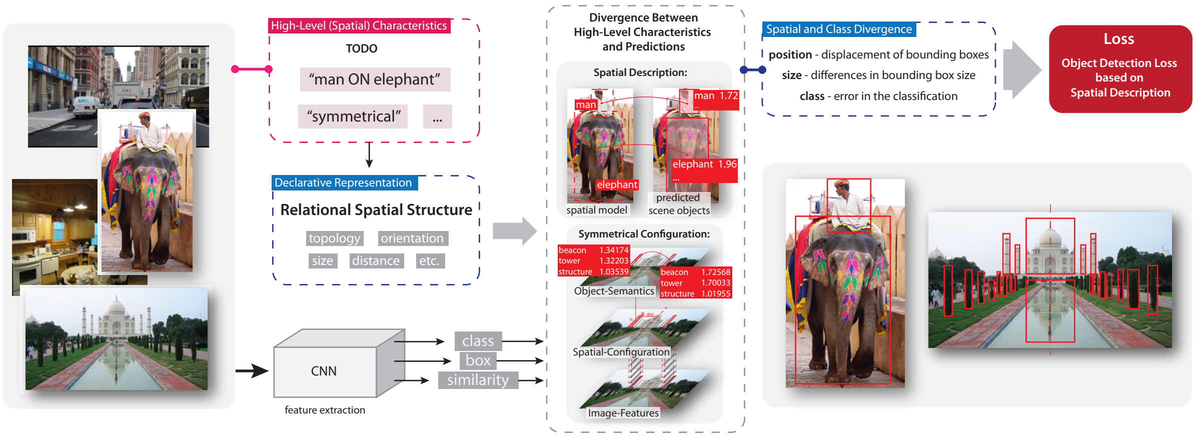

Figure 3: **Semantically Guided Neural Learning** – "*relational visuospatial structure guided optimisation (of the loss function)*".

how much to steer. For further developments, it will be necessary to have a community consensus on aspects such as representation, interoperability, human-centred benchmarks, and data archival & retrieval mechanisms.[2] Ethically driven standardisation & regulation will require addressing challenges in semantic visual interpretation, natural / multimodal human-machine interaction, and high-level data analytics (e.g., for post hoc diagnostics, dispute settlement). This will necessitate –amongst other things– human-centred qualitative benchmarks and multifaceted hybrid AI solutions.

▸ *Realtime Visuospatial Abduction* Consider the *occlusion scenario* in Fig. 2: Car (c) is in-front, and indicating to turn-right; during this time, person (p) is on a bicycle (b) and positioned front-right of c and moving-forward. Car c turns-right, during which the bicyclist $<p, b>$ is not visible. Subsequently, bicyclist $<p, b>$ reappears. This occlusion scenario indicates several challenges concerning aspects such as: identity maintenance, making default assumptions, computing counterfactuals, projection, and interpolation of missing information (e.g., what could be hypothesised about bicyclist $<p, b>$ when it is *occluded*; how can this hypothesis enable in planning an immediate next step). Addressing such challenges —be it realtime or post-hoc— in view of human-centred AI concerns pertaining to ethics, explainability and regulation requires a systematic integration of **Semantics and Vision**, 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.

We showcase a general method for *online* (i.e., incremental, realtime) visual sensemaking using answer set programming is systematically formalised and fully implemented. The method integrates state of the art in (deep learning based) visual computing, and is developed as a modular framework usable within hybrid architectures for perception & control. Our evaluation and demo is based on community established benchmarks KITTIMOD [Geiger *et al.*, 2012] and MOT [Milan *et al.*, 2016]. As a use-case, we focus on the significance of human-centred visual sensemaking —e.g., semantic representation and explainability, question-answering, commonsense interpolation— in safety-critical autonomous driving situations.

