Declarative Reasoning about Space and Motion in Visual Imagery

Theoretical Foundations and Applications
Declarative Reasoning about Space and Motion in Visual Imagery

Theoretical Foundations and Applications

Doctoral Thesis
(Dissertation)

submitted to attain the degree of doctor of natural sciences
(zur Erlangung des Doktorgrades Doktor der Naturwissenschaften)
– Dr. rer. nat. –

Jakob Suchan

University of Bremen, Germany
(Universität Bremen, Deutschland)

July 2022
Date of Defense:  
June 22, 2022

Examiner (1. Gutachter):  Prof. Dr. Mehul Bhatt  
Örebro University, Sweden (Universität Örebro, Schweden)

Co-Examiner (2. Gutachter):  Prof. Dr. Bernd Krieg-Brückner  
University of Bremen, Germany (Universität Bremen, Deutschland)
Abstract

Perceptual sensemaking of dynamic visual imagery, e.g., involving semantic grounding, explanation, and learning, is central to a range of tasks where artificial intelligent systems have to make decisions and interact with humans. Towards this, commonsense characterisations of space and motion encompassing spatio-temporal relations, motion patterns, and events provide an abstraction layer to perform semantic reasoning about (embodied) spatio-temporal interactions observed from visuospatial imagery.

This thesis develops: (1). a general theory about space and motion for representing and reasoning about interactions founded in declaratively grounded models pertaining to space, time, space-time, motion, and events, and (2). a computational cognitive vision framework for perceptual sensemaking with visuospatial imagery, systematically developed to be compliant with declarative programming methods such as Constraint Logic Programming (CLP), Answer-Set Programming (ASP), and Inductive Logic Programming (ILP).

The thesis provides general tools and methods for declarative reasoning with visuospatial imagery, encompassing question-answering, abduction, and integration of reasoning and learning; contributed publications in this thesis focus on:

1. **Grounded Semantic Interpretation and Question-Answering** rooted to expressive declarative models of (embodied) visuospatial semantics to characterise (human) interactions with respect to their relational spatio-temporal structure;

2. **Visuospatial Abduction**, for hypothesising object interactions explaining perceived visuospatial dynamics, tightly integrating low-level (neural) visual processing and high-level (relational) abductive reasoning; and

3. **Declarative Explainability and Inductive Generalisation** based on declarative formalisations of visuospatial image characteristics grounded in (symbolic and subsymbolic) image elements and (neural) image features thereby providing a relational abstraction layer suitable for relational (inductive) learning.

These developed representation and reasoning capabilities are demonstrated and evaluated in the context of real-world applications (with requirements such as real-time processing, robustness against noise, etc.), where the processing and semantic interpretation of (potentially large volumes of) highly dynamic visuospatial imagery is central. Example applications included in this thesis encompass cognitive robotics, autonomous vehicles, and assistive technologies for human behaviour research.
Zusammenfassung


Vor dem Hintergrund dieses Rahmenwerks präsentieren wir allgemeine Methoden und Werkzeuge für deklatives Schließen mit räumlich-visuellen Daten. Im Besonderen umfasst dies: **Beantwortung von Fragen** mit semantischen Informationen, **Abduktion** und die Integration von **logischem Schließen und maschinellem Lernen**.

Die in dieser Arbeit beinhalteten Publikationen legen ihren Fokus dabei auf die folgenden drei Schwerpunktthemen:

1. **Semantische Interpretation und Beantwortung von Fragen** unter Verwendung deklarativer Modelle zur semantischen räumlich-visuellen Charakterisierung (menschlicher) Interaktionen in Bezug auf ihre relationale räumlich-zeitliche Struktur, verankert in der wahrgenommenen Dynamik der Szene;

2. **Räumlich-Visuelle Abduktion zur Hypothesenbildung über Objektinteraktionen**, die wahrgenommene räumlich-visuelle Dynamik erklären. Dies beinhaltet insbesondere die Integration von Methoden zur visuellen Verarbeitung von Bilddaten und zum abduktiven logischen Schließen in Bezug auf die Szenendynamik; und
3. **Deklarative Erklärbarkeit und Induktive Generalisierung**, basierend auf deklarativen Formalisierungen räumlich-visueller Bildeigenschaften, die auf (symbolischen und subsymbolischen) Bildelementen und (neuronalen) Bildmerkmalen beruhen, die eine für (induktives) Lernen geeignete relationale Abstraktionsebene bilden.

Diese entwickelten Repräsentations- und Schlussfolgerungsmethoden werden im Kontext realer Anwendungen demonstriert und evaluiert, bei denen die Verarbeitung und semantische Interpretation von (potenziell großen Mengen) dynamischer räumlich-visueller Daten zentral ist, insbesondere *kognitive Robotik, autonome Fahrzeuge* und Assistenzsysteme für *menschliche Verhaltensforschung*. 
Acknowledgements

I would like to express my deep gratitude towards my advisor Mehul Bhatt for his constant support and guidance during the entire process of writing this thesis, and developing this research. I have always enjoyed the endless days and nights working in the lab, in a cafe, a hotel lobby, or an Indian restaurant.

I would also like to thank Christian Freksa for his support and for giving me the opportunity to start my scientific work in his group and in the SFB/TR8 Spatial Cognition. Discussions with Christian were always inspiring and helped me to shape my research. Further, I would like to thank Bernd Krieg-Brückner for serving as my second examiner.

I would like to thank Michael Spranger for his valuable comments and discussions regarding dynamic spatial relations in natural language processing and for hosting my visit at Sony CSL in Tokyo, Japan, and Stella Yu for her perspectives on semantic sense-making with visual data from a computer vision point of view and for welcoming us at ICSI Berkeley, USA.

I would like to thank the co-authors of the publications included in this thesis for the great collaboration: Mehul Bhatt, Carl Schultz, Przemysław Wałęga, Srikrishna Varadarajan, Michael Spranger, Stella Yu, and Seyed Ali Amirshahi. Further, I would like to thank all the co-workers, interns, and students that have contributed to the overall research agenda of Cognitive Vision. In particular, I want to thank Vasiliki Kondyli for providing insights from the viewpoint of architecture, environmental psychology, and human behaviour in driving, Carl Schultz for sharing his expertise on declarative spatial reasoning systems and for always being available to solve technical problems, and the interns supported by the DAAD Rise program: Richard Addo, Liana Mikaelyan, and Ke (Kathrine) Huang.

Finally, and foremost, I would like to thank my family Sofia, Natalie, Marianne, Martin, Gregor, and Klara for their love, encouragement, and understanding.
## Contents

Abstract v  
Zusammenfassung vii  
Acknowledgements ix  
Select Publications Included in this Thesis xiii  
List of Figures xv  
Acronyms xvii  
List of Symbols xix  

1 Introduction 1  
1.1 Motivation 1  
1.2 Key Contributions 4  
1.3 Organisation of the Thesis 5  

2 Space and Motion in Cognitive Vision Systems 7  
2.1 Commonsense Visual Intelligence. 8  
2.1.1 Cognitive Vision: Integrating KR and Computer Vision 8  
2.1.2 Neurosymbolism and Explainability 9  
2.2 Space, Time, Motion, and Events 10  
2.2.1 Qualitative Spatio-Temporal Reasoning 11  
2.2.2 Ontology and Formal Model 12  
2.3 Perceptual Sensemaking with Dynamic Visuospatial Imagery 16  
2.3.1 Methods and Tools (Included Publications) 18  
2.3.2 Applications and Interdisciplinarity 22  
2.3.3 Datasets (Developed and/or Used) 24  

3 Semantic Q/A with Video 27  
3.1 Visuospatial Semantics 28  
3.1.1 - Declarative Reasoning about Space and Motion with Video – KI Journal 33  
3.1.2 - Semantic Q/A with Video and Eye-Tracking Data – IJCAI 45  
3.1.3 - The Geometry of a Scene – WACV 55
3.2 Declarative Reasoning about Embodied Interaction ........................................... 66
  ■ 4 - Deep Semantic Abstractions of Everyday Human Activities – ROBOT ....... 71
  ■ 5 - Robust Natural Language Processing – IJCAI ........................................... 85

4 Visuospatial Abduction .......................................................................................... 95
  4.1 Abducing Visual Explanations ........................................................................... 96
  ■ 6 - Visual Explanation by High-Level Abduction – AAAI ................................. 99
  4.2 Integrated Abduction and Object-Level Motion ............................................. 109
  ■ 7 - Commonsense Visual Sensemaking for Autonomous Driving – AIJ ........ 113
  ■ 8 - Out of Sight But Not Out of Mind – IJCAI .................................................. 149

5 Learning Explainable Visuospatial Models .......................................................... 159
  5.1 Integrating Declarative Explainability and Inductive Generalization .............. 160
  ■ 9 - Semantic Analysis of (Reflectional) Visual Symmetry – ACS Journal ......... 165
  ■ 10 - Deeply Semantic Inductive Spatio-Temporal Learning – ILP .................. 187

6 Discussion and Outlook ......................................................................................... 197
  6.1 Results .............................................................................................................. 199
  6.2 Future Directions ............................................................................................. 201

Bibliography ............................................................................................................. 205

Appendices ............................................................................................................... 215
  A List of All Publications ..................................................................................... 217
  B Dissemination and Collaboration .................................................................... 221
    B.1 Tutorials ......................................................................................................... 221
    B.2 External Collaboration .................................................................................. 223
Select Publications Included in this Thesis

For a full list of publications see Appendix A.


# List of Figures

1. **Applied Scenarios**: Example domains addressed in the thesis. .......................................................... 2

   2.1 **Geometric Spatial Entities**: Select spatial primitives in (a) 2D, and (b) 3D. ........................................ 13

   2.2 **Space-Time Histories**: Spatio-temporal patterns and events. ......................................................... 14

   2.3 **Reasoning about Occlusion**: Changes in visibility as a result of, a) movement of the observer, and b) movement of the observed object. ................................................................. 15

   2.4 **A Cognitive Vision Framework for Semantic Sensemaking with Dynamic Visuospatial Imagery**: Conceptual overview. ................................................................. 16

   2.5 **Organization of the Thesis**: Conceptual overview and included publications. .................................. 18

   3.1 **Ontological Structure**: Semantically grounded (human) interactions. ........................................... 28

   3.2 **Semantic Scene Structure and Attention**: People movement and fixation clusters. ............................ 30

   3.3 **Relational Grounding of Everyday Activity**: “Making a cup of tea” (ego-centric view from a head-mounted RGB-D capture device) .................................................. 67

   3.4 **Spatio-Temporal Configuration of the Scene**: Spatial entities representing objects in the scene, spatial configuration, and space-time objects representing the dynamics of the scene. ................................................................. 68

   4.1 **Out of Sight but not out of Mind**: The case of hidden entities: e.g., an occluded cyclist. ......................... 96

   4.2 **Anticipatory Driving**: Anticipating lane change. .............................................................................. 111

   5.1 **Analysing Symmetry**: (a) symmetrical, and (b) non-symmetrical elements of an image. .................... 161

   5.2 **Symmetrical Spatial Configuration**: Examples for object level symmetry. ...................................... 162
## Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
<th>Page References</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI</td>
<td>Artificial Intelligence</td>
<td>1, 3, 12, 19, 23, 31, 45, 198, 199, 201, 202, 204</td>
</tr>
<tr>
<td>AOA</td>
<td>Area of Attention</td>
<td>30</td>
</tr>
<tr>
<td>ASP</td>
<td>Answer Set Programming</td>
<td>4, 6, 12, 16, 20, 95, 96, 98, 99, 109, 110, 112, 113, 149, 199, 200, 202</td>
</tr>
<tr>
<td>CLP</td>
<td>Constraint Logic Programming</td>
<td>4, 12, 16, 29, 199</td>
</tr>
<tr>
<td>DNN</td>
<td>Deep Neural Network</td>
<td>5, 6, 160, 165</td>
</tr>
<tr>
<td>ILP</td>
<td>Inductive Logic Programming</td>
<td>4, 12, 16, 21, 160, 162, 187, 199</td>
</tr>
<tr>
<td>IoU</td>
<td>Intersection over Union</td>
<td>109</td>
</tr>
<tr>
<td>KITTI</td>
<td>Karlsruhe Institute of Technology and Toyota Technological Institute</td>
<td>25, 109, 112, 113, 149</td>
</tr>
<tr>
<td>KR</td>
<td>Knowledge Representation and Reasoning</td>
<td>1, 3, 5, 6, 9, 10, 12, 16, 45, 96, 110, 160, 202, 203</td>
</tr>
<tr>
<td>MOT</td>
<td>Multi-Object Tracking</td>
<td>25, 99, 113, 149</td>
</tr>
<tr>
<td>MOTChallenge</td>
<td>Multi-Object Tracking Challenge</td>
<td>25, 98, 109, 112</td>
</tr>
<tr>
<td>OPRA</td>
<td>Oriented-Point Relation Algebra</td>
<td>13</td>
</tr>
<tr>
<td>QSTR</td>
<td>Qualitative Spatio-Temporal Representation and Reasoning</td>
<td>10, 11</td>
</tr>
<tr>
<td>RCC</td>
<td>Region Connection Calculus</td>
<td>13</td>
</tr>
<tr>
<td>RGB-D</td>
<td>combined video (RGB) and depth (D)</td>
<td>20, 67, 71, 200</td>
</tr>
</tbody>
</table>
List of Symbols

\( B \) Background Knowledge for Inductive Learning. 162
\( E \) Positive \((E^+)\) and Negative \((E^-)\) Training Examples. 162
\( H \) Induced Hypotheses on Spatio-Temporal Rules. 162, 163
\( T \) Time Points. 12, 18
\( \Delta \) Time Intervals. 12
\( \Sigma \) Theory about Space and Motion. 12, 15–17, 19, 20, 28, 66, 69, 97, 98, 110, 111, 160, 162, 165, 187, 199
\( \Sigma_{dyn} \) Spatio-Temporal Dynamics. 12, 14
\( \Sigma_{st} \) Spatio-Temporal Ontology. 12
\( \mathcal{E} \) Spatial Entities. 12–14, 16–18, 28, 160, 162, 163
\( \mathcal{H} \) Set of Abduced Hypotheses. 97, 98, 109–111
\( \mathcal{ML} \) Matching Likelihood between \( \mathcal{VO} \) and \( \mathcal{MT} \). 109–111
\( \mathcal{MT} \) Motion Tracks. 109, 111
\( \mathcal{O} \) Domain Objects. 12, 18, 28
\( \mathcal{P} \) Predictions for \( \mathcal{MT} \). 109, 110
\( \mathcal{R} \) Spatio-Temporal Relationships. 13, 14, 28, 162, 163
\( \mathcal{STH} \) Space Time History. 14, 66
\( \mathcal{VO} \) Visual Observations. 96–98, 109, 110
\( \Phi \) Fluents. 14, 20, 97
\( \Theta \) Events. 14, 20, 97
Chapter 1
Introduction

Computational systems designed to assist humans, be it in daily living and everyday tasks, or in professional situations, need to perceive and understand the environment they are situated in, make decisions, and interact / coordinate with humans and other agents, in order to achieve their goals. Being able to make sense of perceived scene dynamics and to connect low-level visual observations with high-level conceptual knowledge, pertaining to object motion, interactions, and events, is a central aspect of such systems. We posit that, for (multimodal) perceptual sensemaking general and systematic integration of (deep learning based) Computer Vision and semantics driven Artificial Intelligence (AI) methods, in particular from the field of Knowledge Representation and Reasoning (KR), is crucial. With this premise, the core technical focus of this thesis is on computational cognitive vision and perception, where commonsense characterisations of space and motion encompassing spatio-temporal relations, motion patterns, and events, facilitate semantic reasoning about perceived dynamics observed from visuospatial imagery.

1.1. Motivation

Perceptual sensemaking of visuospatial imagery is central to a range of cognitive assistive technologies and autonomous perception and interaction systems (Bhatt, 2013; Bhatt and Kersting, 2017; Bhatt and Suchan, 2020), including but not limited to applications in cognitive robotics, autonomous vehicles, or human behaviour research (Figure 1). Commonsense interpretation, explanation, and learning in these applications involve for instance semantic grounding and question answering, projection and interpolation of missing information, the ability to make default assumptions, or represent and reason about (perceived) events.

As an example consider the case of autonomous driving (Figure 1a), to be able to function in complex (urban) traffic situations, an autonomous vehicle has to behave in the environment and interact with other (human) traffic participants, including pedestrians, cyclists, etc. The ability to make default assumptions, e.g., pertaining to persistence of objects and / or object attributes, is an important capability for interpreting human behaviour, and for being able to project and interpolate missing information. For instance, take the case of an occluded object; here the system has to maintain knowledge about this object and has to anticipate where and when it will reappear, e.g., to ensure safety criteria.
I. Introduction

(a) Safety-Criticality in Autonomous Driving

(b) Visual Attention in Movies (Cognitive Film Studies)

(c) Dynamic Spatial Language

(d) Embodied (Human) Interaction

(e) Symmetry Perception in Visual Arts

(f) Human Behaviour Research (Real-World and VR)

Figure 1: Application Scenarios: Example application domains alluded to in this thesis.

Credits: stills in Figure 1.1 b) from the movie “The Bad Sleep Well”, directed and produced by Akira Kurosawa, Toho Studios, Japan, 1960; stills in Figure 1.1 e) from the movie “The Royal Tenenbaums”, directed by Wes Anderson, produced by Wes Anderson, Barry Mendel, and Scott Rudin, Touchstone Pictures and American Empirical Pictures, USA, 2001
Another example is in analysing how the **perception of the moving image** is guided by visuo-auditory cues (i.e., directors choices on cinematographic devices such as, camera movement, character placement, lighting, cuts, etc.). This involves commonsense scene understanding and semantic question-answering (e.g., with image, video, eye-tracking) based on semantically grounded abstractions of scene dynamics for answering questions about the spectators gaze vis-a-vis the moving image. For instance in the example in Figure 1b), an example for such a question could be how the attention of spectators is shifting when Shirai’s gaze is shifting from the bag to Nishi? Answering these kinds of questions involves being able to maintain consistent beliefs with respect to compositionality and indirect effects, space-time continuity, and spatial changes resulting from motion. Further it involves capabilities concerning event perception and interpretation, e.g., detecting events and event boundaries, and reason about their effects.

Further examples of reasoning capabilities relevant for perceptual sensemaking involve embodied grounding of high-level concepts pertaining to spatio-temporal dynamics, e.g., in **cognitive robotics**, to ground interactions with respect to experienced vision, touch, and bodily motion (Figure 1c and 1d); neurosymbolic modelling of visuospatial characteristics, e.g., in considerations about **aesthetics in visual arts**, for interpreting visual symmetry (Figure 1e); or analysing human wayfinding behaviour based on multisensory data coming from **visuo-locomotive human behaviour studies** (Figure 1f).

Addressing such challenges in view of (computational) cognitive vision and perception systems requires a systematic and general integration of **Semantics and Vision**, i.e., robust commonsense representation and reasoning about space-time dynamics on the one hand, and powerful low-level visual computing capabilities, e.g., for detecting scene objects and estimate / track their motion on the other. While computer vision research is approaching visual perception mainly bottom-up (and driven by deep learning) with a focus on low-level processing of (large amounts of) sensor data, **AI** research, and **KR** in particular, has focused on formal methods for high-level, semantics driven representation and reasoning about the world. Within this scope, **KR** has developed a large collection of tools and methods for semantic and commonsense reasoning (Davis, 2017) and within **AI** research, semantics, common sense, and explainability have been recognised to play a crucial role in the development of next generation **AI** systems (Davis and Marcus, 2015; Miller, 2019). Of particular relevance from the viewpoint of this thesis are cognitive and formal (computational) foundations for representing and reasoning about visuospatial dynamics. Towards this, research on Spatial Cognition (Denis, 2018; Nebel and Freksa, 2011) has studied human and artificial spatial representation and reasoning from an interdisciplinary perspective, involving cognitive science, linguistics, psychology, **AI**, and logic. For instance, Freksa (2004) investigates spatial cognition from the viewpoint of **AI** and relational spatial calculi, Krieg-Brückner et al. (2004) have developed a spatial ontology for (robot) navigation in the context of smart environments and ambient assistive living, and Bhatt and Freksa (2012) and Bhatt, Schultz, and Freksa (2013) have investigated spatial representation and reasoning in cognitive assistance systems and (architecture) design.
With the broader aim of bringing together these fields of research and bridging the gap between low-level visual processing and high-level spatio-temporal representation and reasoning, our research focuses on **deep visuospatial semantics**, where we develop declaratively grounded models of perceived interactions based on semantic knowledge and general formalisations of space and motion (i.e., *space, time, space-time*), and incorporate conceptual knowledge about (human) interactions. We present a general declarative theory of space and motion which constitutes the theoretical foundations for developing systematic tools and methods for deep visuospatial reasoning, i.e., involving reasoning capabilities pertaining to *semantic question-answering, visual abduction, and learning*, demonstrated within real-world applications.

### 1.2. Key Contributions

In the context of deep semantic perceptual sensemaking and declarative characterisations of space and motion, the publications included in this thesis (Pgs. xiii) are centred around the following three key contributions:

1. **Human-Centred Representation for Space and Motion**

   We develop a general, domain neutral theory of space and motion, declaratively modelled within *Constraint Logic Programming (CLP), Answer Set Programming (ASP)*, and *Inductive Logic Programming (ILP)*. The theory consists of ontological characterisations of human-centric relational representations that are semantically rooted to commonsense spatio-linguistic primitives pertaining to space and motion as they occur in natural language. Furthermore, the theory constitutes the theoretical foundations for systematic integration of vision and semantics, i.e., knowledge representation and reasoning methods with low-level (deep learning based) visual processing methods.