CASE II. Semantically Guided Neural Learning and (Explainable) Visuospatial Interpretation

We showcase a computational model (Figs. 3–4) with the capability to generate semantic, explainable interpretation models for the analysis of visuospatial symmetry [Suchan *et al.*, 2018a]; more generally, we emphasis the capability of the model wherein the incremental learning process itself may be semantically guided by conceptual visuospatial knowledge (e.g., qualitative description of symmetry, or arbitrary spatial constraints amongst abstract representations of domain entities / visuospatial features by way of points, line-segments, polygons etc). The explainability is founded on a domain-independent, mixed qualitative-quantitive representation of visuo-spatial relations based on which the symmetry is declaratively characterised. From an applied viewpoint, the developed methodology is intended to serve as the technical backbone for assistive and analytical technologies for visual media studies, e.g., from the viewpoint of behavioural research in psychology [Suchan *et al.*, 2016b], empirical aesthetics, cultural heritage.

▸ *Semantics Guided Optimisation* Visuospatial characteristics (e.g., reflectional symmetry, or other arbitrary discriminants) can be declaratively formalised in a semantic model by describing their respective relational structure (see [Suchan *et al.*, 2018a] for the case of reflectional symmetry). Our proposed system utilises such high-level (spatial) scene descriptions for guiding neural learning by utilising a declarative model of spatial divergence to calculate loss; i.e., for object detection the loss for training Faster RCNN may be calculated based on the divergence of a predicted scene structure to a high-level characterisation, e.g., coming from an image description. For assessing the correctness of the predictions from the neural network the predicted scene model (i.e., the detected objects and spatial relations between these objects) can be compared to the spatial characterisation represented as relational spatial structure. This the divergence of scene objects from a relational structure are definable based on at-

[2] Within autonomous driving, the need for standardisation and ethical regulation has most recently garnered interest internationally, e.g., with the Federal Ministry of Transport and Digital Infrastructure in Germany taking a lead in eliciting 20 key propositions (with legal implications) for the fulfilment of ethical commitments for automated and connected driving systems [BMVI, 2018].

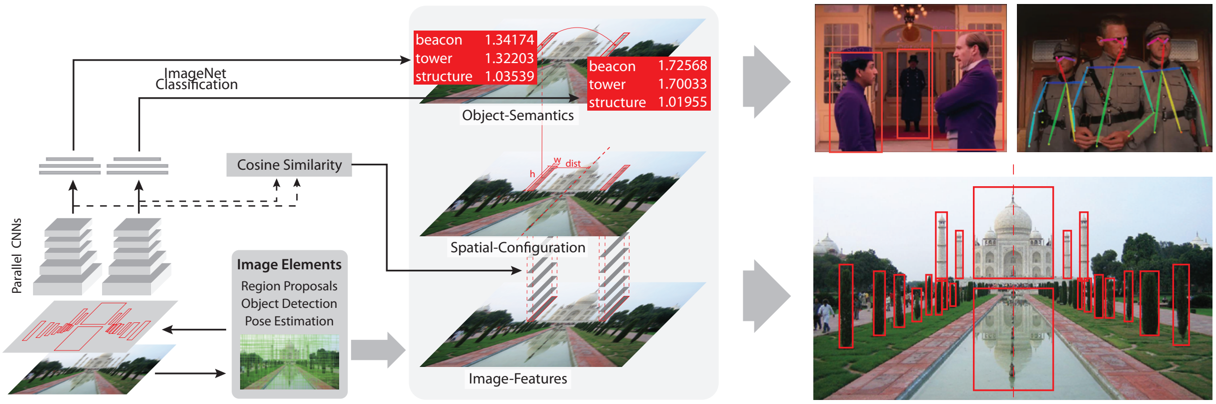

Figure 4: **A Multi-level model of reflectional symmetry for application in visual arts** – "*towards a neurosymbolic explainable interpretation model of multi-level semantic symmetry*".

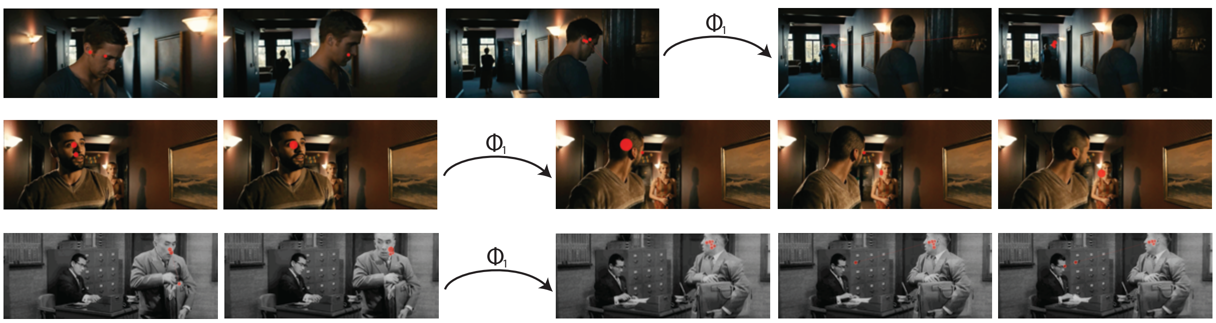

Figure 5: **Learning Axioms of Visual-Auditory Perception** – "*Onscreen gaze transition driven shift of visual attention*". (red markers on each frame correspond to the visual fixation data as obtained via an eye-tracker as part of an visual perception experiment. *Media sources (as per Fair Use)*: Drive (2011 / Director: Nicolas Winding Refn), The Bad Sleep Well (1960 / Director: Akira Kurosawa)

tributes such as *position*, *size*, and *class score*; for instance, e.g., for the description "*man on elephant*" (Fig 3), we calculate the distance of the detected bounding boxes to a configuration that is consistent with the description. Similarly in the case of a highly symmetrical image (e.g., the Taj Mahal, or the the movie scene in Fig. 4), we may calculate the divergence of the detected boxes to a symmetrical configuration. The loss for each detection may then be calculated based on the spatial (position and size) and the class divergence of the detected box to the corresponding box fulfilling the spatial constraints imposed by the image description.

CASE III. Cognitive Vision Foundations for Research in Psychological and Behavioural Sciences

Our research in this area is broadly driven by the need to systematically learning high-level behavioural models of embodied multimodal interaction as applicable in varied contexts such as visual perception (psychology), environmental behaviour studies (environmental psychology), human robot interaction (cognitive robotics). The particular emphasis is on *inducing behavioural models* from the viewpoint of psychology and related behavioural research domains, where data-centred analytical methods in naturalistic experimental settings are now gaining momentum.

▸ *Visuo-Auditory Perception Research* We demonstrate the case of *cognitive media studies* with a focus on (eye-tracking driven) visual perception of visuo-auditory media (e.g., narrative film) [Suchan and Bhatt, 2016a; Suchan and Bhatt, 2016b; Suchan *et al.*, 2016a]. Consider Fig. 5, consisting of three sets of frames from three different film scenes; the frames are superimposed with the eye-tracking data obtained as part of an eye-tracking experiment / dataset ([Suchan and Bhatt, 2016a]) . Here, Θ_1 in each of the three sets of frames corresponding to a *gaze transition* event of one of the on-screen characters (as trackable via a head rotation event; Fig. 6). Corresponding to this the gaze transition is a shift of attention (as measurable via eye-tracking data) from one location in media to another (in this case, towards the direction of the gaze of the tracked character). From the viewpoint of this demo, of particular interest is computational learning of human reception of media (as recorded within large-scale experiments) vis-a-vis visuo-auditory computational narrative structure (i.e., *geometry of the scene* [Suchan and Bhatt, 2016b]) of the medium itself. Indeed, the goal here is to acquire qualitative or high-level knowledge about human behaviour from large-scale experiments / multimodal datasets.[3]

[3] For the purposes of (a possible presentation at) NeSy 2019, we restrict to visual computing foundations. However, it is worth noting that in the domain of visual perception, auditory perception is equally important: the audio analysis focuses on analysing human speech, and in particular on segmenting parts of the audio where people are speaking and identifying the different speakers. Towards this, we first detect speech parts and then cluster the speech parts for speaker diarization in the following way: (**1**). *Speech detection*. We use RNN based sound event detection [Adavanne and Virtanen, 2017] trained on the DCASE dataset [Stowell *et al.*, 2015] to detect speech within the audio track; (**2**). *Speaker diarization*. The detected speech parts are used as a basis for clustering speaker