2. **Deep (Visuospatial) Semantics**

   The core technical contribution is on deep (visuospatial) semantic representations, and general methods and tools developed for declarative reasoning about space and motion with dynamic imagery, developed as a modular framework usable within hybrid architectures for perception & control. These include:

   - Semantically grounded characterisations of (embodied) interactions (e.g. human interactions in movies, human robot interactions, interactions of traffic participants in autonomous driving), rooted in CLP and ASP, and suited for *semantic sensemaking and question-answering* with multimodal human behavioural data.

   - A general method for *online* (i.e., incremental, realtime) **abductive visual explainability** jointly generating hypotheses of object interactions and corresponding motion tracks, based on hypotheses formation and optimisation in ASP and (deep learning based) visual computing.
A neurosymbolic pipeline for developing (declaratively) explainable interpretation models, integrating relational spatio-temporal structure and visual (neural) image features (e.g., visual similarity), extracted from Deep Neural Networks (DNN); together with a prototypical system for (inductive) learning of relational spatio-temporal structure founded in deep semantic representations.

3. Applications

From an applied viewpoint, the developed methodology serves as the technical foundation for assistive and analytical technologies. In this thesis we demonstrate applicability of the developed methods in diverse application domains (Section 2.3.2) in the context of autonomous perception & interaction systems, including autonomous vehicles, cognitive robotics, and cognitive assistance systems.

Aside from this, a particular focus is on interdisciplinary applications for empirical research in visual perception, and human behavioural studies, where cognitive vision systems provide the computational backbone for visuospatial analysis and facilitate large-scale and real world experiments. In this thesis we are highlighting the application for analysing visual perception in human behaviour studies, in particular focusing on cognitive film / media studies.

I.3. Organisation of the Thesis

This thesis is based on ten select, previously published journal and conference publications included within this thesis (A list of Select Publications Included in this Thesis can be found on Pg. xiii).

The remaining chapters are organised as follows: Chapter 2 discusses the overall aim and the broader context of the research presented in this thesis, concerning the integration of vision and semantics; summarises foundations and preliminaries for declarative reasoning about visuospatial dynamics; and relates the included publications to each other. Chapters 3 - 5 contain the select included publications and highlight reasoning capabilities developed within the context of visuospatial sensemaking, focusing on semantic interpretation and question-answering, abductive reasoning with visuospatial imagery, and learning of explainable visuospatial models. Finally, chapter 6 summarises results of this thesis and provides an outlook on future research directions.

In the following we briefly summarise Chapters 2 - 6:

Chapter 2: Space and Motion in Cognitive Vision Systems

In this chapter we are setting the context for the thesis and describe how the research is situated within preliminary work. In particular we are defining deep visuospatial semantics, and highlight our focus on cognitive vision, integrating KR and Vision. Further, we introduce space and motion for visual sensemaking and present a general theory for reasoning about visuospatial dynamics, based on declarative spatial reasoning. Finally
we put the select publications that are included in this thesis into context and describe how they are related to each other. Here we are in particular focusing on the three main areas, namely semantic question-answering, visual abduction, and inductive generalisation and learning.

**Chapter 3: Semantic Question-Answering with Video**

In this chapter we present visuospatial semantics for semantic question-answering with video, using declaratively grounded models of space and motion, and discuss their application for the interpretation of dynamic scenes and query answering in the domains of visual perception research and reasoning about embodied interaction in cognitive robotics. The chapter includes five publications focusing on declarative space and motion, semantic question-answering with video and eye-tracking for cognitive film studies, grounded models of human interaction for cognitive robotics, and spatial language processing for robot interaction.

**Chapter 4: Visuospatial Abduction**

In this chapter we summarise the concept of visuospatial abduction and discuss how visual perception can be modelled as an abductive process using hypotheses formation and optimisation in ASP. Further we present an online method for visual abduction, tightly integrating low-level visual processing and high-level semantic sensemaking. The chapter includes three publications focusing on visual abduction based on hypothesis formation and optimisation, and on joint abduction and object-level motion tracking for online processing in the context of autonomous driving.

**Chapter 5: Learning Explainable Visuospatial Models**

In this chapter we discuss how declarative semantics of space and motion facilitate learning of explainable visuospatial models. In particular we focus on declarative interpretation models of visuospatial characteristics integrating relational structure with visual features extracted from DNN, and inductive learning with qualitative space and motion for learning relational spatio-temporal structures. The chapter includes two publications, focusing on declarative interpretation models of visuospatial characteristics and on inductive (visuospatial) generalisation.

**Chapter 6: Discussion and Outlook**

Finally, we conclude by summarising the key results of the thesis and outline future research directions in the context of declarative reasoning about space and motion in visuospatial imagery, and also in cognitively inspired and KR based methods integrating with state of the art computer vision.
Chapter 2
Space and Motion
in Cognitive Vision Systems

We introduce (computational) cognitive vision in the context of visual intelligence and outline deep visuospatial semantics as alluded to in this thesis. Further we discuss preliminaries and put the included publications into context with respect to the overall framework for perceptual sense-making. We describe the theoretical considerations underlying commonsense visual intelligence; discuss preliminaries in qualitative and declarative spatial reasoning, and summarise the theory of space and motion developed within this thesis; finally, we describe the conceptual framework for perceptual sensemaking with dynamic visuospatial imagery and discuss how the select publications included in this thesis contribute to this framework.

This chapter summarises the intellectual ark of this thesis and is based on the publications originating from this thesis (Appendix A).
2.1. Commonsense Visual Intelligence

Commonsense knowledge (e.g., pertaining to space, motion, events) plays a central role in human understanding of visuospatial imagery and in making sense of perceived motion and interaction. With regard to computational visual perception systems, modelling such commonsense knowledge and developing computational systems capable of commonsense visuospatial sensemaking at a level of descriptive and analytical complexity that matches human performance and expectations is a long term research interest in Artificial Intelligence and Computer Vision research. For instance, Marr (1982) describes the motivation behind his approach to Computer Vision from an integrated perspective, including visual processing and semantic perceptual representation and sensemaking, as follows:

“The study of vision must [...] include not only the study of how to extract from images the various aspects of the world that are useful to us, but also an inquiry into the nature of the internal representations by which we capture this information and thus make it available as a basis for decisions about our thoughts and actions.”

From a psychological point of view, visual perception (Cavanagh, 2011) can be understood as a multi-layered process connecting low-level visual stimuli with high-level conceptual knowledge, e.g., pertaining to object dynamics and motion, events, causality, physics, functionality and affordances, etc. Based on these considerations, we approach Commonsense Visual Intelligence from the viewpoint of visuospatial sensemaking and human-centred representations of space and motion, with a broader focus on integrating Vision and Semantics.


Visual cognition is a central part of human intelligence and a key factor in a multitude of tasks humans have to face both in every day activities, as well as in highly specialised (professional) situations. Visuospatial abstraction and reasoning (Hegarty and Stull, 2012; Shah and Miyake, 2005; Tversky, 2008) is a key ability for perceptual sensemaking and for understanding visuospatial dynamics – e.g., being able to anticipate and simulate possible motion and interactions, to explain perceived changes in terms of object interactions, or to learn visuospatial concepts from perception – is a foundational cognitive ability for human interaction and decision-making. Research in the area of Cognitive Vision (Vernon, 2006, 2008) has focused on exploring how research in cognitive science may contribute towards achieving human level visual perception and visual intelligence, and thus help to close the gap between low-level visual precessing and high-level (cognitive) abilities pertaining to commonsense perceptual sensemaking. Vernon (2008) defines a cognitive vision system in terms of its capabilities as follows:

“A cognitive vision system should be able to engage in purposive goal-directed
From a methodological point of view, we approach cognitive vision research from an interdisciplinary angle, as follows (Bhatt and Suchan, 2020):

“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 premise of the research presented in this thesis is, that commonsense perceptual sensemaking with dynamic visuospatial imagery needs systematically developed general and modular tools and methods integrating high-level knowledge based semantics focused techniques developed within KR research, with low-level (neural) methods capable of computing primitive features of interest in visual data. More precisely, commonsense visual intelligence, as alluded to above, i.e., focusing on reasoning capabilities such as question-answering, abduction, and learning, requires representation and reasoning with human-centred abstractions of space, time, motion, actions, events and interaction (Bhatt, 2012; Bhatt, Guesgen, et al., 2011; Galton, 2000).

Deep Visuospatial Semantics. Deep semantics in the context of this thesis refer to declaratively grounded models of space and motion, based on qualitative representations of spatial, temporal, and spatio-temporal relations and motion patterns. From the viewpoint of cognitive vision systems, deep semantic representations of visuospatial dynamics provide the basis for abstracting and reasoning about perceived (human) interaction. In particular, deep visuospatial semantics (Bhatt, 2012; Bhatt, Guesgen, et al., 2011; Bhatt and Suchan, 2020; Bhatt, Suchan, and Schultz, 2013; Suchan and Bhatt, 2016a) denote:

“The existence of declaratively grounded models pertaining to space, time, spatio-time, motion, actions & events, spatio-linguistic conceptual knowledge and systematic formalisation supporting diverse computational capabilities encompassing question-answering, relational learning, visuospatial abduction, analogical inference, embodied grounding and simulation etc. Formal semantics and computational models of deep semantics manifest themselves in varied declarative methods such as constraint logic programming, inductive logic programming, and answer set programming (modulo theories).”

Such deep visuospatial semantic models may serve as an intermediate abstraction layer between human-understandable (natural language) representations and explanations, and low-level visuospatial imagery, i.e., object positions, motion trajectories, etc.

2.1.2. Neurosymbolism and Explainability

Explainable visual perception from a human-centred and commonsense reasoning point of view is concerned with developing visual perception systems with human-level vi-
sual sensemaking capabilities, i.e., the ability to represent and reason about conceptual symbolic commonsense knowledge (i.e., pertaining to space, time, motion, and events) and ground these concepts with respect to sub-symbolic visual features of involved image elements. In view of this thesis, our focus is on high-level interpretability and explainability, supporting commonsense, semantic reasoning with dynamic visuospatial imagery within declarative programming. This is founded on deep semantic representations, i.e., domain-independent, mixed qualitative-quantitative representation of visuospatial dynamics, as alluded to above. We argue that human-like visual perception systems require a tight integration of such commonsense representation and abstraction with deep learning based visual learning in a neurosymbolic manner. In particular this involves:

- **Commonsense Representations of Visuospatial Knowledge** consisting of KR based semantic abstraction of space, motion, dynamics, etc., where commonsense concepts of visuospatial phenomena, scene dynamics, and events and interactions are declaratively characterised by relational spatio-temporal models founded in general (human-centred) abstractions of space, time, and space-time.

- **Neural Representations of Image Elements and Visual Features** consisting of deep learning based methods for visual processing, including detection of visual scene elements (people, objects, environmental structure, image patches and regions), providing sub-symbolic representations of these elements, including visual features and visual characteristics (e.g. colour, texture, similarity, etc.).

In this context, deep visuospatial semantics can serve as a powerful abstraction mechanism to connect high-level conceptual knowledge with low-level deep learning based visual learning, providing a human-centred commonsense representation of visuospatial scene semantics supporting declarative question-answering, abductive explanation, and inductive generalisation and learning. Broadly speaking, we posit that both, knowledge representation and reasoning, and low-level neural feature learning are essential to realise computational visual intelligence.

### 2.2. Space, Time, Motion, and Events

Deep semantics for commonsense visuospatial reasoning (as presented in Section 2.1.1) are based on (human-centered) abstractions of space and motion capable of representing and reasoning about space, time, motion, and events. Towards this we develop a general and domain neutral theory of space and motion, based on spatio-temporal relations and patterns (Table 2.1; e.g. left-of, touching, part-of, during, approaching), rooted in research on Qualitative Spatio-Temporal Representation and Reasoning (QSTR) (Ligozat, 2013), and implemented in the context of declarative logic programming and declarative spatial reasoning. In the following we discuss preliminaries in QSTR and declarative spatial reasoning, and summarise the theory about space and motion.
2.2. Space, Time, Motion, and Events

2.2.1. Qualitative Spatio-Temporal Reasoning

Qualitative Spatio-Temporal Representation and Reasoning (QSTR) (Bhatt, Guesgen, et al., 2011; A. G. Cohn and Renz, 2008; Freksa, 1991b; Ligozat, 2013) abstracts from an exact numerical representation of space and time by describing the relations between objects using a finite number of symbols, i.e., a set of relations that hold between objects are used to describe a spatial configuration. In this context, qualitative spatial and temporal calculi are relational-algebraic systems pertaining to one or more aspects of space or time such as topology, orientation, direction, size (Ligozat, 2013). Of particular relevance from the viewpoint of this thesis is the work on spatio-temporal dynamics and motion, and in particular research about qualitative continuous spatial change. Within this area of research, Galton (1993, 1995, 2000) investigated movement on the basis of an integrated theory of space, time, objects, and position. Davis (2012) discusses the use of transition graphs for reasoning about continuous spatial change and applies them in physical reasoning problems. Freksa (1991a, 1992) introduced the conceptual neighborhoods to represent the continuity of spatial change, where relations between two entities are considered conceptual neighbors if they can be directly transformed from one relation into the other by continuous change of the scene. Muller (1998) defined continuous change using 4-dimensional regions in space-time. Hazarika (2005) and Hazarika and A. G. Cohn (2002) built on this work using an interval based approach to represent spatio-temporal primitives. Bhatt (2012) and Bhatt and Loke (2008) investigates reasoning about spatio-temporal dynamics in the context of commonsense reasoning about space, actions, and change.

Commonsense Spatial Reasoning. Computational (commonsense) spatial reasoning corresponds to the ability to declaratively specify and perform (mixed) geometric and qualitative visuospatial representation and reasoning pertaining to temporal, spatial, and spatio-temporal things, be it abstract regions of space and time, geometric entities and physical objects, or other spatial artefacts (Bhatt, Lee, and Schultz, 2011). Visual-spatial reasoning as alluded to in this thesis directly builds upon and extends foundations in (declarative) spatial reasoning with constraint logic programming (Bhatt, Lee, and Schultz, 2011), inductive logic programming (Suchan, Bhatt, and Schultz, 2016), and answer set programming (Wałęga, Bhatt, and Schultz, 2015; Wałęga, Schultz, and Bhatt, 2017). A key aspect of these works, also of developmental significance to this research, is that these methods enable mixed quantitative-qualitative commonsense reasoning encompassing inference patterns such as query-answering, non-monotonic spatial reasoning, and inductive generalisation with space, time and space-time. More broadly, the fact that general commonsense spatial reasoning is made possible directly as part of sys-
tematic (declarative) KR methods such as Constraint Logic Programming (CLP) (Jaffar and Maher, 1994), Answer Set Programming (ASP) (Brewka, Eiter, and Truszczyński, 2011), and Inductive Logic Programming (ILP) (Muggleton and Raedt, 1994) is also of significance towards building multi-faceted AI systems consisting of hybrid neurosymbolic methods.

2.2.2. Ontology and Formal Model

Commonsense abstractions of space and motion are central to this thesis and provide the formal foundations for deep visuospatial perceptual sensemaking as presented in Chapters 3 – 5. Within the publications originating from this thesis (Appendix A) we have developed and used a general and domain neutral theory about space and motion (Σ). In this section we summarise the main parts of this theory to provide a central point of reference.

The theory is implemented based on the foundations in declarative spatial reasoning, where spatial, temporal, and spatio-temporal entities are geometrically represented, e.g., as points, line-segments, polygons, etc. directly within logic programming. These declarative representations of spatio-temporal entities facilitate (mixed) geometric and qualitative visuospatial representation and reasoning, i.e., using spatio-temporal configurations, where the primitive entities describing the scene can possibly be grounded, partially grounded, or completely ungrounded. Further, these representations provide the ability to combine different aspects of space and different types of objects. Within this thesis these capabilities are central for representing different spatial aspects of visual elements extracted from the scene, and for being able to perform reasoning about hypothetical and conceptual spatial artefacts (without any real physical manifestation), e.g., hidden entities, shadows, etc.

From a formal point of view, the theory consists of spatio-temporal primitives (Σ_{st}) and spatio-temporal dynamics (Σ_{dyn}), and is defined as follows:

\[
\Sigma \equiv_{def} \Sigma_{dyn} < \Phi, \Theta > \cup \Sigma_{st} < O, E, T, \Delta, STH, R >
\]

Details of Σ_{st} and Σ_{dyn} are described in the following (I. and II.):

I. Σ_{st} – Qualitative Space and Motion

Space and motion is defined based on the relational spatio-temporal structure between domain objects and the changes within these relations.

► Space and Time – Relational Spatio-Temporal Structure. High-level, domain-dependent visual elements of the scene, e.g., people, objects, environmental structures are represented by domain-objects \( \mathcal{O} = \{o_1, o_2, ..., o_i\} \), elements in \( \mathcal{O} \) are geometrically interpreted as basic spatial entities (regions, points, line-segments, etc) denoted by \( \mathcal{E} = \{\varepsilon_1, \varepsilon_2, ..., \varepsilon_i\} \). The temporal aspects of the scene are represented by time points, denoted as \( T = \{t_1, ..., t_n\} \) and time intervals \( \Delta = \{\delta_{t_i, t_j}, ..., \delta_{t_n, t_m}\} \). The spatial configuration
2.2. Space, Time, Motion, and Events

![Geometric Spatial Entities](image)

**Figure 2.1: Geometric Spatial Entities:** Select spatial primitives in (a) 2D, and (b) 3D.

<table>
<thead>
<tr>
<th>SPATIO-TEMPORAL DOMAIN (QS)</th>
<th>Spatial, Time, Motion Relations (R)</th>
<th>Entities (E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mereotopology</td>
<td>disconnected (dc), external contact (ec), partial overlap (po), tangential proper part (tpp), non-tangential proper part (ntpp), proper part (pp), part of (p), discrete (dr), overlap (o), contact (c)</td>
<td>arbitrary rectangles, circles, polygons, cuboids, spheres</td>
</tr>
<tr>
<td>Incidence</td>
<td>interior, on boundary, exterior, discrete, intersects</td>
<td>2D point with rectangles, circles, polygons; 3D point with cuboids, spheres</td>
</tr>
<tr>
<td>Orientation</td>
<td>left, right, collinear, front, back, on, facing towards, facing away, same direction, opposite direction</td>
<td>2D point, circle, polygon with 2D line</td>
</tr>
<tr>
<td>Distance, Size</td>
<td>adjacent, near, far, smaller, equi-sized, larger</td>
<td>rectangles, circles, polygons, cuboids, spheres</td>
</tr>
<tr>
<td>Motion</td>
<td>moving: towards, away, parallel; growing / shrinking: vertically, horizontally; splitting / merging; rotation: left, right, up, down, clockwise, counterclockwise</td>
<td>rectangles, circles, polygons, cuboids, spheres</td>
</tr>
<tr>
<td>Time</td>
<td>before, after, meets, overlaps, starts, during, finishes, equals</td>
<td>time-points, time intervals</td>
</tr>
</tbody>
</table>

Table 2.1: Commonsense relations for abstract representation of space and motion.

**Credits:** Table reprinted from (Suchan, Bhatt, and Varadarajan, 2021)

of the scene and changes thereof are then characterised based on n-ary spatio-temporal relations \( R = \{r_1, r_2, ..., r_n\} \) of a particular logic of space / time, holding between basic entities in \( E \) (Figure 2.1) at a particular time. In this context, spatio-temporal relations in \( R \) are characterised with respect to arbitrary spatial and spatio-temporal domains such as mereotopology, orientation, distance, size, motion (Table 2.1), including established spatial abstraction calculi, such as the Region Connection Calculus (RCC) (Randell, Cui, and A. Cohn, 1992), Rectangle Algebra and Block Algebra (Guesgen, 1989), LR Calculus (Scivos and Nebel, 2004), Oriented-Point Relation Algebra (OPRA) (Moratz, 2006), and Allen’s Interval Algebra (Allen, 1983).
Motion – Qualitative Spatial Change. Spatial relations ($R$) holding between entities ($E$) change as a result of motion of the involved entities. Such spatio-temporal dynamics are represented using Space-Time Histories ($STH$) (Hazarika, 2005; Muller, 1998), i.e., objects with an extent in space and time. The space-time history representing the movement of an entity in the scene is defined by the occurrences of the object, i.e., $\text{sth}(o, \delta) = (\varepsilon_{t_1}, \varepsilon_{t_2}, \varepsilon_{t_3}, ..., \varepsilon_{t_n})$, where $\varepsilon_{t_i}$ to $\varepsilon_{t_n}$ denote the spatial primitives representing the object $o$ at the time points $t_i$ to $t_n$. Continuous movement of objects is described by the changes in the spatial properties of the $\text{sth}$ between consecutive time points $\varepsilon_{t_i}$ and $\varepsilon_{t_{i+1}}$. Dynamic spatio-temporal relations (involving one or more objects), such as approaching, growth, shrinkage, turning, accelerating, etc are defined by making qualitative distinctions on object movement based on changes in the objects position, size, orientation, direction, and the distance between objects, as well as on the rate of these changes. Complex movement patterns and motion characteristics (Figure 2.2) are defined by combining different spatio-temporal aspects, e.g., two objects moving parallel to each other, or one object passing in front or behind another object.

II. $\Sigma_{dyn}$ – Scene Dynamics

Dynamic properties of the scene are characterised based on changes in the (spatio-temporal) properties of the scene objects and the events causing this change. To this end Fluents $\Phi = \{\phi_1, ..., \phi_n\}$ are used to describe properties of the scene, i.e. the predicates $\text{holds-at}(\phi, t)$ and $\text{holds-in}(\phi, \delta)$ denote that the fluent $\phi$ holds at time point $t$, resp. in time interval $\delta$. Spatio-temporal fluents denote that a relation $r \in R$ holds between basic spatial entities $\varepsilon$, e.g., of a space-time history, at a time point $t$. Events $\Theta = \{\theta_1, \theta_2, ..., \theta_i\}$ describe processes that change the value of the fluents, i.e., the spatio-temporal configuration of objects in the scene at a specific time point. These are defined in terms of the involved motion (i.e., movement patterns and motion characteristics, changing the

![Figure 2.2: Space-Time Histories: Spatio-temporal patterns and events.](image)
2.2. Space, Time, Motion, and Events

Example I. Reasoning about Occlusion

Occlusion may be modelled using fluents representing the visibility state of an object and events changing its visibility state. The functional fluent visibility($o_i$) denoting the visibility of object $o_i$ may take the values \{fully_visible, partially_visible, not_visible\}. To represent that the object $o_i$ hides behind another object $o_j$, we define the event hides_behind/2, changing the value of the visibility fluent to not_visible, as follows:

$$\text{causesValue(hides\_behind}(o_i, o_j), \text{visibility}(o_i), \text{not\_visible}, t).$$  \hspace{1cm} (2.1)

Further we define preconditions that have to be true for the event hides_behind/2 to happen at a time point $t$, based on the visibility and the motion of the involved objects with respect to the observer.

$$\text{poss(hides\_behind}(o_i, o_j), t_i) \supset$$

$$\text{holds-}\text{at(visibility}(o_i), \text{partially\_visible}, t_j) \wedge \text{holds-}\text{in(approaching}(o_i, o_j), \delta_i)\wedge$$

$$\text{meets}(t_j, t_i) \wedge \text{meets}(\delta_i, t_i).$$  \hspace{1cm} (2.2)

Such events may be combined with other events, motion patterns, and spatio-temporal relations to define more complex interactions, e.g., an object passing behind another object, or an object getting occluded because of a moving observer (Figure 2.3). Further, such definitions may be used for reasoning, e.g., in the case of perspective taking for reasoning about object position and visibility with multiple observers.

Figure 2.3: Reasoning about Occlusion: Changes in visibility as a result of, a) movement of the observer, and b) movement of the observed object.

spatio-temporal configuration of domain-objects) and changes in the scene properties caused by the event. We use \text{occurs-}\text{at}(\theta, t), and \text{occurs-}\text{in}(\theta, \delta) to denote that an event $\theta$ occurred at a time point $t$ or in a time interval $\delta$. For reasoning about the scene dynamics these definitions may be directly integrated with formalisms for reasoning about action and change, such as Event Calculus (Kowalski and Sergot, 1989), or Situation Calculus (Lin, 2008; McCarthy, 1963).

The presented theory about space and motion ($\Sigma$) provides the foundation to represent and reason about object interactions, by combining motion patterns, events, and beliefs to define complex events, as depicted in the characterisation of occlusion in Example 1.
2.3. Perceptual Sensemaking with Dynamic Visuospatial Imagery

Reasoning about visuospatial imagery in real world applications necessitates a systematic integration of the reasoning capabilities within larger systems involving sensing, reasoning, interpretation, decision making, and control, under the consideration of real-world conditions, such as the need for online and real-time processing, the presence of uncertainty and sensor noise, etc. Towards this, we integrated the domain neutral Theory about Space and Motion ($\Sigma$, Section 2.2.2) within a hybrid framework combining commonsense representation and reasoning about space and motion with visual processing, resulting in declaratively grounded models of scene dynamics, systematically implemented within mainstream KR methods, i.e., CLP, ASP, and ILP.

The framework (Figure 2.4) includes the different modules and the entire pipeline needed for deep semantic analysis of visuospatial imagery (consisting of components C1 - C3).

### C1 Visuospatial Imagery

Visuospatial imagery (obtained from sensor data as applicable to applications in Section 2.3.2) is processed to detect and track visual elements (e.g. people, objects, motion), using state of the art (deep learning based) computer vision methods (Visual Processing, Pg. 17). The extracted elements are representing the visual structure of the input data and are interpreted as spatial entities ($\mathcal{E}$), i.e., points, line-segments, regions, space-time histories, etc. for further analysis.
Within this thesis visual processing is applied for extracting and analysing scene elements (i.e., people, body-structure, and objects in the scene) and motion (i.e., object motion and scene motion), encompassing methods for:

- **Image Classification** and **Feature Learning** – based on Big Data, (e.g., ImageNet (J. Deng et al., 2009; Russakovsky et al., 2015)), using neural network architectures such as AlexNets (Krizhevsky, Sutskever, and Hinton, 2012), VGG (Simonyan and Zisserman, 2015), or ResNet (Kaiming He et al., 2016).

- **Detection**, i.e., of **people and objects** (Bochkovskiy, Wang, and Liao, 2020; Redmon, Divvala, et al., 2016; Redmon and Farhadi, 2018; Ren et al., 2017), and **faces** (Jiankang Deng et al., 2020; Hu and Ramanan, 2017).

- **Pose Estimation**, i.e., of body pose (Cao et al., 2019) (including fine grained hand pose), face and gaze analysis (Baltrusaitis et al., 2018).

- **Segmentation**, i.e., semantic segmentation (Chen et al., 2018) and instance segmentation (K. He et al., 2020).

- **Motion Analysis**, i.e., optical flow based motion estimation (Ilg et al., 2017) and movement tracking (Bergmann, Meinhardt, and Leal-Taixé, 2019; Bewley et al., 2016).

These methods are used to analyse the visual imagery and extract the “Geometry of a Scene” consisting of geometrical objects representing the scene structure.

### C2 Deep Semantics of Space and Motion

Declarative characterisations of events and movement patterns are used to represent the spatio-temporal dynamics of domain objects, based on the declarative theory of space and motion ($\Sigma$) consisting of qualitative spatial and temporal relations ($R$) holding between basic entities ($E$). Further, this component provides general reasoning capabilities pertaining to semantic grounding, abductive explainability, and inductive generalization (these reasoning capabilities constitute the technical focus of this thesis and are discussed in more detail in Section 2.3.1).

### C3 Declarative Scene Structure

The scene structure is analysed based on the semantic interpretation and sensemaking capabilities rooted in qualitative reasoning about space, motion, and events. Further, these capabilities are made available for applications involving semantic analysis of dynamic visuospatial imagery, e.g., for query answering, decision-making, complex data visualisation, summarisation, etc.
For reasoning and semantic sensemaking with dynamic visuospatial imagery, objects in the scene are represented based on their spatial and temporal properties, and abstracted using spatial entities ($E$), such as regions, points, oriented points, line segments as per needs. For instance, a person may be represented as a point representing its location, as a rectangle using its bounding box, or as a polygon representing the perceived shape of the object. Basic entities are extracted directly from the sensor data and characterised by numerical values using functions mapping from an object to a spatial entity with respect to a particular spatial modality, i.e., the spatial entity representing the object $o$ at a time point $t$, entity: $O \times T \rightarrow E$; or the space-time history representing the object $o$ in a time interval $\delta$, sth: $O \rightarrow E \times T$.

2.3.1. Methods and Tools (Included Publications)

Reasoning about visuospatial dynamics in the context of the cognitive vision framework (and, as being investigated in this thesis), pertains to general tools and methods for semantic question-answering, visuospatial abduction, and declarative explainability and inductive generalisation. This thesis contains ten peer-reviewed articles published in conferences and journals, which are focusing on these three main research directions, respectively.

In the following we provide an overview of the conducted research within the respective research components (R1 – R3) and position the included publications with respect to the conceptual framework (Figure 2.5).
Semantic Question-Answering (Chapter 3)

Semantic interpretation and question-answering with visuospatial imagery is based on declaratively grounded models of (embodied) visuospatial semantics. In this context, the theory of space and motion ($\Sigma$) is used to characterise rich (human) interactions with respect to their relational spatio-temporal structure, i.e., the spatial, temporal, and spatio-temporal relations involved within the interaction. These characterisations are declaratively grounded within observed spatio-temporal primitives obtained from state of the art computer vision methods, e.g., for people and object detection, motion estimation, etc., and provide queryable declarative structures of scene dynamics suitable for semantic query answering.

Semantic interpretation and question-answering is covered in five publications included in this thesis, encompassing the theoretical foundations for deep semantic characterisation of visuospatial interactions and their applications in the context of visual perception studies focusing on the reception of the moving image, and embodied interactions in cognitive robotics:

- [6], Pg. 33 (Suchan, 2017), presenting the broader perspective and formal characterisation of declarative space and motion;
- [7], Pg. 45 (Suchan and Bhatt, 2016a), focusing on declarative space and motion for semantic question-answering with video and eye-tracking;
- [8], Pg. 55 (Suchan and Bhatt, 2016b), presenting semantic question-answering about the perception of video from the viewpoint of applied computer vision;
- [9], Pg. 71 (Suchan and Bhatt, 2017b), developing a declarative characterisation of embodied human interactions; and
- [10], Pg. 85 (Spranger, Suchan, and Bhatt, 2016), exploring the use of spatio-temporal relations as an underlying representation for natural language processing.

[6] (Suchan, 2017) is a journal article published in the German Journal of Artificial Intelligence (KI), partly based on (Suchan and Bhatt, 2016a,b; Suchan, Bhatt, and Santos, 2014). It presents a broader perspective on the formal characterisation of space and motion that can be used to represent and reason about spatio-temporal dynamics in visual imagery. In [7] (Suchan and Bhatt, 2016a) and [8] (Suchan and Bhatt, 2016b) we apply the declaratively grounded models for semantic question-answering about the perception of movies, encompassing scene semantics and eye-tracking data from spectators. In particular, [8] (Suchan and Bhatt, 2016b) has a focus on the application of computer vision, used to extract scene elements and semantic structure of the scene, and its application for semantic sensemaking for cognitive film studies, and [7] (Suchan and Bhatt, 2016a) discusses semantic sensemaking for cognitive film studies from the perspective of declarative AI and knowledge representation and reasoning, and the use of AI to assist visual perception research.
Space and Motion in Cognitive Vision Systems

4 (Suchan and Bhatt, 2017b) and 5 (Spranger, Suchan, and Bhatt, 2016) focus on declarative (visuospatial) reasoning about embodied interaction in two different settings. 4 (Suchan and Bhatt, 2017b) explores the use of declarative semantics and embodied grounding in cognitive robotics and everyday activities and develops a declarative characterisation of embodied human interactions involving a declarative model of the human body and the representational means to ground human object interactions within human skeleton data obtained from combined video (RGB) and depth (D) (RGB-D) sensing. Aspects of this work are also presented in (Suchan and Bhatt, 2017a).

Further, 5 (Spranger, Suchan, and Bhatt, 2016) explores embodied interactions in the context of space and motion in natural language, focusing on declarative abstractions of space and motion for natural language interactions of robots. Preliminary considerations on declarative spatio-temporal representations for natural language processing are published in (Spranger, Suchan, Bhatt, and Eppe, 2014).

Visuospatial Abduction

Visuospatial abduction denotes reasoning from visual observations to explanations, consisting of spatio-temporal belief (states) and events, grounded in low-level (object) motion. In particular, this means hypothesising relational spatio-temporal structures, pertaining to visual phenomena, such as object persistence, occlusion, etc., that are represented by object interactions and motion patterns, and explain the perceived object motion. Towards this, we present a general method for explaining visuospatial observations by integrating low-level visual processing and high-level abductive reasoning in the backdrop of ASP, where we generate hypotheses on abducible events (Θ) and beliefs (Φ) that are characterised based on the theory about space and motion (Σ), and that explain the perceived spatio-temporal changes in the environment, e.g., the disappearance of an object. The resulting set of hypotheses is then ranked based on the abduced event sequence and the corresponding object motion.

Visual abduction for computing declarative visual explanations is subject of three publications included in this thesis:

- 6, Pg. 99 (Suchan, Bhatt, Wałęga, et al., 2018), introducing the general concept of visual abduction;
- 7, Pg. 113 (Suchan, Bhatt, and Varadarajan, 2021), integrating abduction and object-level motion for online visual explanation in autonomous driving; and
- 8, Pg. 149 (Suchan, Bhatt, and Varadarajan, 2019), introducing online visual abduction for autonomous driving.

The concept of visual abduction is introduced in 6 (Suchan, Bhatt, Wałęga, et al., 2018), where we present the general method for generating visual explanations driven by ASP and based on hypotheses formation and optimisation. In 8 (Suchan, Bhatt,
2.3. Perceptual Sensemaking with Dynamic Visuospatial Imagery

and Varadarajan, 2019) and 7 (Suchan, Bhatt, and Varadarajan, 2021) we are then extending this method to integrate abduction with object-level motion for jointly abducting event sequences and tracking scene dynamics, which facilitates iterative solving and online (real-time) execution. 8 (Suchan, Bhatt, and Varadarajan, 2019) is a conference contribution published at IJCAI 2019 and recognised by the distinguished paper award honorable mention. It develops a general method for online visual abduction and demonstrates it in the context of safety-critical situations in autonomous driving. 7 (Suchan, Bhatt, and Varadarajan, 2021) is an invited, peer-reviewed journal publication, which elaborates on the developed method in 8 (Suchan, Bhatt, and Varadarajan, 2019) and provides additional examples demonstrating visual abduction in human-centred autonomous driving.

R3 Declarative Explainability and Inductive Generalisation (Chapter 5)

(Explainable) visuospatial interpretation is based on declarative formalisations of visuospatial image characteristics grounded in (symbolic and sub-symbolic) image elements and (neural) image features. Image characteristics (e.g., symmetry) are declaratively modelled based on the configuration of image elements and their visual features. The model is then used for explainable interpretation of image characteristics based on the divergence of the image structure from the model, i.e., differences in the spatio-temporal configuration of image elements (e.g. size, position, etc.) and the visual features of these elements (e.g., visual similarity, etc.). In the context of (inductive) generalisation and learning such interpretation models constitute the basis for learning relational visuospatial structures by integrating the declarative characterisation into the ILP framework. Furthermore, the incremental learning process of (neural) visual features 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), and facilitate (neurosymbolic) visuospatial learning (Bhatt, Suchan, and Varadarajan, 2019). In this respect, this thesis contains two publications:

- 9, Pg. 165 (Suchan, Bhatt, Vardarajan, et al., 2018), focusing on declarative explainability of image characteristics; and
- 10, Pg. 187 (Suchan, Bhatt, and Schultz, 2016), presenting spatio-temporal structure learning based on inductive generalisation.

In 9 (Suchan, Bhatt, Vardarajan, et al., 2018) we develop (declaratively) explainable interpretation models in a neurosymbolic pipeline. We demonstrate these models for the case of (reflectional) visuospatial symmetry in visual arts to characterise symmetrical image structure and to learn subjective symmetry measures based on human data. In 10 (Suchan, Bhatt, and Schultz, 2016), we present a general framework for relational spatio-temporal (inductive) learning, implemented within the framework of ILP and demonstrate its application for learning spatio-temporal rules on symmetrical configuration of image elements, directly from visual imagery.
2.3.2. Applications and Interdisciplinarity

The Cognitive Vision framework described above is demonstrated in the backdrop of diverse application domains, which serve as a testbed for developed methods and tools, and are used for demonstration and evaluation purposes. A particular emphasis in this thesis is on **real-world applications** where the processing and semantic interpretation of large amounts of dynamic visuospatial imagery is central, and where human-centred considerations play a role, including applications in the area of **autonomous systems**, and **assistive technology**. Aside from the core technical considerations, these applications also constitute the basis for **interdisciplinary research**, focusing on **human-centred empirical research**, and **cognitive studies of human behaviour**. These interdisciplinary aspects have been partly conducted in the context of (international) collaborations we have engaged in during the course of this thesis (see Appendix B.2).

In the following we briefly summarise the application areas **A1 – A3** that are of interest from the viewpoint of this thesis and that serve as the basis for application examples in Chapters 3 – 5. In the context of behavioural studies and visual perception research (**A3**), examples in this thesis are focusing on studies of the moving image, and visual arts and aesthetics. Studies on driving and human-environment interaction are included for completeness, given their significance (Bhatt, Suchan, Schultz, et al., 2016; Suchan, Bhatt, and Varadarajan, 2020) as a potential area of application. However, these lines of application are not addressed or further developed in this thesis.

**A1  Autonomous Vehicles**

Research on autonomous vehicles focuses of developing robust methods for perception and control designed to function in real-world situations, i.e., capable of real-time processing, handle noise and unexpected situations, etc. From the viewpoint of this thesis, we approach cognitive vision and perception for autonomous driving from a human-centred perspective, where the interaction between the car and its environment, e.g. other traffic participants, pedestrians and cyclists, etc., are of interest. Here, visuospatial sensemaking and reasoning about space and motion is applied in the context of semantic visual interpretation and explanation. In particular, commonsense characterisations of dynamic visuospatial phenomena are used to track and explain scene dynamics and to provide a declarative basis for interpreting safety-critical situations in driving.

From an interdisciplinary viewpoint this research also relates to empirical studies of driving behaviour, as detailed in **A3**.

**A2  Commonsense Cognitive Robotics**

Cognitive robotics focus on human-centred and cognitively inspired methods and tools targeted at cognitive abilities, such as embodied grounding, commonsense abstraction and reasoning, learning by demonstration or description, natural (human-)interaction. In this context we focus on embodied interactions, i.e., for spatio-temporal grounding of
human-object interactions in everyday activities, and for dynamic spatial representation and reasoning in natural language (human-robot / robot-robot) interactions. Here declaratively characterised image schematic representations are used to provide an abstraction for the semantic interpretation of visuospatial imagery, directly obtained from the robots perception systems (e.g., video, RGB-D, point-cloud data, etc.). In this context embodied abstractions of space and motion, as developed within this thesis, provide a representational middle-layer for reasoning about perceived motion events and for generating and parsing of dynamic spatial language.


An interdisciplinary focus of the work in this thesis is on empirical human behaviour research, targeted from an interdisciplinary viewpoint, involving behavioural studies, cognitive science, psychology, and AI. Within this thesis and connected to our research on cognitive vision and perception, we are developing assistive technology to support analysis of human behaviour and visual perception data, in particular within the context of natural “in-the-wild” stimuli, i.e., experiments conducted in (possibly unconstraint) real-world situations. Areas of significance for this thesis are the reception of visuo-auditory media, symmetry in the context of visual aesthetics and art, visual complexity and multimodal interaction in driving, and visual search in navigation tasks.

**Moving Image.** In the context of analysing human interaction with media, we are focusing on (eye-tracking driven) visual perception of visuo-auditory media (e.g., narrative film), where declarative perceptual sensemaking capabilities constitute the basis for empirical research concerned with the human reception of the moving image. The technical focus herein is on deep semantic analysis and semantic question-answering with visual attention (as recorded within large-scale eye-tracking experiments) vis-a-vis visuo-auditory computational narrative structure (i.e., the geometry of a scene (Suchan and Bhatt, 2016b)). This is applied for cognitive film studies, e.g. concerning the use of symmetry in the editing of Wes Anderson (Suchan, Bhatt, and Yu, 2016).

**Visual Arts and Aesthetics.** Perception and reception of visual arts and aesthetics differ a lot between individuals as a result of individual differences in perception and interpretation and the socio-cultural background of the perceiver. Within this thesis we use visuospatial symmetry as an example application for developing computational cognitive models of visual perception, where declarative interpretation models of image characteristics are used to analyse and learn subjective factors in the perception of (reflectional) symmetry. Here, neurosymbolic models of image characteristics aiming at human-centered interpretability and explainability provide an abstraction layer for characterising symmetry as a multi-layered perceptual phenomenon, involving spatial organization, visual features, semantics, and individual differences.

**Driving.** An interdisciplinary focus of our work on (autonomous) driving is on behavioural studies concerning empirical (eye-tracking based) studies of visual percep-
tion while driving (Kondyli, Bhatt, and Suchan, 2020), i.e. focusing on visual complexity, and on multimodal interaction between traffic participants in typical (urban) traffic situations. Such studies may also provide a basis for evaluating visuospatial sensemaking methods for autonomous vehicles.

**Human-Environment Interaction.** Environment-behaviour research is focusing on empirical studies on how humans interact with the environment, e.g., during wayfinding tasks, in emergency situations, etc. Here, deep visuospatial characterisations of human interactions with the environment are used for visuo-locomotive human behaviour analysis (e.g., concerning aspects such as user behaviour during wayfinding / navigation in large-scale builtup environments), based on data collected during large-scale behavioural studies, e.g., (mobile) eye-tracking, egocentric video, floor plan data, etc. (Bhatt, Suchan, Kondyli, et al., 2016; Bhatt, Suchan, Schultz, et al., 2016).

### 2.3.3. Datasets (Developed and/or Used)

For the examples and evaluation in this thesis we are using the following datasets (D1 – D4). Datasets D1 and D2 were developed in the course of this thesis; datasets D3 and D4 are community established benchmark datasets which serve as evaluation benchmarks for multi-object tracking. In the following we provide a brief description of the respective datasets:

**D1. Movie Dataset** (developed in this thesis)

The Movie Dataset (Suchan and Bhatt, 2016a) consists of 16 select scenes from 12 films with corresponding eye-tracking data of spectators, collected in a large-scale eye-tracking experiment with a total of 31 subjects. Each scene ranges between 0 : 38 minute to a max. of 9 : 44 minutes in duration. Most of the scenes involve multiple moving objects and moving camera(s). In addition to the eye-tracking data, the dataset contains high quality annotations of a large range of objects in the scenes, which are directly usable for evaluating object tracking and can also be used for interpreting interactions between people and objects.