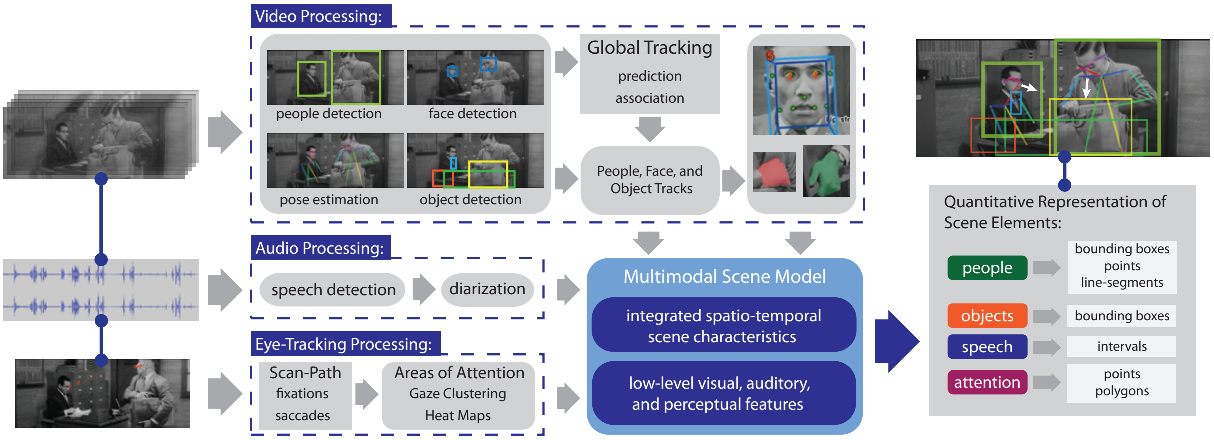

Figure 6: Visuo-Auditory Computing Pipeline

Conclusion

Driven by a systematic integration of *knowledge representation* and *computer vision*, we report on an established line of research in computational cognitive vision and perception focussing on general, systematic methods for the semantic interpretation of dynamic visual imagery. Our experience suggests that deep learning based computer vision is highly powerful; with a little bit of semantics: (1) the performance of low-level visual computing (e.g., tracking & detection) can be improved; (2). neural feature learning can be influenced by high-level semantics, and that semantic models are necessary for fulfilling capabilities for high-level introspection / explanation / inductive model-building; (3) both knowledge representation and low-level vision are essential to realise computational visual intelligence.

References

[Adavanne and Virtanen, 2017] Sharath Adavanne and Tuomas Virtanen. Sound event detection using weakly labeled dataset with stacked convolutional and recurrent neural network. In *Detection and Classification of Acoustic Scenes and Events 2017*, pages 12–16, November 2017.

[Bewley *et al.*, 2016] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468, 2016.

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

[Bhatt *et al.*, 2011] Mehul Bhatt, Jae Hee Lee, and Carl P. L. Schultz. CLP(QS): A declarative spatial reasoning framework. In Max J. Egenhofer, Nicholas A. Giudice, Reinhard Moratz, and Michael F. Worboys, editors, *Spatial Information Theory - 10th International Conference, COSIT 2011, Belfast, ME, USA, September 12-16, 2011. Proceedings*, volume 6899 of *Lecture Notes in Computer Science*, pages 210–230. Springer, 2011.

[Bhatt *et al.*, 2013] Mehul Bhatt, Jakob Suchan, and Carl P. L. Schultz. Cognitive Interpretation of Everyday Activities - Toward Perceptual Narrative Based Visuo-Spatial Scene Interpretation. In *2013 Workshop on Computational Models of Narrative, CMN 2013, August 4-6, 2013, Hamburg, Germany*, pages 24–29, 2013.

identities using the LIUM speaker diarization toolbox [Meignier and Merlin, 2010] for separating segments of different people speaking and assign a unique ID to each speaker.

[Bhatt, 2012] Mehul 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.

[Bhatt, 2013] Mehul Bhatt. Between Sense and Sensibility: Declarative narrativisation of mental models as a basis and benchmark for visuo-spatial cognition and computation focussed collaborative cognitive systems. *CoRR*, abs/1307.3040, 2013.

[Bhatt, 2018] Mehul Bhatt. Cognitive Media Studies: Potentials for Spatial Cognition and AI research. *Cognitive Processing*, 19(Suppl. 1):S6–S6, Sep 2018. as part of: Spatial Cognition in a Multimedia and Intercultural World Proceedings of the 7th International Conference on Spatial Cognition (ICSC 2018).

[BMVI, 2018] BMVI. Report by the ethics commission on automated and connected driving., bmvi: Federal ministry of transport and digital infrastructure, germany, 2018.

[Brewka *et al.*, 2011] Gerhard Brewka, Thomas Eiter, and Miroslaw Truszczyński. Answer set programming at a glance. *Commun. ACM*, 54(12):92–103, December 2011.

[Cao *et al.*, 2018] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018.

[Cordts *et al.*, 2016] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[DBL,]

[Dubba *et al.*, 2011] Krishna Sandeep Reddy Dubba, Mehul Bhatt, Frank Dylla, David C. Hogg, and Anthony G. Cohn. Interleaved Inductive-Abductive Reasoning for Learning Complex Event Models. In Stephen Muggleton, Alireza Tamaddoni-Nezhad, and Francesca A. Lisi, editors, *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*, pages 113–129. Springer, 2011.

[Dubba *et al.*, 2015] Krishna Sandeep Reddy Dubba, Anthony G. Cohn, David C. Hogg, Mehul Bhatt, and Frank Dylla. Learning Relational Event Models from Video. *J. Artif. Intell. Res. (JAIR)*, 53:41–90, 2015.

[Geiger *et al.*, 2012] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[Hu and Ramanan, 2016] Peiyun Hu and Deva Ramanan. Finding tiny faces. *CoRR*, abs/1612.04402, 2016.

[Jaffar and Maher, 1994] Joxan Jaffar and Michael J Maher. Constraint logic programming: A survey. *The journal of logic programming*, 19:503–581, 1994.

[Lee and Meng, 2013] Joohyung Lee and Yunsong Meng. Answer set programming modulo theories and reasoning about continuous changes. In *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*, 2013.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[Meignier and Merlin, 2010] Sylvain Meignier and Teva Merlin. Lium spkdiarization: an open source toolkit for diarization. In *in CMU SPUD Workshop*, 2010.

[Milan *et al.*, 2016] Anton Milan, Laura Leal-Taixé, Ian D. Reid, Stefan Roth, and Konrad Schindler. MOT16: A benchmark for multi-object tracking. *CoRR*, abs/1603.00831, 2016.

[Muggleton and Raedt, 1994] Stephen Muggleton and Luc De Raedt. Inductive logic programming: Theory and methods. *JOURNAL OF LOGIC PROGRAMMING*, 19(20):629–679, 1994.

[Redmon and Farhadi, 2016] Joseph Redmon and Ali Farhadi. YOLO9000: better, faster, stronger. *CoRR*, abs/1612.08242, 2016.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Annual Conference on Neural Information Processing Systems 2015, Canada*, 2015.

[Schultz *et al.*, 2018] Carl P. L. Schultz, Mehul Bhatt, Jakob Suchan, and Przemyslaw Andrzej Walega. Answer set programming modulo 'space-time'. In Christoph Benzmüller, Francesco Ricca, Xavier Parent, and Dumitru Roman, editors, *Rules and Reasoning - Second International Joint Conference, RuleML+RR 2018, Luxembourg, September 18-21, 2018, Proceedings*, volume 11092 of *Lecture Notes in Computer Science*, pages 318–326. Springer, 2018.