**D2. Subjective Symmetry Perception** (developed in this thesis)

The Symmetry Dataset (Suchan, Bhatt, Vardarajan, et al., 2018) is based on human-generated data from subjective, qualitative assessments of symmetry. It includes 150 images consisting of landscape and architectural photography, and movie scenes. The images range from highly symmetric images showing very controlled symmetric patterns to completely non symmetric images. We presented 50 images to each participant, selected randomly from the dataset and the participants had to rank the images by selecting one of four categories: not_symmetric, somewhat_symmetric, symmetric, and highly_symmetric. Each image was presented to approximately 100 participants, and the symmetry value is calculated as the average of all responses.
D3. KITTI Tracking Dataset

The Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) Vision Benchmark Suite (Geiger, Lenz, Stiller, et al., 2013) is a real-world benchmark dataset for autonomous driving, it is collected from a driving car equipped with cameras, laser scanners, and GPS localisation. In addition to the raw sensor data, it provides benchmarks for different perception tasks, including highly accurate annotations. In this thesis we are using the KITTI object tracking dataset (Geiger, Lenz, and Urtasun, 2012), which focuses on multi-object tracking in traffic situations, filmed from the egocentric perspective of a moving car. It consists of 21 training and 29 test scenes, and provides accurate track annotations for 8 object classes (e.g., car, pedestrian, van, cyclist).

D4. MOT Challenge

The Multi-Object Tracking Challenge (MOTChallenge) (Dendorfer et al., 2020; Milan et al., 2016) is a collection of datasets, providing a benchmark for Multi-Object Tracking (MOT). It consists of a large number of datasets with challenging situations, common detections for the datasets, and highly accurate annotations of object tracks. Furthermore, it provides an evaluation framework for measuring the performance of multi-object tracking approaches. Within this thesis we use the MOT 2017 dataset (Milan et al., 2016) consisting of 7 training and 7 test scenes which are highly unconstrained videos, filmed with both static and moving cameras, and the MOT 2020 dataset (Dendorfer et al., 2020) consisting of 4 training and 4 test scenes filmed in crowded environments.
Chapter 3
Semantic Q/A with Video

We present the foundations for semantic question answering with visuospatial imagery based on declaratively grounded models of visuospatial dynamics. We develop visuospatial semantics consisting of declarative characterisations of dynamic interactions, grounded in the visual structure of the imagery, as obtained from sensor data and extracted using computer vision methods. In this context, semantic question-answering is demonstrated in the area of cognitive film studies for analysing spectators eye-tracking behaviour with respect to the structure of the scene. Further, we discuss the application of declarative representations of space and motion to characterise (embodied) interactions in cognitive robotics, i.e., for modelling everyday human activities in human robot interaction and for perceptual abstraction and sensemaking in natural language interactions.

Included Publications:


3. Semantic Q/A with Video

3.1. Visuospatial Semantics

Semantic interpretation and question-answering with visuospatial imagery is based on declarative abstractions of space and motion. To this end, (embodied) visuospatial semantics of the scene, i.e., pertaining to space, time, motion, and events, are declaratively grounded within its visual structure. Here, the theory of space and motion (Σ, Section 2.2.2) is used to model rich (human) interactions with respect to their relational spatio-temporal characteristics, i.e., the motion patterns and events involved in the interaction. In this context, domain objects (O) represent the visuospatial elements of the scene, e.g., people and objects, motion, etc., extracted from the visual imagery using state of the art computer vision methods. These elements are then abstracted based on their spatial and temporal properties using spatial entities in E, e.g., regions, points, line-segments, etc. The spatio-temporal configuration of the scene and changes within it are characterised using qualitative spatio-temporal relationships (R) between these entities.

Figure 3.1 presents a conceptual overview of the multiple levels of abstraction necessary to model the relational structure of (embodied) interactions from visuospatial imagery. These consist of (L1 – L3):

L1. the **multimodal visuo-auditory features** characterising (human) interaction in a domain-independent manner, including and / or pertaining to people, objects, gazing direction, body pose, visual fixation, sound and speech, environmental structure and affordances etc;
3.1. Visuospatial Semantics

L2. the spatio-temporal movement patterns characterising the dynamics of (human) interaction using relational primitives of space, time, and motion (also in a fully domain-neutral manner);

L3. the high-level visuospatial concepts (e.g., pertaining to activities, or specific behavioural phenomena) that are applicable for perceptual sensemaking from the viewpoint of the specific domains of interest being investigated.

Within the framework for visuospatial sensemaking (Section 2.3), multi-level semantic characterisations (pertaining to abstraction layers L1 – L3) describing spatio-temporal dynamics may be formulated within logic programming, i.e., building on CLP, to reason about semantics of visuospatial dynamics in applications where computational systems have to be able to query and make sense of (human) interactions. Examples include domains such as scene and movie understanding, cognitive robotics, human behaviour analysis, or smart environments.

Declarative Semantics for Cognitive Media Studies. As a testbed for semantic interpretation and question-answering we apply the developed framework in the context of empirical research concerned with the human reception of films, where visual question-answering is demonstrated in support of visual perception research in Cognitive Film Studies. The main applied focus herein is on empirical studies on the influence of cinematographic aids such as cuts, long takes, the use of symmetry, etc., on spectators attention whilst watching a movie. In this context, declarative semantics serve as a basis for developing assistive technology facilitating the semantic analysis of spectators eye-tracking behaviour vis-a-vis the moving image. To facilitate this analysis, relational characterisations of visual perception are developed and declaratively grounded with respect to primitive spatio-temporal entities and motion patterns, pertaining to the structure of the scene and the dynamics of spectators gaze, i.e., obtained from a computational pipeline for visual processing and gaze analysis. In particular we extract geometric entities representing the following scene elements and perceptual artefacts:

- Structural elements of the scene, e.g., object / character identity and placement, shots and shot types, categories of camera movement, extracted from the movie scene using computer vision methods and constituting the “Geometry of a Scene”.

- Perceptual artefacts, e.g., visual fixation, saccadic movements, gaze patterns, pertaining to the perception and reception of the moving image. These are obtained based on spectators attention available from eye-tracking data collected in visual perception experiments.

Semantic sensemaking with declarative characterisations of the scene structure and visual attention is demonstrated in the context of a large scale visual perception experiment conducted with the Movie Dataset (see Section 2.3.3, D1.), where we collected gaze data from 31 participants on 16 diverse movie scenes. In this context declarative scene semantics are used to answer questions about the reception of the movie scenes, such as the following (Q1 – Q3):
Example 2. SEMANTIC Q/A WITH VIDEO AND EYE-TRACKING

Considering the depiction of the sample movie scene in Figure 3.2, here scene dynamics are represented using space-time histories of people tracks and spectators eye-tracking data aggregated to Areas of Attention (AOA). Based on this one could, for instance, formulate a query to determine what happened when the AOA following Jack and Francis merged?

```
?- Int = interval(_ , _), TP = timepoint(_),
| sth(jack, STH_jack), sth(francis, STH_francis),
| sth(attention(AOA_1), STH_AOA_1), sth(attention(AOA_2), STH_AOA_2),
| occurs_in(attn_following(STH_AOA_1, STH_jack), _),
| occurs_in(attn_following(STH_AOA_2, STH_francis), _),
| occurs_at(merge([STH_AOA_1, STH_AOA_2], _), TP),
| occurs_in(Obs, Int), time(TP, Int, during).
```

results in:

```
Obs = approaching(francis, jack),
Int = interval(25, 30),
TP = 28;
... 
```

I.e., Francis is approaching Jack when the respective AOA merge.

Q1. How does the spectators gaze shift when a new character is entering the scene / how long does it take till spectators attention is on the newly appearing character?

Q2. How are spectators following people / object movement during a long shot / camera movement?

Q3. How is spectators gaze affected by spatio-temporal configuration of objects and actors in the scene?

For instance, consider the semantic question-answering and resulting declarative models presented in Example 2. Such models may provide a basis for analysing the correla-
tion between scene structure and eye-movements, and serve as an intermediate semantic abstraction layer for summarising and externalising the dynamics of visual attention vis-a-vis the moving image. Moreover, the queryable declarative structures of scene dynamics may form the basis for externalisation, e.g., natural language summarisation, statistical analysis, complex data visualisation, etc. (Suchan, Bhatt, and Jhavar, 2015).

INCLUDED PUBLICATIONS: Semantic question-answering with video was published in the following three publications:

- A journal article published in the *German Journal of Artificial Intelligence* (KI) (Suchan, 2017, 1), partly based on previously published work (Suchan and Bhatt, 2016a,b; Suchan, Bhatt, and Santos, 2014). It focuses on declarative reasoning about visuospatial dynamics with video, based on the general, domain neutral theory about space and motion.

- A conference paper published in the *International Joint Conference on Artificial Intelligence* (IJCAI 2016) (Suchan and Bhatt, 2016a, 2), targeted at the AI community and focusing on declarative modelling of scene semantics for assistive technology in the context of empirical research on cognitive film studies.

- A conference paper published in the *Winter Conference on Applications of Computer Vision* (WACV 2016) (Suchan and Bhatt, 2016b, 3) aiming at computer vision research and discussing the application of computer vision foundations for declarative reasoning about the “Geometry of a Scene” and its reception based on eye-tracking data.

*Copies of the above stated publications in the given order follow in the next sections.*
I - Declarative Reasoning about Space and Motion with Video

Published in:
The German Journal of Artificial Intelligence (KI - Künstliche Intelligenz)

Citation:

This journal paper was published in the German Journal of Artificial Intelligence (KI) and is partly based on previously published work (Suchan and Bhatt, 2016a,b; Suchan, Bhatt, and Santos, 2014). It presents the commonsense theory of space and motion for representing and reasoning about motion patterns in video data, and to perform declarative (deep) semantic interpretation of visuospatial sensor data, e.g., coming from object tracking, eye-tracking data, movement trajectories, etc. The paper defines qualitative spatio-temporal relations, relations on motion, and motion patterns as part of the theory. Furthermore, it describes how these relations and patterns are used to define high-level actions, events, and interactions.

The application of the domain independent theory is discussed in a range of domains, in which the capability to semantically interpret motion in visuospatial data is central. In particular, we are showcasing examples from the area of cognitive film studies for analysing visual perception of spectators by integrating the visual structure of a scene and spectators gaze, acquired from eye-tracking experiments.
Declarative Reasoning about Space and Motion with Video

Jakob Suchan

Received: date / Accepted: date

Abstract We present a commonsense theory of space and motion for representing and reasoning about motion patterns in video data, to perform declarative (deep) semantic interpretation of visuo-spatial sensor data, e.g., coming from object tracking, eye tracking data, movement trajectories. The theory has been implemented within Constraint Logic Programming to support integration into large scale AI projects.

The theory is domain independent and has been applied in a range of domains, in which the capability to semantically interpret motion in visuo-spatial data is central. In this paper, we demonstrate its capabilities in the context of Cognitive Film Studies for analysing visual perception of spectators by integrating the visual structure of a scene and spectators gaze acquired from eye tracking experiments.

Keywords Cognitive Vision · Spatio-temporal Dynamics · Semantic Interpretation of Video

1 Introduction

Semantic interpretation of dynamic visuo-spatial imagery is central to a broad range of applications where computational systems have to make sense of human interactions, such as scene and movie understanding, cognitive robotics, human behaviour analysis, or smart environments. In this paper we focus on declarative methods and tools for high-level semantic interpretation of human behaviour from dynamic visuo-spatial data, such as video or RGB-D, obtained from visual sensing devices [Bhatt et al., 2013b; Suchan et al., 2014]. As an example consider the scene depicted in Fig. 1, showing a person taking a slice of bread in the course of making a sandwich. A natural language description of the (spatial) interactions involved in this activity could be the following.

The person reaches for the bread, picks up a slice of bread, and moves the hand together with the bread back.

To make sense of the interactions in this example, a system has to be able to interpret the spatio-temporal dynamics involved in the interactions, i.e. the spatio-
Jakob Suchan

Fig. 2: Cinematographic Scene Semantics., Grand Budapest Hotel by Wes Anderson
temporal and conceptual semantics of the perceived scene.

In this context, we present a general theory of Space and Motion for the high-level semantic interpretation of dynamic scenes based on space-time histories, that can be used to reason about such spatio-temporal dynamics using movement patterns, and the events and mutual interactions that accrue in this context. The theory has been implemented within declarative logic programming based on formalisations of space and spatial reasoning within constraint logic programming (CLP(QS)) [Bhatt et al., 2011; Schultz and Bhatt, 2014]. As such, the framework can directly be integrated with standard methods from knowledge representation and reasoning (KR) (e.g., using Prolog) to provide declarative spatial reasoning about spatio-temporal dynamics for large scale AI projects. The broader focus of our research is on deep visuo-spatial semantics [Suchan and Bhatt, 2016a], where we develop declaratively grounded models of perceived interactions based on semantic knowledge and general formalisations of space and motion (i.e., space, time, space-time), and incorporate conceptual knowledge about human interactions. These models can be used for reasoning (e.g. explaining observed events) and question answering about the perceived interactions, relational learning of grounded event models and conceptual knowledge, or embodied grounding and simulation etc.

In this paper we present a general theory of space and motion for declarative reasoning about spatio-temporal dynamics and demonstrate its capabilities by utilising it in a larger framework for deep semantic analysis of (subject’s) attention in the context of visuo-spatial perception focused cognitive film studies. Commonsense Space and Motion is used for (deep) semantic interpretation and qualitative analysis of visual perception data, including peoples gaze data and the semantic structure of a scene. We use visual processing to extract the semantic structure of a scene (Fig. 2), e.g., consisting of cinematographic devices such as cuts, camera movements, character placement and object motion, and combine it with visual attention data, obtained from eye tracking experiments, for individual spectators as well as grouped areas of attention based on gaze data from multiple subjects.

2 Related Work

There is an increased interest in the computer vision community to synergise with cognitively motivated methods [Aloimonos and Fermüller, 2015], in particular relevant to this paper is the research on semantic interpretation of visual imagery, e.g., for combining information from video analysis with textual information for understanding events and answering queries about video data [Tu et al., 2014], and perceptual grounding and inference [Yu et al., 2015; Yang et al., 2015; Zamponiannis et al., 2015]. In the context of semantic interpretation of movies, Rohrbach et al. [2017] investigates generating natural language descriptions from movies, Rohrbach et al. [2016] investigates the recognition of fine grained interactions from video. Probabilistic graphical models and probabilistic logics have been used for recognising high-level activities from videos, E.g., Tran and Davis [2008] present an approach to the use of spatio-temporal relations within first-order probabilistic logic for the analysis of video sequences from a parking lot. Song et al. [2013] present a general framework for recognising events in RGB-D data using probabilistic first-order logic in the context of kitchen activities. Research in the area of Cognitive Vision [Vernon, 2008, 2006; Cohn et al., 2006] has focused on enhancing computer vision systems with cognitive abilities. Within this line of work qualita-
tive spatio-temporal reasoning was investigated, e.g., for event modelling [Dubba et al., 2010; Sridhar et al., 2011] and event learning using interleaved abductive-inductive reasoning [Dubba et al., 2014], or grounding natural language in vision data for robot interaction tasks [Al-Omari et al., 2016]. From the perspective of cognitive systems, semantic sense making of visuo-spatial imagery requires the representation and reasoning within logic programming. In this paper, the high-level semantic interpretation and qualitative analysis of visuo-spatial imagery requires the representational and inferential mediation of (declarative) qualitative abstractions of the spatio-temporal dynamics, encompassing space, time, motion, and interaction. We use a first-order typed language (\(\mathcal{L}\)) with the following alphabet: \(\{\neg, \land, \lor, \forall, \exists, \supset, \equiv\}\) (respectively meaning negation, conjunction, disjunction, universal quantification, existential quantification, implication, and equivalence).

3 A Theory of Space and Motion

The high-level semantic interpretation and qualitative analysis of visuo-spatial imagery requires the representational and inferential mediation of (declarative) qualitative abstractions of the spatio-temporal dynamics, encompassing space, time, motion, and interaction. We use a first-order typed language (\(\mathcal{L}\)) with the following alphabet: \(\{\neg, \land, \lor, \forall, \exists, \supset, \equiv\}\) (respectively meaning negation, conjunction, disjunction, universal quantification, existential quantification, implication, and equivalence).

3.1 Space – Qualitative Spatial Relations

Spatio-temporal relationships (\(\mathcal{R}\)) between the basic entities in \(\mathcal{E}\) are characterised with respect to arbitrary spatial and spatio-temporal domains such as mererotopology, orientation, distance, size, motion.

Objects in the scene are represented by their spatial and temporal properties, and abstracted with spatial entities such as regions, points, oriented points, line segments as per needs. E.g., a person in the video data can be represented as a point, representing its location, as a rectangle using its bounding box, or as a polygon, representing the perceived shape of the object. Basic entities are extracted from directly from the sensor data and characterised by numerical values using functions mapping from a spatial object and a time point to a particular spatial entity, entity: \(\mathcal{O} \times \mathcal{T} \rightarrow \mathcal{E}\) as appropriate for the spatial properties of the objects e.g. its position as a point, or the region as a bounding box.

- points are a pair of reals \(x, y\),
- oriented-points consists of a point \(p\) and a vector \(v\),
- line-segments consists of two points \(p_1, p_2\) denoting the start and the end point of the line-segment,
- axis-aligned rectangles consists a point \(p\) and its width and height \(w, h\),
- polygons consists of list of points \(p_1, \ldots, p_n\) defining the boundary of the polygon.

The spatial configuration of objects in the scene is represented by basic spatial entities (regions, points etc): \(\mathcal{E} = \{\epsilon_1, \epsilon_2, \ldots, \epsilon_i\}\); and \(n\)-ary spatial relations \(\mathcal{R} = \{r_1, r_2, \ldots, r_n\}\) of a particular logic of space / time. The relations include established spatial abstraction calculi implemented in the declarative spatial reasoning system CLP(QS) Bhatt et al. [2011], such as the topological relations of the RCC8 fragment of the Region Connection Calculus [Randell et al., 1992], \(\mathcal{R}_{top} \equiv \{\text{dc, ec, po, eq, tpp, ntp, tpp}^{-1}, \text{ntpp}^{-1}\}\) or orientation relations of the \(\mathcal{LR}\) calculus [Scivos and Nebel, 2004] \(\mathcal{R}_{orient} \equiv \{\ldots\}^\).
Further spatial calculi include the Rectangle Algebra and Block Algebra [Guesgen, 1989], and the Oriented-Point Relation Algebra (OPRA) [Moratz, 2006]. However, the reasoning framework supports a broad range of spatial abstractions and can easily be extended based on the needs of the application.

The spatio-temporal configuration of objects is represented using spatial fluents describing the spatial relations holding between objects at a particular point in time. Towards this, we define $\Phi = \{ \phi_1, \phi_2, ..., \phi_n \}$ as a set of propositional and functional fluents, e.g. $\phi(\varepsilon_1, \varepsilon_2)$ denotes the spatial relationship between $\varepsilon_1$ and $\varepsilon_2$. Predicates holds-at($\phi(r,t)$ and holds-in($\phi, r, \delta$) are used to denote that the fluent $\phi$ has the value $r$ at time $t$, resp. during a time interval $\delta$. Allen’s interval algebra [Allen, 1983] $\mathcal{R}_{\text{time}} \equiv \{ \text{before, after, during, contains, starts, started by, finishes, finished by, overlaps, overlapped by, meets, met by, equal} \}$ for representing temporal aspects.

3.2 Motion – Qualitative Spatial Change

Spatial relations holding between objects change as a result of motion of the individuals in the scene (see Fig. 6). These spatio-temporal dynamics are represented by Space-Time Histories [Muller, 1998; Hazarika and Cohn, 2002] (depicted in Fig. 5), i.e. objects with an extend in space and time. The space-time history ($sth$) representing the movement of an object in the scene is defined by the occurrences of the object and the motion between these occurrences, i.e., the $sth$ of an object $o$ is given by the function $sth: \mathcal{O} \mapsto \mathcal{E} \times \mathcal{T}$, which maps the object to its appearance in space and time. $sth(o, \delta) = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)$, where $\varepsilon_1$ to $\varepsilon_n$ denotes the spatial primitive representing the object $o$ at the time points $t_1$ to $t_n$. Space-time histories serve as basic primitives to represent and reason about the spatio-temporal dynamics in a perceived scene, by defining movement patterns (dynamic spatio-temporal relations), and actions and events, based on the perceived object movement. As such, continuous spatial change interpreted as movement becomes available as a first class object within the theory.

Continuous movement of objects is described by the changes in the spatial properties of the $sth$ between consecutive time points $\varepsilon_i$ and $\varepsilon_{i+1}$. We define qualitative relations on object movement based on the objects position, size, and the distance between objects.

- position $x, y$ position of an objects;
- size width and height $w, h$ of an object;
- distance the distance between two objects.

Dynamic spatio-temporal relations are defined by making qualitative on distinctions these parameters.