[Spranger *et al.*, 2014] Michael Spranger, Jakob Suchan, Mehul Bhatt, and Manfred Eppe. Grounding Dynamic Spatial Relations for Embodied (Robot) Interaction. In *PRICAI 2014: Trends in Artificial Intelligence - 13th Pacific Rim International Conference on Artificial Intelligence, Gold Coast, QLD, Australia, December 1-5, 2014. Proceedings*, volume 8862 of *Lecture Notes in Computer Science*, pages 958–971. Springer, 2014.

[Spranger *et al.*, 2016] Michael Spranger, Jakob Suchan, and Mehul 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, July 2016.

[Stowell *et al.*, 2015] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M.D. Plumbley. Detection and classification of acoustic scenes and events. *Multimedia, IEEE Transactions on*, 17(10):1733–1746, Oct 2015.

[Suchan and Bhatt, 2012] Jakob Suchan and Mehul Bhatt. Toward an Activity Theory Based Model of Spatio-Temporal Interactions - Integrating Situational Inference and Dynamic (Sensor) Control. In *STAIRS 2012 - Proceedings of the Sixth Starting AI Researchers' Symposium, Montpellier, France, 27-28 August 2012*, pages 318–329, 2012.

[Suchan and Bhatt, 2016a] 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 *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 2633–2639, 2016.

[Suchan and Bhatt, 2016b] Jakob Suchan and Mehul 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*, pages 1–9, 2016.

[Suchan and Bhatt, 2017a] Jakob Suchan and Mehul Bhatt. Commonsense Scene Semantics for Cognitive Robotics: Towards Grounding Embodied Visuo-Locomotive Interactions. In *ICCV 2017 Workshop: Vision in Practice on Autonomous Robots (ViPAR), International Conference on Computer Vision (ICCV)*, 2017.

[Suchan and Bhatt, 2017b] Jakob Suchan and Mehul 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.

[Suchan *et al.*, 2014] Jakob Suchan, Mehul Bhatt, and Paulo E. Santos. Perceptual Narratives of Space and Motion for Semantic Interpretation of Visual Data. In *Computer Vision - ECCV 2014 Workshops - Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part II*, pages 339–354, 2014.

[Suchan *et al.*, 2015] Jakob Suchan, Mehul Bhatt, and Harshita Jhavar. Talking about the Moving Image: A Declarative Model for Image Schema Based Embodied Perception Grounding and Language Generation. *CoRR*, abs/1508.03276, 2015.

[Suchan *et al.*, 2016a] Jakob Suchan, Mehul Bhatt, and Carl P. L. Schultz. Deeply semantic inductive spatio-temporal learning. In James Cussens and Alessandra Russo, editors, *Proceedings of the 26th International Conference on Inductive Logic Programming (Short papers), London, UK, 2016.*, volume 1865 of *CEUR Workshop Proceedings*, pages 73–80. CEUR-WS.org, 2016.

[Suchan *et al.*, 2016b] Jakob Suchan, Mehul Bhatt, and Stella Yu. The Perception of Symmetry in the Moving Image: Multi-level Computational Analysis of Cinematographic Scene Structure and its Visual Reception. In *Proceedings of the ACM Symposium on Applied Perception, SAP 2016, Anaheim, California, USA, July 22-23, 2016*, page 142, 2016.

[Suchan *et al.*, 2018a] 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.

[Suchan *et al.*, 2018b] 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 Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1965–1972. AAAI Press, 2018.

[Suchan *et al.*, 2019] Jakob Suchan, Mehul Bhatt, and Srikrishna Vardarajan. 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 and Thomas Eiter, editors, *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI 2019), Macao, August 10-16*, 2019. (To appear).

[Suchan, 2017] Jakob Suchan. Declarative Reasoning about Space and Motion with Video. *KI*, 31(4):321–330, 2017.

[Walega *et al.*, 2015] Przemyslaw Andrzej Walega, Mehul Bhatt, and Carl P. L. Schultz. ASPMT(QS): non-monotonic spatial reasoning with answer set programming modulo theories. In Francesco Calimeri, Giovambattista Ianni, and Miroslaw Truszczynski, editors, *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*, pages 488–501. Springer, 2015.