Changes in Position Movement of objects is defined based on changes in the objects positions.

$$\text{holds-in}(\text{moving}(o), \text{true}, \delta) \supset \text{during}(t_i, \delta) \land \text{during}(t_j, \delta) \land \text{before}(t_i, t_j) \land (\text{position}(o, t_i) \neq \text{position}(o, t_j)).$$

(1)

Changes in Distance Relative Movement of objects, such as approaching and receding, is defined based on
on changes in distance between objects. E.g. *approaching* is defined as follows:

\[
\begin{align*}
\text{holds-in(approaching}(o, o_j), \text{true}, \delta) & \supset \text{during}(t_1, \delta) \land \text{during}(t_2, \delta) \land \\
& \text{before}(t_1, t_2) \land \text{distance}(o, o_j, t_1) > \text{distance}(o, o_j, t_2).
\end{align*}
\]

(3)

\[
\begin{align*}
\text{holds-in(receding}(o, o_j), \text{true}, \delta) & \supset \text{during}(t_1, \delta) \land \text{during}(t_2, \delta) \land \\
& \text{before}(t_1, t_2) \land \text{distance}(o, o_j, t_1) < \text{distance}(o, o_j, t_2).
\end{align*}
\]

(4)

\[
\begin{align*}
\text{holds-in(same_distance}(o, o_j), \text{true}, \delta) & \supset \text{during}(t_1, \delta) \land \text{during}(t_2, \delta) \land \\
& \text{before}(t_1, t_2) \land \text{distance}(o, o_j, t_1) = \text{distance}(o, o_j, t_2).
\end{align*}
\]

(5)

### Changes in Size

Accordingly, *growth* and *shrinkage* of an object is defined based on the changes in size of an object, in one or more dimensions.

\[
\begin{align*}
\text{holds-in(growing}(o), \delta) & \supset \text{during}(t_1, \delta) \land \text{during}(t_2, \delta) \land \\
& \text{before}(t_1, t_2) \land \text{size}(o, t_1) < \text{size}(o, t_2).
\end{align*}
\]

(6)

\[
\begin{align*}
\text{holds-in(shrinking}(o), \delta) & \supset \text{during}(t_1, \delta) \land \text{during}(t_2, \delta) \land \\
& \text{before}(t_1, t_2) \land \text{size}(o, t_1) > \text{size}(o, t_2).
\end{align*}
\]

(7)

\[
\begin{align*}
\text{holds-in(same_size}(o), \delta) & \supset \text{during}(t_1, \delta) \land \text{during}(t_2, \delta) \land \\
& \text{before}(t_1, t_2) \land \text{size}(o, t_1) = \text{size}(o, t_2).
\end{align*}
\]

(8)

### Complex movement patterns

These are defined by combining different spatio-temporal aspects, e.g. a pattern describing that two objects are moving parallel to each other could then be defined as follows:

\[
\begin{align*}
\text{holds-in(parallel}(o, o_j), \text{true}, \delta) & \supset \\
\text{holds-in(same_distance}(o, o_j), \text{true}, \delta) & \supset \\
\text{holds-in(xspace}(o, o_j), \delta).
\end{align*}
\]

(9)

### 3.3 Actions, Events, and Interactions

For representing and reasoning about interactions of people and objects we define actions and events, using the predicate *occurs-in*(\(\theta, \delta\)) denotes that an event \(\theta\) occurred in a time interval \(\delta\).

**Movement Events** defined on qualitative space and motion describe changes in the spatio-temporal configuration of the perceived scene as result of a particular movement. These events are defined by changes in the spatio-temporal relations holding between the moving object and the other objects in the scene.

**Appearance and Disappearance** describes the cases where the existence status of an object changes, i.e. the time point, where the \(s\)th starts to exists, resp. ends to exist.

\[
\begin{align*}
\text{occurs-in(apppearance}(o), \delta) & \supset \\
& \text{starts}(t_1, \delta) \land \text{finishes}(t_2, \delta) \land \text{meets}(t_1, t_2) \land \\
& \text{holds-at(exists}(o), \text{true}, t_1) \land \text{holds-at(exists}(o), \text{false}, t_2).
\end{align*}
\]

(10)

\[
\begin{align*}
\text{occurs-in(disappearance}(o), \delta) & \supset \\
& \text{starts}(t_1, \delta) \land \text{finishes}(t_2, \delta) \land \text{meets}(t_1, t_2) \land \\
& \text{holds-at(exists}(o), \text{false}, t_1) \land \text{holds-at(exists}(o), \text{true}, t_2).
\end{align*}
\]

(11)

**Object Interactions** The event *moves_into*, describing that something is moving into a region (e.g. representing some container) is defined as follows.

\[
\begin{align*}
\text{occurs-in(moves_into}(o_1, o_2), \delta) & \supset \\
& \text{holds-at}(\text{position}(o_1), \text{region}(o_2)), \text{outside}, t_1) \land \\
& \text{holds-at}(\text{position}(o_1), \text{region}(o_2)), \text{inside}, t_2) \land \\
& \text{holds}(\text{time}(t_1, \delta), \text{starts}) \land \text{holds}(\text{time}(t_2, \delta), \text{finishes}) \land \\
& \text{holds}(\text{time}(t_1, t_2), \text{meets}).
\end{align*}
\]

(12)

\[
\begin{align*}
\text{occurs-in(moves_out}(o_1, o_2), \delta) & \supset \\
& \text{holds-at}(\text{position}(o_1), \text{region}(o_2)), \text{inside}, t_1) \land \\
& \text{holds-at}(\text{position}(o_1), \text{region}(o_2)), \text{outside}, t_2) \land \\
& \text{holds}(\text{time}(t_1, \delta), \text{starts}) \land \text{holds}(\text{time}(t_2, \delta), \text{finishes}) \land \\
& \text{holds}(\text{time}(t_1, t_2), \text{meets}).
\end{align*}
\]

(13)
Complex interactions, e.g. a person passing in front, or behind another person, or a person passing between two persons, can be described by combining multiple actions and events.

**Event semantics** can be used as a basis for narrative level reasoning about perceived events, i.e. for explanation of spatio-temporal observations, or prediction and anticipation of events, e.g. in dynamic control tasks such as decision making in robot control, autonomous driving, or smart environments. In this context preconditions and effects of events can be used for reasoning about action and change based on the perceived events.

4 Deep Semantics for Multi-Modal Human Behaviour Analysis

The theory about *space and motion* presented above has been implemented declaratively within constraint logic programming. For deep semantic interpretation of human behaviour data, we integrated the theory in a larger framework combining commonsense space and motion with visual processing. The proposed system (fig. 7) includes the different modules and the entire pipeline needed for deep semantic analysis of human behaviour data (consisting of the components C1 - C3).

**C1) Multi-Modal Human Behaviour Data.** Visuo-spatial imagery is processed to detect and track visual elements, e.g. people, objects, motion. These extracted elements are representing the visual structure of the input data and are interpreted as Space-Time Histories for further analysis.

**C2) Spatio-Temporal Dynamics.** The declarative theory of space and motion consists of qualitative spatial and temporal relations and defines events and movement pattern representing the spatio-temporal dynamics of the data.

**C3) Deep Semantics of Human Behaviour.** This component leads to the semantic interpretation and sense-making of the spatio-temporal observations by qualitative reasoning about space, action, and change, and makes these capabilities available for applications involving semantic analysis of human behaviour data.

Within this framework, it is possible to define high-level rules and execute queries in the logic programming language PROLOG to reason about multi-modal human behaviour data, however, the complete logical reasoning engine of PROLOG (i.e., also our space-time history extensions implemented in PROLOG) can be embedded as a reasoning component within larger projects where semantic sense-making of human interactions and declarative reasoning about spatio-temporal dynamics is central. I.e. we have applied the theory in the context of:

- *Cognitive Film Studies* for analysing spectators perception of the moving image [Suchan and Bhatt, 2016a,b];
- *Visuo-Locomotive Experience* for analysing wayfinding behaviour and the influence of architectural artefacts, such as landmarks or signage [Bhatt et al., 2016a,b];
– *Dynamic Control & Decision Making* for supporting decision-making in dynamic control tasks [Suchan et al., 2014; Bhatt et al., 2013a];

– *Grounded Language Learning for Cognitive Robotics* for generating and parsing dynamic spatial phrases in robotic interactions [Spranger et al., 2016, 2014].

Besides these domains, declarative reasoning about spatio-temporal dynamics is central to a range of applications concerned with interpreting visuo-spatial dynamics, such as human-robot interaction, smart environments, or autonomous driving. In those domains, deep semantics may be used for dynamically interpreting and reacting to human behaviour by integrating general formalisations of space and motion with conceptual knowledge about human behaviour and evidence from human behaviour studies, e.g. for understanding and adequately reacting to human interactions or anticipating driving behaviour. Towards this ground models of human behaviour combining high-level event semantics with low-level movement patterns can be used to predict / anticipate low-level motion based on high-level event sequences, which may be obtained from commonsense reasoning about action and change. Grounded models of human behaviour may be learned based on evidence from human behaviour studies, using the theory of space and motion for inductive learning of spatio-temporal dynamics with the Aleph ILP system [Srinivasan, 2001] to learn logical rules from examples. E.g., axioms of perception can be learned by the spatio-temporal dynamics of the gaze data and the corresponding facts about the structure of the scene [Suchan et al., 2016]. Further, in many of these applications it is important to handle uncertainty resulting from noisy sensor data or from ambiguities due to incomplete knowledge. The presented theory is concerned with “crisp” qualitative abstractions of spatio-temporal dynamics, however, to handle noise in the data or to represent uncertainties in the interpretation of the data, it may be integrated with stochastic methods for generating beliefs, e.g., by using probabilistic graphical models [Suchan et al., 2014] or probabilistic logics [Schultz et al., 2016].

4.1 The Dynamics of Spectator Attention in Films

As an exemplary application for the theory of space and motion presented in this paper, we investigate studies in visual perception of films, computationally analysing the perceptual experience of spectators when watching a movie. Towards this we combine the spatio-temporal regions of attention with perceptual elements from the films, such as *shot types, camera movement, people and object movement*, for analysing the effects of these cinematographic aids on the perceptual experience of the spectators. The semantic model of the perception of the moving image is based on:

– **Geometry of a Scene.** content-level deep semantic analysis of scene structure and semantics — object / character identity and placement, shots and shot types, categories of camera movement — pertaining to the moving image, obtained from low-level visual processing algorithms founded in state-of-the-art outcomes for computer vision research on detection and analysis of people, objects, and motion [Suchan and Bhatt, 2016b].

– **Perception & Reception** visual perception analysis of spectator behaviour and engagement with the medium, e.g., visual fixation on film characters, gaze patterns co-related with influence of cinematographic aids such as cuts, long takes, symmetry on attention and whilst watching a film, available from eye tracking data from visual perception experiments.

The spatio-temporal objects obtained from visual processing and the eye tracking study form the basis for semantic analysis of the spatio-temporal dynamics of spectators attention. Consider the following instances aimed at reflecting the kinds of Q/A capabilities that might be needed from the viewpoint of cognitive film studies:

**Q1.** how is the spectator attention shifting, when the camera is moving / after a cut / during a long shot?

**Q2.** which movement / characters / objects is the spectators attention following in a spatio-temporal sense?

**Q3.** are there individual or aggregate regularities with respect to the shift in spectator attention at a certain time?
Our space-time history model and its integration with low-level visual processing supports such question-answering based on the visual analysis of the scene and the eye movement data. Looking at the space-time history of the aggregated area of attention of all participants, the system is able to answer queries concerning the focus of the attention of the spectators and also involving people and objects in the scene, e.g. at which times was the attention on a certain character.

**The Darjeeling Limited (2007). Dir: Wes Anderson.** As an example consider the scene from The Darjeeling Limited (2007) as depicted in Fig. 8 The following attention predicate is true if the space-time history of an object is topologically connected, i.e. *inside* or *overlapping*, with the space-time history of attention.

```
atts_on(Obj, Int) :- sth(Obj, ST_Obj),
                  holds_in(inside(ST_Obj, ST_AOA), Int);  
                  holds_in(overlapping(ST_AOA, ST_Obj), Int).
```

Given the above rule, a query where the spatio-temporal history of the character Jack is compared with the aggregated Area of Attention of all participants would be the following:

```
?- Int = interval(_,_), atts_on(jack, Int).
```

The query results in all time intervals during which spectator attention is on the character Jack:

```
Int = interval(5, 30); ...
```

Further, one could formulate a query to determine what happened when the areas of attention following Jack and Francis merged:

```
?- Int = interval(_,_), TP = timepoint(1),
     sth(jack, st_jack), sth(francis, st_francis),
     atts_following(ST_AOA_1, st_jack, Int),
     atts_following(ST_AOA_2, st_francis, Int),
     occurs_at(merge([ST_AOA_1, ST_AOA_2], _), TP),
     occurs_in(Obj, Int), time(TP, Int, during).
```

The result of the query is that Francis is approaching Jack when the respective areas of attention Merge:

```
Obs = approaching(st_francis, st_jack),
TP = 28; ...
```

Similarly one could use the model for analysing the effects of the scene structure and cinematographic devices on spectators gazing patterns.

**Drive (2011). Dir: Nicolas Winding Refn** As an example for these kinds of queries, consider the scene of the movie Drive (2011) (see Fig. 9). The domain-specific input data for this scene is the following:

```
% Basic scene structure (scene, shots, and cuts)
at(scene(scene1), true, interval(0, 1214)).
at(scene(scene2), true, interval(603, 1214)).
in(shot(shot1), true, interval(0, 602)).
in(shot(shot2), true, interval(603, 1214)).
```

To semantically analyse the cinematographic characteristics of a scene using film analysis techniques, e.g., one could formulate a query to reflect the shifting attention between the left and the right half of the frame while the camera is tracking Irene and the Driver.

```
?- I = interval(_, _), I1 = interval(_, _),
    tracking(can, ['Irene', 'The Driver'], I1),
    sth(gaze(Spectator), ST_GP),
    inside(ST_GP, quadrant(half(Half)), I1),
    time(during, I, I1).
```

The answer to this is a sequence of attention shifts between the right and the left half of the quadrant system.

```
Spectator = subject(s1),
I = interval(639, 646),
Half = left;
Spectator = subject(s1),
I = interval(647, 746),
Half = right;
Spectator = subject(s1),
I = interval(747, 833),
Half = left;
Spectator = subject(s1),
I = interval(834, 894),
Half = right;
false.
```

This way, semantic question answering becomes possible within spatio-temporal entities of visual attention as well as domains-specific perceptual elements within the
scene; both categories exist as generic first class objects within our logic programming framework.

5 Conclusion and Outlook

We have presented a general theory about space and motion aimed at (deep) semantic interpretation and qualitative analytics driven visuo-spatial computing (e.g., in visual perception studies). The proposed framework has been fully modelled and implemented declaratively within constraint logic programming (CLP). We emphasize that the level of declarativeness within logic programming is such that each aspect pertaining to the overall framework can be seamlessly customised and elaborated, and that question-answering & query can be performed using the primitives of the theory—things, space and motion, events, perceptual objects (e.g., people, objects, cinematographic aids, eye-tracking / gaze points, regions of attention etc)—as native objects within CLP. The theory has been implemented in a modular way, such that it can be easily extended and integrated with conceptual knowledge and cognitive theories, e.g. implementing image schematic representations [Mandler and PagCvas, 2014] based on spatio-temporal dynamics or integrating deep semantics with the theory of conceptual spaces [Gärdenfors, 2000]. Broadly, this work is driven by a tighter integration of AI with the state of the art in computer vision, in this context, we believe that high-level semantics based on knowledge representation and reasoning and AI are crucial to develop systems that are able to interpret and analyse dynamic visuo-spatial data, and in particular to make sense of human behaviour based on multi-modal sensor data. In this line of work our long term goal is towards human level semantic understanding of human interactions and involved spatio-temporal dynamics, including general concepts and patterns utilised by humans to understand perceived dynamics.

References


2 - Semantic Q/A with Video and Eye-Tracking Data

Published in: 
International Joint Conference on Artificial Intelligence (IJCAI 2016)

Citation:

This conference paper was published in the International Joint Conference on Artificial Intelligence (IJCAI 2016) and is targeted at the AI community. It discusses semantic question-answering as an assistive technology in visual perception research, focusing on cognitive film studies. The key focus is on deep visuospatial semantics integrating visual computing and KR foundations for the development of human-centred assistive technology supporting high-level interpretation and qualitative analysis, involving video and eye-tracking based attention. The paper develops deep semantic models of space and motion (implemented in constraint logic programming) for semantic inference and query answering with scene dynamics and attention, based on a visual processing pipeline for extracting visual scene structure from video.

Semantic question-answering is demonstrated in the context of a large-scale experiment concerned with analysis of visual perception and reception of the moving image. This involves high-level analysis of eye-tracking based attention, i.e., fixations, saccadic movements, gaze patterns, with respect to the dynamics of the scene.
Semantic Question-Answering with Video and Eye-Tracking Data:
AI Foundations for Human Visual Perception Driven Cognitive Film Studies

Jakob Suchan and Mehul Bhatt
Human-Centred Cognitive Assistance Lab. (hcc.uni-bremen.de)
Cognitive Vision (www.cognitive-vision.org)
University of Bremen, Germany
{jsuchan, bhatt}@cognitive-vision.org

Abstract
We present a computational framework for the grounding and semantic interpretation of dynamic visuo-spatial imagery consisting of video and eye-tracking data. Driven by cognitive film studies and visual perception research, we demonstrate key technological capabilities aimed at investigating attention & recipient effects vis-a-vis the motion picture; this encompasses high-level analysis of subject’s visual fixation patterns and correlating this with (deep) semantic analysis of the dynamic visual data (e.g., fixation on movie characters, influence of cinematographic devices such as cuts). The framework and its application as a general AI-based assistive technology platform —integrating vision & KR— for cognitive film studies is highlighted.

1 Introduction
Research in visual perception is predominantly an empirical or evidence-based research initiative aimed at the formation or confirmation of hypotheses, theories etc. In recent years, eye-tracking has emerged as an increasingly powerful means for analysing visual and visuo-locomotive human behaviour in general settings, as well as in specialised areas of everyday life and professional activity. Within eye-tracking based visual perception research, statistical data analytics and complex data visualisation have received significant interest in both academia and industry [Blascheck et al., 2014]; this is typically done in synchrony with manual questionnaire based subject-experimenter interactions, think-aloud protocols etc. As for eye-tracking methodology itself, a key emphasis and primary concern from a technological perspective has been on computational and algorithmic foundations aimed at evaluating the distribution and dynamics of eye-movement patterns [Holmqvist et al., 2011]. Our research extends these lines of work, but is a departure from dominant approaches in its focus on high-level semantic interpretation, qualitative analysis, and multi-modality at the interface of AI, HCI, and Visual-Spatial Computing:

» Assistive technologies (applications). From the applied perspective of human-centred cognitive assistive technologies for evidence-based studies in human perception, we present an AI based computational backbone —encompassing computer vision and KR methods— for next-generation software and services in (eye-tracking driven) visual perception research.

» Integrating Vision and KR. From the theoretical perspective of vision and KR research, we focus on developing general methods for the integration of visual processing with (logic-based) declarative reasoning about space and motion in the context of constraint logic programming.

The key emphasis in this paper is on human-centred semantic interpretation and qualitative analysis of multi-modal perceptual data encompassing vision and eye-tracking. Whereas visual perception provides a compelling applied backdrop for the development and demonstration of vision and KR-centric general methods and tools for visuo-spatial computing, the broader orientation of the particular line of research (presented in this paper) is geared toward tighter integration of KR with state of the art in computer vision, contributing to the agenda of what has been attributed as cognitive vision at the interface of language, logic, and artificial intelligence [Cohn et al., 2003; Vernon, 2008; Bhatt et al., 2013b]. This, we posit, impacts several AI application areas (e.g., vision and robotics) beyond the focus of this paper.

Cognitive Film Studies (CFS) Cognitive studies of the moving image —film, digital media etc— has emerged as an area of research at the interface of disciplines as diverse as aesthetics, psychology, neuroscience, film theory, and cognitive science.1 Within CFS, the role of mental activity of observers (e.g., subjects / spectators) has been regarded as one of the most central objects of inquiry [Nannicelli and Taberham, 2014; Aldama, 2015; Sobchack, 2004]. Principal research questions addressed pertain to the systematic study and generation of evidence that can characterise and establish correlates between principles for the synthesis of the moving image, and its cognitive (e.g., embodied visuo-auditory, emotional) recipient effects on observers [Suchan and Bhatt, 2016].

Our technological focus within CFS is on the high-level analysis of subject’s visual fixation or saccadic eye-movement patterns whilst watching a film and correlating this

with semantic analysis of the visuo-auditory data (e.g., fixation on movie characters, influence of cinematographic devices such as cuts and sound effects on attention etc).

**Integrated Vision and KR for Visual Perception** This paper focusses on an integration of computer vision and KR for semantic question answering with video and eye-tracking data in the domain of film. We present a formal model and general methods & tools focussing on (F1–F3):

(F1). **Visual Processing** an integrated pipeline for visual processing of video and eye-tracking data from the viewpoint of high-level feature extraction encompassing spatio-temporal gaze data clustering, people tracking, and (for the film domain) identification of scene structure, camera movements, and character identity.

(F2). **Space - Motion - Histories** a framework for the semantic interpretation of dynamic visuo-spatial imagery encompassing video and eye-tracking data; here, we especially highlight one aspect of the framework concerned with ontologically and computationally elevating perceptual and analytical entities like moving objects, areas of attention and interest, visuo-perceptual saliency, heatmaps as primitive spatio-temporal objects that can be qualitatively and declaratively reasoned about within constraint logic programming.

(F3). **Semantic Question-Answering** running examples of the underlying constraint logic programming implementation with sample queries in the context of a film & eye-tracking dataset. The examples focus on question-answering pertaining to the geometry of a scene [Suchan and Bhatt, 2016] (from a cinematographic viewpoint) in synergy with visual attention predicates related to eye-tracking.

The overall framework (Fig. 2) includes several modules and a pipeline needed for the semantic analysis of visual perception: eye movement and corresponding video datasets are obtained from experiments in visual perception and processed for qualitative spatio-temporal analysis and semantic interpretation. The key modules in the pipeline include the general declarative representations and the inference and query capability based on constraint logic programming. In the backdrop of (F1–F3), we demonstrate the manner in which the integrated visual computing and KR foundations may be applied for the development of human-centred assistive technology supporting high-level interpretation and qualitative analysis. As one instance, we illustrate how results may be used on-demand with question answering, or via a (semantic) database that can be used for applications such as natural language summarisation of experiments.

2 Visual Processing:

**Perception — Scene Structure**

Visuo-spatial semantics for cognitive film studies (from the viewpoint of this paper) include scene objects (people, objects in the scene), cinematographic aids (camera movement, shot types, cuts and scene structure), and perceptual artefacts (eye-tracking / gaze points, areas of attention). In the following, we summarise the visual processing module(s) of Fig. 2 with respect to the cinematographic scene structure of Fig. 1 and Alg. 1.

**Perceptual Artefacts** Visual attention may be estimated based on the dynamics and distribution of eye movement data [Holmqvist et al., 2011]. Gaze data can be grouped for an individual, or may be aggregated from multiple subjects, to Areas of Attention (AOA), via the calculation of eye movement primitives, e.g. scan-path of single spectator including detection of gaze types such as saccadic movement, fixations, smooth pursuit etc; heatmaps based on aggregate gaze; clustering of gaze points. We estimate regions of high attention for a group of people using density based clustering on the gaze points of all participants at a single time point. We also estimate subject attention by calculating a heat map from the gaze points, in a static way, using all gaze points at one time point, and additionally dynamically, using motion compensated gaze points for consecutive time points: (1) estimate the motion in the video data at the position of the gaze point based on Lucas-Kanade optical flow [Lucas and Kanade, 1981]; (2) afterwards the heat map is generated by weighted addition of the gaussian for the motion compensated gaze points for $n$ consecutive time points.
Scene Structure  Computer vision (CV) research has resulted in a variety of methods for detecting humans, body structure, interactions [Hoai and Zisserman, 2014; Bojanowski et al., 2013; Laptev and Perez, 2007], as well methods for estimating facing directions [Marin-Jimenez et al., 2014], or recognising the identity of characters in movies [Tapaswi et al., 2012]. The low-level visual processing algorithms that we utilise for high-level semantic analysis are founded in state-of-the-art outcomes for detection and tracking of people, objects, and motion [Farnebäck, 2003; Dalal and Triggs, 2005; Felzenszwalb et al., 2010; Rodriguez-Molina and Marin-Jimenez, 2011; Jia et al., 2014].

Analysing the structure of the scene involves identifying cuts, i.e., segmenting [Apostolidis and Mezaris, 2014] the scene into its basic elements. This results in single shots, which are used for further cinematographic analysis of the scene. Subsequently, estimation of camera movement (i.e., up, down, left, right, forward, backward) is based on Fernaback’s dense optical flow [Farnebäck, 2003]; estimating the horizontal and vertical camera movement is done by calculating the average movement of all sample points in the x and y direction. For estimating forward and backward movement, we normalise the direction of movement for each sample point with respect to the centre of the frame and calculate the average movement for the normalised samples. We use histograms of oriented gradients (HOG) [Dalal and Triggs, 2005] for face detection and deformable part models (DPM) [Felzenszwalb et al., 2010; Rodriguez-Molina and Marin-Jimenez, 2011] to detect people and upper bodies. For tracking, we use particle filters for each potential track in the scene. We use optical flow [Lucas and Kanade, 1981] and color histograms to track the movement of the detected entities. Thus, we obtain space-time histories for all detected entities in the scene (Fig. 4, and Alg. 1). Finally, for character identification, we use Convolutional Neural Networks (CNN) based deep learning as implemented and made available in the Caffe framework [Jia et al., 2014]; we train the network on pictures of the faces of the characters in the movie, to associate the character names to the extracted people tracks, obtained by the detection and tracking algorithms.

3 Space, Motion, Histories

Commonsense spatial, temporal, and spatio-temporal relations and patterns (e.g., “left”, “overlap”, “during”, “between”, “separation”, “collision”) serve as powerful abstractions for the spatio-linguistic grounding of visual perception and embodied action & interaction [Bhatt et al., 2013a; Suchan et al., 2014]; such spatio-linguistic primitives constitute the basic ontological building blocks of visuo-spatial computing in diverse areas, especially those involving the processing and interpretation of potentially large volumes of highly dynamic spatio-temporal data and commonsense reasoning about space, action, and change [Bhatt, 2012].

Notation: Spatial and temporal objects may be abstracted with primitives such as regions, points, oriented points, line segments. We use a first-order language with sorts for: objects: \( O = \{ o_1, o_2, \ldots , o_i \} \); space-time primitives (regions, points etc): \( E = \{ e_1, e_2, \ldots , e_i \} \); time points: \( T = \{ t_1, t_2, \ldots , t_j \} \); 1D intervals: \( \Delta = \{ \delta_1, \delta_2, \ldots , \delta_l \} \); fluents: \( \Phi = \{ \phi_1, \phi_2, \ldots , \phi_m \} \); actions and events: \( \Theta = \{ \theta_1, \theta_2, \ldots , \theta_n \} \). The spatial configuration of objects in the scene is represented using \( n \)-ary spatial relations \( R = \{ r_1, r_2, \ldots , r_n \} \) of a particular logic of space / time. \( \Phi = \{ \phi_1, \phi_2, \ldots , \phi_n \} \) is a set of propositional and functional fluents, e.g. \( \phi(e_1, e_2) \) denotes the spatial relationship between \( e_1 \) and \( e_2 \). We use functions that map from the object to the corresponding spatial primitive – extend: \( O \times T \mapsto e_\phi \) where \( O \) is the object and \( e_\phi \) is the spatial primitive denoting a spatial property of the object at time \( t \). Predicates \( \text{holds-at}(\phi, r, t) \) and \( \text{holds-in}(\phi, r, \delta) \) are used to denote that the fluent \( \phi \) has the value \( r \) at time \( t \), resp. in time interval \( \delta \). Accordingly, we use \( \text{occurs-at}(\theta, t) \), and \( \text{occurs-in}(\theta, \delta) \) to denote that an event or action \( \theta \) occurred at a time point \( t \) or in an interval \( \delta \).
Space and Time Spatial and temporal relations are used to represent the perceived dynamics in a scene. The spatio-temporal domain is modelled using the mereotopological relations of the RCC8 fragment of the RCC calculus [Randell et al., 1992], which consists of the eight base relations \( R_{\text{top}} \equiv \{ \text{dc, ec, po, eq, tp, ntp, tpp, tpp^{-1}}, \text{ntpp} \} \), the positional relations using the rectangle algebra which uses the relations of Allen’s interval algebra [Allen, 1983]. \( R_{\text{int}, \text{eral}} \equiv \{ \text{before, after, during, contains, starts, started_by, finishes, finished_by, overlaps, overlapped_by, meets, met_by, equal} \} \), for representing position for each dimension (horizontal and vertical) separately. We use ordering relations \( \{ <, =, > \} \) to compare properties of spatial objects, i.e. size and distance. Further, Allen’s intervals algebra is used for representing temporal relations between events and actions, where we consider time points to be intervals where the start point is equal to the end point.

Space-Time Histories These are regions in space-time [Muller, 1998] (depicted in Fig. 3). The space-time history \( sth \) of an object \( o \) is given by the function \( sth: O \rightarrow E \times T \), which maps the object to its appearance in space and time. \( sth(o, \delta) = (\varepsilon_1, \varepsilon_2, \varepsilon_3, \ldots, \varepsilon_n) \), where \( \varepsilon_1 \) to \( \varepsilon_n \) denotes the spatial primitive representing the object \( o \) at the time points \( t_1 \) to \( t_n \). Space-time histories serve as basic primitives to represent and reason about the spatio-temporal dynamics in a perceived scene, by defining movement patterns (dynamic spatio-temporal relations), and actions and events, based on the perceived object movement. We define movement relations based on changes in object positions.

\[
\text{holds-in}(\text{moving}(o, \delta), \text{true}, \delta) \supset \text{during}(t_i, \delta) \land \text{during}(t_j, \delta) \\
\text{before}(t_i, t_j) \land (\text{position}(o, t_i) \neq \text{position}(o, t_j)).
\]

\[
\text{holds-in}(\text{stationary}(o, \delta), \text{true}, \delta) \supset \text{during}(t_i, \delta) \land \text{during}(t_j, \delta) \\
\text{before}(t_i, t_j) \land (\text{position}(o, t_i) = \text{position}(o, t_j)).
\]

Accordingly, growth and shrinkage of an object is defined based on the changes in size of an object, in one or more dimensions.

\[
\text{holds-in}(\text{growing}(o, \delta), \text{true}, \delta) \supset \text{during}(t_i, \delta) \land \text{during}(t_j, \delta) \\
\text{before}(t_i, t_j) \land (\text{size}(o, t_i) < \text{size}(o, t_j)).
\]

\[
\text{holds-in}(\text{shrinking}(o, \delta), \text{true}, \delta) \supset \text{during}(t_i, \delta) \land \text{during}(t_j, \delta) \\
\text{before}(t_i, t_j) \land (\text{size}(o, t_i) > \text{size}(o, t_j)).
\]

![Figure 3: Commonsense Spatial Reasoning with Spatio-Temporal Entities. Illustrated are: Space-Time Histories, and Spatio-Temporal Pattern and Events, i.e. discrete, overlapping, inside, parallel movement, merge, and split](image)

Algorithm 1: SceneSemantics\((O, PA, \Delta_\delta)\)

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( STH_{PA, o} \leftarrow \varnothing )</td>
</tr>
<tr>
<td>2</td>
<td>for ( p_o \in PA ) do</td>
</tr>
<tr>
<td>3</td>
<td>( sth_{p_o} \leftarrow \varnothing )</td>
</tr>
<tr>
<td>4</td>
<td>for ( t \in T ) do</td>
</tr>
<tr>
<td>5</td>
<td>( sth_{p_o} \leftarrow sth_{p_o} \cup \text{p}t_{p_o} )</td>
</tr>
<tr>
<td>6</td>
<td>( STH_{PA} \leftarrow STH_{PA} \cup sth_{p_o} )</td>
</tr>
<tr>
<td>7</td>
<td>for ( \delta \in \Delta_\delta ) do</td>
</tr>
<tr>
<td>8</td>
<td>for ( o \in O ) do</td>
</tr>
<tr>
<td>9</td>
<td>( sth_{o, \delta} \leftarrow \varnothing )</td>
</tr>
<tr>
<td>10</td>
<td>for ( t \in \delta ) do</td>
</tr>
<tr>
<td>11</td>
<td>( sth_{o, \delta} \leftarrow sth_{o, \delta} \cup \text{extend}(o, t) )</td>
</tr>
<tr>
<td>12</td>
<td>( STH_o \leftarrow STH_o \cup sth_{o, \delta} )</td>
</tr>
<tr>
<td>13</td>
<td>( STH \leftarrow STH_o \cup STH_{PA} )</td>
</tr>
<tr>
<td>14</td>
<td>return ( STH )</td>
</tr>
</tbody>
</table>

Movement Pattern (MP) describe spatio-temporal dynamic, by combining arbitrary spatial and temporal relation. The space of possible movement patterns is huge and there are many patterns that are useful to describe visuo-spatial phenomena. E.g. the following pattern describes that one object moves inside another object.

\[
\text{holds-in}(\text{inside}(o_1, o_2), \text{true}, \delta) \supset \text{holds-in}(\text{moving}(o_1, \delta), \text{true}, \delta) \land \text{holds-in}(\text{moving}(o_2, \delta), \text{true}, \delta) \land \\
\text{holds-in}(\text{p}o_{o_2}(o_1, o_2), (\text{pp}, \text{ntpp, eq}), \delta).
\]

Relative Movement of objects, such as approaching and receding, is defined based on changes in distance between objects. E.g. approaching is defined as follows:

\[
\text{holds-in}(\text{approaching}(o_1, o_2), \text{true}, \delta) \supset \text{during}(t_i, \delta) \land \text{during}(t_j, \delta) \\
\text{before}(t_i, t_j) \land (\text{distance}(o_1, o_2, t_i) > \text{distance}(o_1, o_2, t_j)).
\]

Complex movement patterns are defined by combining different spatio-temporal aspect, e.g. a pattern describing that two objects are moving parallel to each other could then be defined as follows:
4 Semantic Question-Answering:
Moving Image and its Reception

From the viewpoint of semantic question-answering for the analysis of the visual reception of the moving image, consider the instances in (Q1–Q3) reflecting the kinds of Q/A capabilities necessary from the viewpoint of cognitive film studies:

(Q1) how is the spectator attention shifting, when the camera is moving / after a cut / during a long shot?
(Q2) which movement / characters / objects is the spectators attention following in a spatio-temporal sense?
(Q3) are there individual or aggregate regularities with respect to the shift in spectator attention at a certain time?

As one use-case, consider again the scene depicted in Fig. 4; using our framework, it is possible to define (manually, or using other UI means) high-level rules and execute queries in the logic programming language PROLOG to reason about spectator attention; details follow:

Attention Predicates and Queries (sample). The set of rules characterising different kinds of attention and fixation behaviours via-a-vis video analysis is in principle extensive, and open-ended. Some examples include:

- $\text{attn_on}(\text{Obj}, \text{Int})$ – attention $\text{Att}$ is overlapping or covering object $\text{Obj}$ during time interval $\text{Int}$
- $\text{attn_following}(\text{Att}, \text{Obj}, \text{Int})$ – attention $\text{Att}$ is following the movement of object $\text{Obj}$ during time interval $\text{Int}$
- $\text{attn_shift}(\text{Att}, T)$ – attention $\text{Att}$ shifts at time point $T$
- $\text{attn_focusing}(\text{Att}, \text{Int})$ – attention $\text{Att}$ becomes more focused during the time interval $\text{Int}$

We illustrate some select sample encodings given the backdrop of Q/A needs such as in (Q1–Q3). The following attention predicate is true if the space-time history of an object is topologically connected, i.e. inside or overlapping, with the space-time history of attention.

$$\text{attn_on}(\text{Obj}, \text{Int}) :- \text{attn}(\text{Obj}, \text{ST.Obj}), \text{attn_aggregate}(\text{spectator.set.top_list}), \text{ST.AOA}, \text{holds_in(inside(\text{ST.Obj}, \text{ST.AOA}), \text{Int})};$$
$$\text{attn_on(overlapping(\text{ST.AOA}, \text{ST.Obj}), \text{Int})}. $$
Figure 5: Interaction Taxonomy

Given the above rule, a query where the spatio-temporal history of the character Jack is compared with the aggregated Area of Attention of all participants would be the following:

?- Int = interval(_, _), attn_on(jack, Int).

The query results in all time intervals during which spectator attention is on the character Jack:

Int = interval(5, 30);
...

One could also analyse the dynamics of spectator attention based on movement patterns and events. For instance, consider the st-histories of Fig. 4b: here, a rule determining how the attention follows the objects in the scene is:

attn_following(Att, Obj, Int) :- sth(Obj, ST_Obj),
sth(attn_following(Att, ST_Obj), Act),
occures_in(following(Att, ST_Obj), Int).

This can be used to query objects the attention is following:

?- Int = interval(_, _), attn_on_following(_, Obj, Int).

This results in the objects the attention is following, i.e., the main characters of the scene:

Obj = jack,
Int = interval(5, 30);
Obj = francis,
Int = interval(13, 30);
...

Further, one could formulate a query to determine what happened when the areas of attention following Jack and Francis merged?

?- Int = interval(_, _), TP = timepoint(_),
sth(jack, st_jack), sth(francis, st_francis),
sth(attn_following(ST_AOA_1, st_jack), _),
sth(attn_following(ST_AOA_2, st_francis), _),
occures_in(stmerge(ST_AOA_1, ST_AOA_2), _, TP),
occures_in(Rs, Int), time(TP, Int, during).

The result of the query is that Francis is approaching Jack when the respective areas of attention merge:

Hence, semantic Q&A becomes possible with spatio-temporal entities of visual attention as well as domain-specific perceptual elements; both categories exist as native entities within the (Prolog based) constraint logic programming framework.

Analytical Summarisation The declarative representations and the inference and query capability provided by the framework (Fig. 2) can be used as a basis for (language-based) analytical summarisation. Listing L1 is a select part of a summary corresponding to the scene in (Fig. 4); the summary has been generated using a (spatio-temporal feature based) natural language generator. Note that the semantics for spatial, temporal, and behavioural information is grounded to relations in the underlying theory of space and motion. This manner of natural language based analytical summarisation of experiments—to the best of our knowledge—presents a novel user interaction paradigm and functional benchmark in visual perception research.

5 Summary

We presented a visuo-spatial computing framework consisting of integrated formal KR and low-level visual processing foundations, including the algorithms & data-structures, and resulting general methods & tools that serve as the computational backbone for next-generation software and services aimed at semantic interpretation and qualitative analytics (for visual perception studies). As examples, we focused on the capability to perform semantic Q/A about the dynamics of space-time histories and their mutual interactions within (constraint) logic programming.

This work is driven by a tighter integration of KR and computer vision; cognitive vision as an area of research has gained prominence, with recent initiatives addressing the topic from the perspectives of language, logic, and AI. There has also been recent interest from the computer vision community to synergise with cognitively motivated methods for perceptual grounding and inference with visual imagery. We posit that KR+Vision can serve a crucial role for the development of hybrid AI & cognitive interaction technologies where processing and human-centred semantic interpretation of dynamic visuo-spatial imagery are central.

3 NLG [Reiter and Dale, 2000] is beyond the scope of this paper; we have used the specialised (Prolog based) NL generator provided by [Suchan et al., 2015].
References


3.1. Visuospatial Semantics

3 - The Geometry of a Scene

Published in:
IEEE Winter Conference on Applications of Computer Vision (WACV 2016)

Citation:

This conference paper was published at the Winter Conference on Applications of Computer Vision (WACV 2016) and is aiming at computer vision research. It presents declaratively grounded models for semantic question-answering about the perception of video. A particular focus is on the application of computer vision foundations for semantic reasoning about the “Geometry of a Scene”, i.e., scene elements and semantic structure of the scene, and its reception based on eye-tracking data. It presents knowledge-based qualitative reasoning (e.g., about object / character placement, scene structure) based on the dynamic scene structure obtained from state of the art computer vision methods for detecting, tracking, and recognition of people, objects, and cinematographic devices such as cuts, shot types, types of camera movement.

The paper demonstrates declarative reasoning and semantic question-answering with video and eye-tracking data based on deep semantic representations of space and motion. In this context, it showcases semantic interpretation and inference with spatial primitives, and highlights the role of powerful (deep learning based) visual processing for visuospatial sensemaking.
THE GEOMETRY OF A SCENE

On Deep Semantics for Visual Perception Driven Cognitive Film Studies

Jakob Suchan Mehul Bhatt

Human-Centred Cognitive Assistance Lab. | hcc.uni-bremen.de/
University of Bremen, Germany
{jsuchan, bhatt}@uni-bremen.de

Abstract

We present a general computational narrative model encompassing primitives of space, time, and motion from the viewpoint of deep knowledge representation and reasoning about visuo-spatial dynamics, and (eye-tracking based) visual perception of the moving image. The declarative model, implemented within constraint logic programming, integrates knowledge-based qualitative reasoning (e.g., about object / character placement, scene structure) with state of the art computer vision methods for detecting, tracking, and recognition of people, objects, and cinematographic devices such as cuts, shot types, types of camera movement. A key feature is that primitives of the theory—things, time, space and motion predicates, actions and events, perceptual objects (e.g., eye-tracking / gaze points, regions of attention etc)— are available as first-class objects with deep semantics suited for inference and query from the viewpoint of analytical Q&A or studies in visual perception.

We present the formal framework and its implementation in the context of a large-scale experiment concerned with analysis of visual perception and reception of the moving image in the context of cognitive film studies.

1. INTRODUCTION

Cognitive studies of the moving image—film, digital media etc—has emerged as an area of research at the interface of disciplines as diverse as aesthetics, psychology, neuroscience, film theory, and cognitive science.\(^1\)\(^2\) Within cognitive film theory, the role of mental activity of observers (e.g., subjects / spectators, analysts / critics) has been regarded as one of the most central objects of inquiry [1, 28].

Principal research questions that emerge in the context of cognitive film theory pertain to the systematic study and generation of evidence that can characterise and establish strong correlates between principles for the synthesis of the moving image (Listing L1), and its cognitive (e.g., embodied visual, auditory, aesthetic, emotional) recipient effects and influences on observers.

Visual Semantics of the Moving Image Driven by cognitive studies of cinema, and cognitive film theory in particular, we interpret the moving image in a broad sense to encompass: multi-modal visuo-auditory perceptual signals (also including depth sensing, haptics, and empirical observational data) where basic concepts of semantic or content level coherence, and spatio-temporal continuity and narrativity are applicable. With this as a basis, this paper focusses on methods for investigating the visuo-spatial semantics of the moving image at the interface of artificial intelligence based spatial representation and reasoning, visuo-spatial cognition, and computational models of narrative [8]. In particular, we develop and demonstrate foundational methods focussing on cognitively-driven qualitative analysis of dynamic visuo-spatial imagery encompassing:

- **geometry of a scene.** content-level deep semantic analysis of scene structure and semantics—object / character identity and placement, cuts, shot types, categories of camera movement—pertaining to the moving image.
• **perception & reception.** Visual perception analysis of spectator behaviour and engagement with the medium, e.g., visual fixation on film characters, gaze patterns co-related with influence of cinematographic aids such as cuts, long takes, symmetry on attention and whilst watching a film.

Our research addresses space and spatio-temporal dynamics from the viewpoint formal representation and computational reasoning about space, events, actions, and change, especially focussing on space and motion as interpreted within artificial intelligence and knowledge representation and reasoning (KR) in general, and declarative spatial reasoning [5, 36] in particular.

**Declarative Narrativisation and Deep Semantics** With respect to a broad-based understanding of the moving image (as aforediscussed), we define dynamic visuo-spatial perceptual narratives as declarative models of visual, auditory, haptic and other (e.g., qualitative, analytical) observations in the real world that are obtained via artificial sensors and/or human input. Deep semantics denotes:

the existence of declaratively grounded models (e.g., for spatial and temporal knowledge) and systematic formalisation that can be used to perform reasoning and query answering, relational learning, or more broadly, even embodied simulation.  

Deep semantics, founded on declarative representation and inference, serves as basis to externalise explicit inferred knowledge, e.g., using modalities such as diagrammatic representations (e.g., Fig. 3), natural language (e.g., Listing L1), complex (dynamic) data visualisation (e.g., Fig. 2) etc.

**Evaluation: An Experimental Case-Study** We demonstrate the model by its application to the domain of cognitive film studies for analysing visual experience combining deep visual analysis of the “geometry of a scene” with analysis of eye movement behaviour. Examples and empirical evaluation are presented in the context of large-scale experiment with a total of 31 subjects, and involving 16 scenes (per subject) from 12 films, with each scene ranging between 0:38 minute to max. of 9:44 minutes in duration (Table 1).

<table>
<thead>
<tr>
<th>Film / Director</th>
<th>Scenes</th>
<th>Duration (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jaws (1975)</td>
<td>1</td>
<td>4:36</td>
</tr>
<tr>
<td>Steven Spielberg</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Untouchables (1987)</td>
<td>1</td>
<td>9:44</td>
</tr>
<tr>
<td>Brian De Palma</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paprika (2006)</td>
<td>1</td>
<td>1:49</td>
</tr>
<tr>
<td>Satoshi Kon</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wes Anderson</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moonrise Kingdom (2012)</td>
<td>1</td>
<td>1:56</td>
</tr>
<tr>
<td>Wes Anderson</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Darjeeling Limited (2007)</td>
<td>1</td>
<td>1:25</td>
</tr>
<tr>
<td>Wes Anderson</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Hunger Games (2012)</td>
<td>1</td>
<td>2:48</td>
</tr>
<tr>
<td>Gary Ross</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solaris (1972)</td>
<td>1</td>
<td>7:46</td>
</tr>
<tr>
<td>Andrey Tarkovsky</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Shining (1980)</td>
<td>2</td>
<td>[2:26, 0:38]</td>
</tr>
<tr>
<td>Stanley Kubrick</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drive (2011)</td>
<td>3</td>
<td>[2:59, 0:51, 1:59]</td>
</tr>
<tr>
<td>Nicolas Winding Refn</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Bad Sleep Well (1960)</td>
<td>1</td>
<td>2:46</td>
</tr>
<tr>
<td>Akira Kurosawa</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goodfellas (1990)</td>
<td>1</td>
<td>3:03</td>
</tr>
<tr>
<td>Martin Scorsese</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 1:** Experiments in Deep Semantics and Eye-Tracking Based Visual Perception. (case-study developed in this paper is part of this experiment with 31 subjects)

**CORE CONTRIBUTIONS**

We present a computational narrative model for performing Q/A centered deep semantic analysis of the geometry—structure and semantics—of the moving image and its visual perception and reception by the audience:

1. **Space & Motion** a domain-independent formal framework encompassing primitives of space, time, and motion for commonsense representing and reasoning about dynamic visuo-spatial imagery. The framework is founded in logic programming such that a corresponding implementation is seamlessly usable as a generic library of space & motion via declarative programming frameworks based on logic programming (e.g., Prolog based CLP(QS) [5]) and answer-set programming (e.g., ASPMT(QS) [36]).

2. **Commonsense Cognitive Vision** integration of the formal KR-based commonsense theory of space, time and motion with state of the art computer vision methods that have been customised herein for the film domain. This encompasses detection, tracking, and recognition of people, objects, cinematographic devices such as (camera) motion, cuts, shot types, object / character placement & scene structure. Whereas our application of state of the art computer vision is film domain specific, the integration with KR methods serves as a model for other areas in AI, e.g., vision.

---

3Whereas this paper alludes to logic programming, the broader agenda of “deep semantics” indeed relates to “deep KR” also encompassing other declarative KR frameworks such as description logic based (spatioterminalogical) reasoning, answer-set programming based non-monotonic (spatial) reasoning, or even other specialised commonsense reasoners based on expressive action description languages for handling space, events, action, and change [4].

4We conducted the experiments with the stationary Tobi X2-60 Eye Tracker, collecting eye movement data with a rate of 60 Hz.
& robotics, where commonsense reasoning about space and motion is crucial.

(3). Implementation  The proposed framework has been fully modelled and implemented declaratively within constraint logic programming (CLP). We emphasize that the level of declarativeness within logic programming is such that each aspect pertaining to the overall framework can be seamlessly customised and elaborated, and that question-answering & query can be performed using the primitives of the theory —things, space and motion, actions and events, perceptual objects (e.g., eye-tracking / gaze points, regions of attention etc)— as first class objects within the CLP environment.

2. SPACE, MOTION, HISTORIES

Commonsense spatial, temporal, and spatio-temporal relations and patterns (e.g., “left”, “overlap”, “during”, “between”, “separation”, “collision”) serve as powerful abstractions for the spatio-linguistic grounding of visual perception and embodied action & interaction [6]. Such spatio-linguistic primitives constitute the basic ontological building blocks of visuo-spatial computing in diverse areas, especially those involving the processing and interpretation of potentially large volumes of highly dynamic spatio-temporal data: architecture design [7], geographic information systems [10], cognitive vision and robotics [9, 31, 33].

The high-level semantic interpretation and qualitative analysis of visual attention in the context of visual perception studies requires the representational and inferential mediation of (declarative) qualitative abstractions of the visuo-spatial dynamics, encompassing space, time, motion, and interaction. We use a first-order typed language (L) with the following alphabet: \{\neg, \land, \lor, \exists, \forall, \equiv\} (respectively meaning negation, conjunction, disjunction, universal quantification, existential quantification, implication, and equivalence).

Notation: Spatial and temporal objects may be abstracted with primitives such as regions, points, oriented points, line segments as per needs. We use a first-order language with sorts for: objects: \mathcal{O} = \{o_1, o_2, \ldots, o_n\}; space-time primitives (regions, points etc): \mathcal{E} = \{e_1, e_2, \ldots, e_i\}; time points: \mathcal{T} = \{t_1, t_2, \ldots, t_i\}; 1D intervals: \Delta = \{\delta_1, \delta_2, \ldots, \delta_i\}; fluents: \Phi = \{\phi_1, \phi_2, \ldots, \phi_i\}; actions and events: \Theta = \{\theta_1, \theta_2, \ldots, \theta_i\}. The spatial configuration of objects in the scene is represented using n-ary spatial relations \mathcal{R} = \{r_1, r_2, \ldots, r_n\} of a particular logic of space / time. \Phi = \{\phi_1, \phi_2, \ldots, \phi_i\} is a set of propositional and functional fluents, e.g. \phi(e_1, e_2) denotes the spatial relationship between e_1 and e_2. We use functions that map from the object to the corresponding spatial primitive — extend: \mathcal{O} \times \mathcal{T} \mapsto \varepsilon_\phi where \mathcal{O} is the object and \varepsilon_\phi is the spatial primitive denoting a spatial property of the object at time t. Predicates holds-at(\phi, r, t) and holds-in(\phi, r, t) are used to denote that the fluent \phi has the value r at time t. We use occurs-at(\theta, t), and occurs-in(\theta, \delta) to denote that an event or action \theta occurred at a time point t or in an interval \delta.

SPACE AND TIME

Spatial and temporal relations (Fig. 1) are used to represent the perceived dynamics in a scene. The spatio-temporal domain is modelled using the topological relations of the RCC8 fragment of the RCC calculus [29] (Fig. 1a), which consists of the eight base relations \mathcal{R}_{\text{base}} \equiv \{dc, de, ec, po, eq, tpp, ntpp, tpp^{-1}, ntppp^{-1}\}, the positional relations using the rectangle algebra which uses the relations of Allen’s interval algebra [2] \mathcal{R}_{\text{interval}} \equiv \{\text{before, after, during, contains, starts, started by, finishes, finished by, overlaps, overlapped by, meets, met by, equal}\}, for representing position for each dimension (horizontal and vertical) separately (Fig. 1b). We use ordering relations \mathcal{R}_{\text{ord}} \equiv \{\prec, \sim, >\} to compare properties of spatial objects, i.e. size and distance. Further, we also use Allen’s intervals for representing temporal relations between events and actions, where we consider time points to be intervals where the start point is equal to the end point, i.e. I = interval(t, t).

\[\text{These characterisations are sufficient for the examples of this paper; an elaborate set of spatio-temporal relations to represent and reason about} \]
SPACE-TIME HISTORIES

These are defined as regions in space-time. The space-time history of an object is given by the function sth : O → S × T, which maps the object to its appearance in space and time. For representing connectedness of space-time histories, we appeal to spatial and temporal connectedness (s-connected and t-connected) [19, 27]. If two space-time histories are connected in space and time we say they are st-connected. Space-time histories serve as basic primitives to represent and reason about the spatio-temporal dynamics in a perceived scene, by defining movement patterns (dynamic spatio-temporal relations), and actions and events.

Movement Pattern (MP) describe spatio-temporal dynamics, by combining relations, MP = r_1 × r_2 × ... × r_i where r_i ∈ R for arbitrary spatial and temporal relation. The space of possible movement patterns is huge and there are many patterns that are useful to describe visuo-spatial phenomena. E.g. the following pattern describes that one object moves inside another object.

\[
\begin{align*}
\text{holds-in}(\text{inside}(o_1, o_2), t, t_1) & \; \rightarrow \\
\text{holds-in}(\; & \text{holds}(\text{move}(o_1, o_2), \{\text{top, up, right, down, left,} \}; t_1, t_2) ; t_1 < t_2) \; (1)
\end{align*}
\]

Relative Movement of objects, such as approaching and receding, is defined based on changes in distance between objects. E.g. approaching is defined as follows:

\[
\begin{align*}
\text{holds-in}(\text{approaching}(o_1, o_2), t, t_1) & \; \rightarrow \\
\text{holds}(\; & \text{holds}(\text{move}(o_1, o_2), \{\text{top, up, right, down, left,} \}; t_1, t_2) ; t_1 < t_2) \; (2)
\end{align*}
\]

Accordingly growth and shrinkage of an object is defined based on the changes in size of an object, in one or more dimensions. Complex movement patterns are defined by combining different spatio-temporal aspects, e.g. a pattern describing that two objects are moving parallel to each other could then be defined as follows.

\[
\begin{align*}
\text{holds-in}(\text{parallel}(o_1, o_2), t, t_1) & \; \rightarrow \\
\text{holds}(\; & \text{holds}(\text{move}(o_1, o_2), \{\text{top, up, right, down, left,} \}; t_1, t_2) ; t_1 < t_2) \; (3)
\end{align*}
\]

Actions and Events describe processes that change the spatio-temporal configuration of objects in the scene, at a time point or in a time interval ; these are defined by the involved spatio-temporal dynamics in terms of changes in the status of st-histories caused by the action or event, i.e. the description consists of spatio-temporal relations and movement patterns of the involved st-histories, before, during and after the action or event.

- Appearance and Disappearance describes the cases where the existence status of an object changes, i.e. the time point, space, time, and motion in the context of dynamic visuo-spatial imagery can be utilised as per [33].

where the st-history starts to exists, ends to exist.

\[
\begin{align*}
\text{occurs-at}(\text{appearing}(o_1), t, t_1) & \; \rightarrow \\
\text{holds}(\; & \text{exists}(o_1), \text{false}; t_1) \; \land \\
\text{holds}(\; & \text{exists}(o_1), \text{true}; t_1) \; \land \\
\text{holds}(\; & \text{meets}(t_1, t, \text{true}) \; (4)
\end{align*}
\]

- Movement Events change describes in the spatial state of the space-time histories, due to movement of individuals in the scene, e.g. crossing describes the events that two objects, i.e. st-histories of detected persons cross each other. This happens, for example, when the movement of two persons crosses each other.

\[
\begin{align*}
\text{occurs-at}(\text{crossing}(o_1, o_2), t, t_1) & \; \rightarrow \\
\text{holds}(\; & \text{exists}(o_1, o_2), \text{true}; t_1) \; \land \\
\text{holds}(\; & \text{meets}(t_1, t, \text{true}) \; (5)
\end{align*}
\]

Complex interactions, e.g. a person passing in front, or behind another person, or a person passing between two persons, can be described by combining multiple actions and events. We define a range of actions and events, for describing the dynamics of human interactions, visual attention, and cinematography. E.g. consider the cinematic device of a Tracking Shot describes the action, that the camera is tracking the movement of some objects in the scene.

\[
\begin{align*}
\text{occurs-in}(\text{tracking}(\text{cam}_i, \{o_1, o_2, ..., o_n\}), t, t_1) & \; \rightarrow \\
\text{exists}(\; & \text{meets}(\text{dwell}(\text{cam}_i, \text{true}; t_1, t_2), \text{true}; t_1) \; (7)
\end{align*}
\]

3. VISUAL PROCESSING: PERCEPTION AND SCENE STRUCTURE

Visuo-spatial semantics for cognitive film studies (from the viewpoint of this paper) include scene objects (people, objects in the scene), cinematographic aids (camera movement, shot types, cuts and scene structure), and perceptual objects (eye-tracking / gaze points, areas of attention). The obtained individuals are represented as space-time objects in the context of the presented visuo-spatial narrative model (see Alg. 1).

SCENE STRUCTURE

Detecting visual elements in movies is a key focus in computer vision research and resulted in a variety of methods for detecting humans (including body structure), and their interactions[11, 20, 24], as well methods for estimating facing directions [26] or recognising the identity of characters in movies [34]. The low-level visual processing algorithms that we utilise for high-level semantic analysis are founded in state-of-the-art outcomes from the computer vision community for detection and tracking of people, objects, and motion in the context of film analysis.
Identifying Cuts. Analysing the structure of the scene, includes, identifying cuts [3], i.e. segmenting the scene into its basic elements. In this way, we obtain single shots, that are used for further analysis of the semantics of the scene.

3D Camera Movement. Estimation of camera movement is done using Fernabach’s dense optical flow [17] for two consecutive frames where we reject samples in homogeneous regions, based on the eigen values of the sample points. Estimating the horizontal and vertical camera movement is done by calculating the average movement of all sample points in the x and the y direction. For estimating forward and backward movement, we normalise the direction of movement for each sample point with respect to the centre of the frame and calculate the average movement for the normalised samples.

People Detection and Tracking. We are using histograms of oriented gradients (HOG) [13] for face detection and deformable part models (DPM) [18, 30] to detect people, and upper bodys. To associate detections over time and to generate tracks of people movement, we use particle filters for each potential track in the scene. We use optical flow [25] and color histograms to track the movement of the detected entities. In this way we obtain space-time histories for all detected entities in the scene (see Figure 2).

Character Identification by Deep Learning. We use Convolutional Neural Networks (CNN) based deep learning as implemented and made available in the Caffe framework [22]; we train the network on pictures of the faces of the characters in the movie, to associate the character names to the extracted people tracks, obtained by the detection and tracking algorithms.

Algorithm 1: \(\text{STH(GP, V)}\)

<table>
<thead>
<tr>
<th>Data</th>
<th>(\text{Gaze data (GP) given as gaze point at time point } t, ) and (\text{Video sequence (V) given as frame at time point } t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Result</td>
<td>(\text{A set of Space-Time Histories (STH) where each } sth \in \text{STH is a set of detected regions / points for consecutive time points } t)</td>
</tr>
</tbody>
</table>

1. \(\text{STH} \leftarrow \emptyset\)
2. \(\text{for } t \in \text{GP} \text{ do} \)
   3. \(\text{Att.regions} \leftarrow \text{detect.attention(GP, } t)\)
   4. \(\text{STH} \leftarrow \text{associate(Att.regions)}\)
   5. \(\text{Shots} \leftarrow \text{detect.shots(V)}\)
   6. \(\text{for shot } \in \text{Shots} \text{ do} \)
   7. \(\text{for } t \in \text{Shots} \text{ do} \)
   8. \(\text{Faces} \leftarrow \text{DPM.detect(V)}\)
   9. \(\text{Upper.Bodys} \leftarrow \text{DPM.detect(V)}\)
   10. \(\text{People} \leftarrow \text{DPM.detect(V)}\)
   11. \(\text{Individuals} \leftarrow \text{Individuals} \cup \{(\text{Faces, } t), (\text{Upper.Bodys, } t), (\text{People, } t)\}\)
   12. \(\text{for individual } \in \{\text{Faces, Upper.Bodys, People}\} \text{ do} \)
   13. \(\text{Tracks} \leftarrow p.f_.tracking(individual)\)
   14. \(\text{for track } \in \text{Tracks} \text{ do} \)
   15. \(\text{ID} \leftarrow \text{cnn.identify(track)}\)
   16. \(\text{STH} \leftarrow \text{STH} \cup \{\text{track, ID}_{\text{track}}\}\)
17. \(\text{return } \text{STH}\)

PERCEPTUAL ARTEFACTS

Visual attention may be estimated based on the dynamics and distribution of eye movement data [15, 21]. Gaze data can be grouped for an individual, or may be aggregated from multiple subjects, to Areas of Attention (AOA), via the calculation of eye movement primitives, e.g. scan-path of single spectator including detection of gaze types such as saccadic movement, fixations, smooth pursuit etc; heat.
maps based on aggregate gaze; clustering of gaze points.

Spatial Clustering of Gaze Points. We estimate regions of high attention for a group of people using density based clustering (DBSCAN) [16] on the gaze points of all participants at a single time point.

Regular and Dynamic Heat Maps. We also estimate subject attention by calculating a heat map from the gaze points, in a static way, using all gaze points at one time point, and additionally dynamically, using motion compensated gaze points for consecutive time points: (1) estimate the motion in the video data at the position of the gaze point based on Lucas-Kanade optical flow [25]; (2) afterwards the heat map is generated by weighted addition of the gaussian for the motion compensated gaze points for n consecutive time points.

4. DEEP (VISUO-SPATIAL) SEMANTICS WITH THE MOVING IMAGE

Consider the instances in (Q1–Q3) reflecting the kinds of Q/A capabilities necessary from the viewpoint of cognitive film studies:

(Q1). how is the spectator attention shifting, when the camera is moving / after a cut / during a long shot?

(Q2). which movement / characters / objects is the spectators attention following in a spatio-temporal sense?

(Q3). are there individual or aggregate regularities with respect to the shift in spectator attention at a certain time?

Our space-time history model and its integration with low-level visual processing supports such Q/A on the visual analysis of the scene and the eye movement data. Looking at the space-time history of the aggregated area of attention of all participants, the system is able to answer queries concerning the focus of the attention of the spectators and also involving people and objects in the scene, e.g., at which time(s) was the attention fixated on a certain character.

Drive (2011). Dir: Nicolas Winding Refn As a use-case, consider the scene of the movie Drive (2011) (see Fig. 3), using our framework, it is possible to define (manually, or using other UI means) high-level rules and execute queries in the logic programming language PROLOG to reason about spectator attention;

The domain-specific input data for this scene is as follows:

Given this data we calculate different kinds of geometric representations, e.g. points, regions, line-segments, etc., which serve as a basis for analysing the spatio-temporal dynamics of the scene.

Sample Predicates and Queries. The set of rules characterising different kinds of attention and fixation behaviours vis-a-vis deep video analysis is in principle extensive, and open-ended. Here, we illustrate some select sample encodings (Table 2) given the backdrop of Q/A needs such as in (Q1–Q3). The following attention predicate is true if the space-time history of an object is topologically connected, i.e. inside or overlapping, with the space-time history of attention.

Given the above rule, a query where the spatio-temporal history of a character, e.g. Irene, is compared with the aggregated Area of Attention of all participants would be the following:

attn_on(Obj, Int) :- sth(Obj, ST_Obj), 
sth[aggregate_aoa(spectator_set(gp_list)), ST_AOA],
holds_in(inside(ST_OBJ, ST_AOA), Int); 
holds_in(overlapping(ST_AOA, ST_OBJ), Int).

---

*Within a usable product, it is expected to have UI modalities that will facilitate the creation of user - domain specific rules (i.e., rules need not be predefined, and may be created easily).  
*Within PROLOG [12], ‘;’ corresponds to conjunction, ‘,’ to a disjunction, and ‘a :- b, c.’ denotes a rule where ‘a’ is true if both ‘b’ and ‘c’ are true; capitals are used to denote variables, whereas lower-case refers to constants; ‘.’ (i.e., the underscore) is a “don’t care” variable, i.e., denoting placeholders for variables in cases where one doesn’t care for a resulting value.
The query results in all time intervals in which the spectator's attention is on the character Irene:

\[
\text{Int} = \text{interval}(643, 741);
\]

To semantically analyse the cinematographic characteristics of a scene as in Listing 1.1 using film analysis techniques, i.e. the quadrant system, and to validate the claim about shifting attention while the camera is tracking Irene and the Driver, one could formulate as follows:

\[
\begin{align*}
7- & \ \
& \text{Spectator = subject(s1),} \\
& \text{I = interval(639, 646),} \\
& \text{Half = left;} \\
& \text{Spectator = subject(s1),} \\
& \text{I = interval(647, 746),} \\
& \text{Half = right;} \\
& \text{Spectator = subject(s1),} \\
& \text{I = interval(747, 833),} \\
& \text{Half = left;} \\
& \text{Spectator = subject(s1),} \\
& \text{I = interval(834, 894),} \\
& \text{Half = right;} \\
& \text{false.}
\end{align*}
\]

The answer to this is a sequence of attention shifts between the two right and the left half of the quadrant system.

A Note on Software Integration. Whereas the sample queries in this section have been exemplified using the interactive capabilities of PROLOG, note that it is not necessary to manually use the framework as such; the complete logical reasoning engine of PROLOG (i.e., also our space-time history extensions implemented in PROLOG) can be embedded as a reasoning component within larger software frameworks / middleware etc for online processing, or reasoning results may be serialised within a database for offline / processing of sets of experiments.

5. SUMMARY AND OUTLOOK

Cognitive vision as an area of research has already gained prominence, with several recent initiatives addressing the topic from the perspectives of language, logic, and artificial intelligence [9, 14, 31, 35]. There has also been an increased interest from the computer vision community to synergise with cognitively motivated methods for perceptual grounding and inference with visual imagery [23, 37]. We posit that knowledge representation and reasoning can serve a crucial role for the development of next-generation methods and tools for large-scale experiments in visual perception in cognitive science and psychology. Driven by this, our research has laid out the conceptual, formal, and computational foundations for a general, declarative model of representing and reasoning with deep semantics about visuo-spatial narrative primitives identifiable with respect
to a broad-based interpretation of “the moving image”. Our narrative model and approach can directly provide the foundations that are needed for the development of novel assistive technologies in areas where high-level qualitative analysis and perceptual sensemaking of dynamic visuo-spatial imagery are central.

ACKNOWLEDGEMENTS

Tobii Pro (Sweden) donated eye-trackers | J. Sutton (Cal Poly, USA) helped with visual processing | S. Goyal (Amity University, IN) supported film experiments and data collection, and programming of clustering algorithms.

References


3.2. Declarative Reasoning about Embodied Interaction

Embodied grounding refers to the ability to link high-level conceptual knowledge, e.g., pertaining to vision, touch, and bodily motion, to (spatio-temporal) experiences in everyday life (Hampe and Grady, 2008). In this context, image schemas refer to abstract recurring structures within human cognitive processes, establishing patterns of thought, understanding, and reasoning (George Lakoff and Johnson, 1980, 1999; Mandler, 1992; Mandler and Pagán Cánovas, 2014). In this thesis we address embodied grounding in the context of cognitive robotics, where image schematic representations of embodied human-object interactions provide an abstract, pre-linguistic cognitive structure (Mandler, 1992; Mandler and Pagán Cánovas, 2014) for the description of perceptual events, e.g., pertaining to space, motion, actions, events, change.

Within the context of deep visuospatial semantics, we posit that image schematic visuospatial concepts can be characterised using declarative abstractions of space, time, space-time, motion (as available within the theory of space and motion $\Sigma$, Section 2.2), and may provide a general foundation for image schematic perceptual analysis in developmentally driven cognitive interaction technologies, such as cognitive robotic systems. In particular, we present grounded representations of embodied interactions within the area of everyday activities, and showcase qualitative space and motion as an intermediate (image schematic) abstraction layer for the processing of dynamic spatial language in robot interaction scenarios.

Example 3. RELATIONAL (ACTIVITY) GROUNDING

Consider Fig. 3.3, consisting of a sample human activity —“making a cup of tea”— as captured from an egocentric viewpoint with a head-mounted RGB-D capture device. From a commonsense viewpoint, this activity may be represented as a sequence of dynamic visuospatial interactions, such as the following:

- opening the tea-box, removing a tea-bag from the box and putting the tea-bag inside a tea-cup filled with water while holding the tea-cup.

Such interactions correspond to high-level spatial and temporal relationships between the agent and other involved objects, constituting an embodied grounding, e.g., involving conceptual representations of contact and containment that hold across specific time-intervals. In this context, manipulation and control actions $((\theta_1(\vec{t})), ... \theta_n(\vec{t}))$ cause state transitions in the world, modelled as changes in the spatio-temporal relations amongst involved domain entities.

Declarative Modelling of Human Interaction. Dynamics of activities and interactions are represented using space-time histories ($STH$) of involved people and objects, and are characterised based on spatio-temporal relations and motion patterns, as part of the theory about space and motion ($\Sigma$). These ontological characterisations serve as
a bridge between high-level conceptual categories (e.g., pertaining to human-object interactions) on the one-hand, and low-level / quantitative sensory-motor data on the other (Example 3). In this regard, we developed a declarative model of the human body grounded in 3D-data of skeleton joints and body-parts obtained from combined video (RGB) and depth (D) (RGB-D) sensing, and abstracted using basic entities in $E$, i.e., body-parts are abstracted using regions and line-segments, and joints are abstracted using points. Within the declarative model of the human body, body pose can be modelled based on the spatio-temporal configuration of the body-parts (using spatio-temporal relationships in $R$), i.e., the position of body-parts and their angle at specific skeleton joints. Complex human interaction may then be abstracted based on changes in these configurations. From the viewpoint of embodied human-object interactions we then characterise everyday human activities based on relational structures and motion pattern, pertaining to embodied concepts such as containment, attachment, support, etc.

**Space and Motion in Natural Language.** We target space and motion in natural language in the context of interacting robots that use dynamic spatial language to describe and communicate about perceived visuospatial dynamics. Understanding and externalising such spatial language and the involved spatio-temporal relations is a key ability of robots in order to interact with humans and other robots, e.g. when faced with commands, and descriptions. For instance, robots have to be able to produce and parse
Example 4. MOVEMENT DESCRIPTION

Given the scene depicted in Figure 3.4, the robots need to be able to generate and understand descriptions, similar to the one below, describing the perceived movement of the green block.

The green block starts moving at the right of the green region. The green block moves straight across the green region. The green block moves curved across the red region. The green block moves straight into the white region. The green block stops moving inside the white region.

Such descriptions contain a multitude of (embodied) visuospatial concepts, which need to have a semantic meaning in terms of involved space and motion, in order to support generation and interpretation of dynamic spatial relations in natural language, with respect to the visually perceived scene.
incomplete or possibly wrong observations, coming from a potentially noisy perception system (Spranger, Loetzsch, and Steels, 2012). To this end, the theory of space and motion $\Sigma$ provides an abstraction layer for reasoning about the spatial configuration and spatio-temporal dynamics within the perceived scene, e.g., for perspective transformation (between the two robots), or interpolating and correcting incomplete / faulty observations. This enables the robots to generate meaningful language descriptions of the observed dynamics and parse such descriptions, in order to compare them with the own observations.

The described method is applied and evaluated using a language game, where two robots are situated in an environment with a moving block, and one robot (the speaker) has to describe the perceived scene to the other robot (the listener), who has to either agree or disagree on the description. The representational focus in this setting is on spatial dynamics of the moving block with respect to the environment, based on image schematic abstractions using containment and source-path-goal schemes. Spatio-temporal relation, e.g., topology, extrinsic and intrinsic orientation, are used to characterise source and goal locations; motion patterns and motion characteristics, e.g., moving_into, moving_out_of, are used to describe the path of the block.

INCLUDED PUBLICATIONS: Declarative visuospatial reasoning about embodied interaction was published in various publications, out of which we include the following two in this thesis:

- A conference contribution published at the Iberian Robotics Conference (ROBOT 2017) (Suchan and Bhatt, 2017b, §4), focusing on grounded representations of human interactions. This work was also presented at the Vision in Practice on Autonomous Robots (ViPAR) workshop held at the International Conference on Computer Vision (ICCV 2017) (Suchan and Bhatt, 2017a).

- A conference paper published at the International Joint Conference on Artificial Intelligence (IJCAI 2016) (Spranger, Suchan, and Bhatt, 2016, §5), focusing on declarative abstractions of space and motion for natural language interactions with robots. This work was conducted in collaboration with Michael Spranger, partly during a research visit at the Sony Computer Science Laboratory (Sony CSL), Tokyo, Japan. Preliminary aspects are published in the Pacific Rim Conference on Artificial Intelligence (PRICAI 2014) (Spranger, Suchan, Bhatt, and Eppe, 2014).

Copies of the above stated publications in the given order follow in the next sections.
This conference paper was published at the Iberian Robotics Conference (ROBOT 2017). It explores the use of declarative semantics for grounding embodied human-object interactions in the context of cognitive robotics and everyday activities. Towards this, the paper develops a declarative characterisation and formalisation of human interactions, involving a declarative model of human pose and human-object interactions based on the general theory of space and motion. A particular focus is on the representation of embodied human interaction (e.g., body motion and pose, environmental interaction, etc.) based on relational visuospatial structure (i.e., spatio-temporal relations, motion patterns, and events) directly grounded within visuospatial imagery obtained from the robots perception system.

The paper demonstrates a declarative model of human interaction with select examples of everyday activities, focusing on human-object interaction data obtained from RGB-D sensors. In this context embodied grounding and declarative spatio-temporal reasoning about everyday activities is highlighted.
Deep Semantic Abstractions of Everyday Human Activities
On Commonsense Representations of Human Interactions

Jakob Suchan¹ and Mehul Bhatt¹,²

¹ Spatial Reasoning. www.spatial-reasoning.com
EASE CRC: Everyday Activity Science and Engineering., http://ease-crc.org
University of Bremen, Germany

² Machine Perception and Interaction Lab., https://mpi.aass.oru.se
Centre for Applied Autonomous Sensor Systems (AASS)
Orebro University, Sweden

Abstract. We propose a deep semantic characterisation of space and motion categorically from the viewpoint of grounding embodied human-object interactions. Our key focus is on an ontological model that would be adept to formalisation from the viewpoint of commonsense knowledge representation, relational learning, and qualitative reasoning about space and motion in cognitive robotics settings. We demonstrate key aspects of the space & motion ontology and its formalisation as a representational framework in the backdrop of select examples from a dataset of everyday activities. Furthermore, focussing on human-object interaction data obtained from RGBD sensors, we also illustrate how declarative (spatio-temporal) reasoning in the (constraint) logic programming family may be performed with the developed deep semantic abstractions.

1 Introduction

Cognitive robotics technologies and machine perception & interaction systems involving an interplay of space, dynamics, and (embodied) cognition necessitate capabilities for explainable reasoning, learning, and control about space, events, actions, change, and interaction [Bhatt, 2012]. A crucial requirement in this context pertains to the semantic interpretation of multi-modal human behavioural data [Bhatt, 2013, Bhatt and Kersting, 2017], with objectives ranging from knowledge acquisition and data analyses to hypothesis formation, structured relational learning, learning by demonstration etc. Towards this, the overall focus & scope of our research is on the processing and semantic interpretation of dynamic visuo-spatial imagery with a particular emphasis on the ability to abstract, reason, and learn commonsense knowledge that is semantically founded in qualitative spatial, temporal, and spatio-temporal relations and patterns. We propose that an ontological characterisation of human-activities — e.g., encompassing (embodied) spatio-temporal relations and motion patterns — serves
as a bridge between high-level conceptual categories (e.g., pertaining to human-object interactions) on the one-hand, and low-level / quantitative sensory-motor data on the other.

**Deep Semantics – The Case of Dynamic Visuo-Spatial Imagery**

The high-level semantic interpretation and qualitative analysis of dynamic visuo-spatial imagery requires the representational and inferential mediation of commonsense abstractions of *space, time, action, change, interaction* and their mutual interplay thereof. In this backdrop, *deep visuo-spatial semantics* denotes the existence of declaratively grounded models — e.g., pertaining to *space, time, space-time, motion, actions & events, spatio-linguistic conceptual knowledge* — and systematic formalisation supporting capabilities such as: (a) mixed quantitative qualitative spatial inference and question answering (e.g., about consistency, qualification and quantification of relational knowledge); (b) non-monotonic spatial reasoning (e.g., for abductive explanation); (c) relational learning of spatio-temporally grounded concepts; (d) integrated inductive-abductive spatio-temporal inference; (e) probabilistic spatio-temporal inference; (f) embodied grounding and simulation from the viewpoint of cognitive linguistics (e.g., for knowledge acquisition and inference based on natural language).

Recent perspectives on deep (visuo-spatial) semantics encompass methods for declarative (spatial) representation and reasoning — e.g., about *space and motion* — within frameworks such as constraint logic programming (rule-based spatio-temporal inference [Bhatt et al., 2011b, Suchan et al., 2014]), answer-set programming (for non-monotonic spatial reasoning [Walega et al., 2015, Bhatt and Loke, 2008]), description logics (for spatio-terminological reasoning [Bhatt et al., 2009]), inductive logic programming (for inductive-abductive spatio-temporal learning [Dubba et al., 2011, 2015]) and other specialised forms of commonsense reasoning based on expressive action description languages for modelling *space, events, action, and change* [Bhatt, 2012, Bhatt and Loke, 2008].

In general, deep visuo-spatial semantics driven by declarative spatial representation and reasoning pertaining to dynamic visuo-spatial imagery is relevant and applicable in a variety of cognitive interaction systems and assistive technologies at the interface of (spatial) language, (spatial) logic, and (visuo-spatial) cognition.

**Deep Semantics, and Reasoning about Human-Robot Interactions**

The starting point of our work is from formal commonsense representation and reasoning techniques developed in the field of Artificial Intelligence. Here, the core focus of the overall research goal is on the question:

*How can everyday activity tasks be formally represented in terms of spatio-temporal descriptions (that are augmented by knowledge about objects and environments) such that it enables robotic agents to execute everyday manipulation tasks appropriately?*
We particularly focus on an ontological and formal characterisation of space and motion from a human-centered, commonsense formal modeling and computational viewpoint, i.e., space, as it is interpreted within the AI subdiscipline of knowledge representation and reasoning, commonsense reasoning, spatial cognition & computation, and more broadly, within spatial information theory [Aiello et al., 2007, Bhatt et al., 2011a, Bhatt, 2012, Bhatt et al., 2013, Cohn and Renz, 2007, Renz and Nebel, 2007]. Whereas the main focus of this paper is on the ontological and representational aspects, we emphasise that this is strongly driven by computational considerations focussing on: (a) developing general methods and tools for commonsense reasoning about space and motion categorically from the viewpoint of commonsense cognitive robotics in general, but human-object interactions occurring in the context of everyday activities in particular; (b) founded on the established ontological model, developing models, algorithms and tools for reasoning about space and motion, and making them available as extensions knowledge representation (KR) based declarative spatio-temporal reasoning systems, e.g., constraint logic programming based CLP(QS) [Bhatt et al., 2011b], or answer-set programming based ASPMT(QS) [Walega et al., 2015].

2 Commonsense Reasoning about Space and Change: Background and Related Work

Commonsense spatio-temporal relations and patterns (e.g. left, touching, part of, during, collision) offer a human-centered and cognitively adequate formalism for logic-based automated reasoning about embodied spatio-temporal interactions involved in everyday activities such as flipping a pancake, grasping a cup, or opening a tea box [Bhatt et al., 2013, Worgotter et al., 2012, Spranger et al., 2014, 2016].

Qualitative, multi-modal, and multi-domain\(^3\) representations of spatial, temporal, and spatio-temporal relations and patterns, and their mutual transitions can

\(^3\) Multi-modal in this context refers to more than one aspect of space, e.g., topology, orientation, direction, distance, shape. Multi-domain denotes a mixed domain ontology involving points, line-segments, polygons, and regions of space, time, and space-time [Hazarika, 2005a].
provide a mapping and mediating level between human-understandable natural language instructions and formal narrative semantics on the one hand [Eppe and Bhatt, 2013, Bhatt et al., 2013], and symbol grounding, quantitative trajectories, and low-level primitives for robot motion control on the other (see Fig. 1). By spatio-linguistically grounding complex sensory-motor trajectory data (e.g., from human-behaviour studies) to a formal framework of space and motion, generalized (activity-based) qualitative reasoning about dynamic scenes, spatial relations, and motion trajectories denoting single and multi-object path & motion predicates can be supported [Eschenbach and Schill, 1999]. For instance, such predicates can be abstracted within a region based 4D space-time framework [Hazarika, 2005a, Bennett et al., 2000a,b], object interactions [Davis, 2008, 2011], and spatio-temporal narrative knowledge [Tyler and Evans, 2003, Eppe and Bhatt, 2013, Davis, 2013]. An adequate qualitative spatio-temporal representation can therefore connect with low-level constraint-based movement control systems of robots [Bartels et al., 2013], and also help grounding symbolic descriptions of actions and objects to be manipulated (e.g., natural language instructions such as cooking recipes [Tellex, 2010]) in the robots perception.

3 Embodied Interactions in Space-Time: Towards Commonsense Abstractions of Everyday Activities

3.1 Humans, Objects, and Interactions in Space-Time

Activities and interactions are characterised based on visuo-spatial domain-objects \( \mathcal{O} = \{ o_1, o_2, \ldots, o_i \} \) representing the visual elements in the scene, i.e, people and objects.

The Qualitative Spatio-Temporal Ontology (QS) is characterised by the basic spatial and temporal entities (\( \mathcal{E} \)) that can be used as abstract represen-
Deep Semantic Abstractions of Everyday Human Activities

Fig. 2: Declarative Model of Human-Body Posture

Fig. 2: Declarative Model of Human-Body Posture

Annotations of domain-objects and the relational spatio-temporal structure ($\mathcal{R}$) that characterises the qualitative spatio-temporal relationships amongst the entities in ($\mathcal{E}$). Towards this, domain-objects ($\mathcal{O}$) are represented by their spatial and temporal properties, and abstracted using the following basic spatial entities:

- **points** are triplets of reals $x, y, z$;
- **oriented-points** consisting of a point $p$ and a vector $v$;
- **line-segments** consisting of two points $p_1, p_2$ denoting the start and the end point of the line-segment;
- **poly-line** consisting of a list of vertices (points) $p_1, ..., p_n$, such that the line is connecting the vertices is non-self-intersecting;
- **polygon** consisting of a list of vertices (points) $p_1, ..., p_n$, (spatially ordered counter-clockwise) such that the boundary is non-self-intersecting;

and the temporal entities:

- **time-points** are a real $t$
- **time-intervals** are a pair of reals $t_1, t_2$, denoting the start and the end point of the interval.

The dynamics of human activities are represented by 4-dimensional regions in space-time (sth) representing people and object dynamics by a set of spatial entities in time, i.e. $STH = (\varepsilon_{t_1}, \varepsilon_{t_2}, \varepsilon_{t_3}, ..., \varepsilon_{t_n})$, where $\varepsilon_{t_1}$ to $\varepsilon_{t_n}$ denotes the spatial primitive representing the object $o$ at the time points $t_1$ to $t_n$.

**Spatio-Temporal Characteristics of Human Activities**

Dynamics of human activities are abstracted using 4-dimensional regions in space-time, i.e.
space-time histories (sth) representing people and object dynamics. Based on the space-time histories of domain-objects, we define the following functions for spatio-temporal properties of objects:

- **position**: $O \times T \rightarrow \mathbb{R} \times \mathbb{R} \times \mathbb{R}$, gives the 3D position $(x,y,z)$ of an object $o$ at a time-point $t$;
- **size**: $O \times T \rightarrow \mathbb{R}$, gives the size of an object $o$ at a time-point $t$;
- **distance**: $O \times O \times T \rightarrow \mathbb{R}$, gives the distance between two objects $o_1$ and $o_2$ at a time-point $t$;
- **angle**: $O \times O \times T \rightarrow \mathbb{R}$, gives the angle between two objects $o_1$ and $o_2$ at a time-point $t$;

for static spatial properties, and

- **movement velocity**: $O \times T \times T \rightarrow \mathbb{R}$, gives the amount of movement of an object $o$ between two time-points $t_1$ and $t_2$;
- **movement direction**: $O \times T \times T \rightarrow \mathbb{R}$, gives the direction of movement of an object $o$ between two time-points $t_1$ and $t_2$;
- **rotation**: $O \times T \times T \rightarrow \mathbb{R}$, gives the rotation of an object $o$ between two time-points $t_1$ and $t_2$;

for dynamic spatio-temporal properties.

Spatio-temporal relationships ($R$) between the basic entities in $E$ may be characterised with respect to arbitrary spatial and spatio-temporal domains such as mereotopology, orientation, distance, size, motion, rotation (see Table 1 for a list of considered spatio-temporal abstractions).
Deep Semantic Abstractions of Everyday Human Activities

Table 2: Sample Interactions Involved in Everyday Human Activities

<table>
<thead>
<tr>
<th>Interaction</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>pick up(P;O)</td>
<td>a person P picks up an object O.</td>
</tr>
<tr>
<td>put down(P;O)</td>
<td>a person P puts down an object O.</td>
</tr>
<tr>
<td>reach for(P;O)</td>
<td>a person P is reaching for an object O.</td>
</tr>
<tr>
<td>passing over(P1;P2;O)</td>
<td>a person P1 is passing an object O to another person P2.</td>
</tr>
</tbody>
</table>

Declarative Model of Human Body Pose

The human body is represented using a declarative model of the human body (see Fig. 2), within this model we ground the human body in 3d-data of skeleton joints and body-parts obtained from RGB-D sensing. Body-parts may be abstracted using regions and line-segments, and joints may be abstracted using points. As such, Body pose can be declaratively abstracted by the spatio-temporal configuration of the body-parts, using the position of body-parts and the angle between skeleton joints.

Spatio-temporal fluents are used to describe properties of the world, i.e. the predicates holds-at(φ, t) and holds-in(φ, δ) denote that the fluent φ holds at time point t, resp. in time interval δ. Fluents are determined by the data from the depth sensing device and represent qualitative relations between domain-objects, i.e. spatio-temporal fluents denote, that a relation r ∈ R holds between basic spatial entities ε of a space-time history at a time-point t. Dynamics of the domain are represented based on changes in spatio-temporal fluents (see Fig. 3), e.g., two objects approaching each other can be defined as follows.

\[
\text{holds-in(approaching}(o_i, o_j), \delta) \supset \text{during}(t_i, \delta) \land \text{during}(t_j, \delta) \land \text{before}(t_i, t_j) \land (\text{distance}(o_i, o_j, t_i) > \text{distance}(o_i, o_j, t_j)).
\] (1)

Interactions

Interactions Θ = {θ1, θ2, ..., θl} describe processes that change the spatio-temporal configuration of objects in the scene, at a time point t or in a time interval δ; these are defined by the involved spatio-temporal dynamics in terms of changes in the status of st-histories caused by the interaction, i.e. the description consists of (dynamic) spatio-temporal relations of the involved st-histories, before, during and after the interaction (See Table 2 for exemplary interactions). We use occurs-at(θ, t), and occurs-in(θ, δ) to denote that an interaction θ occurred at a time point t or in an interval δ, e.g., a person reaching for an object can be defined as follows.

\[
\text{holds-in(reach for}(o_i, o_j), \delta) \supset \text{person}(o_i) \land \\
\text{holds-in(approaching(body part}(\text{hand}, o_i), o_j), \delta) \land \\
\text{holds-in(touches(body part}(\text{hand}, o_i), o_j), \delta) \land \\
\text{meets}(\delta_i, \delta_j) \land \text{starts}(\delta_i, \delta) \land \text{ends}(\delta_j, \delta).
\] (2)
### 4 Application: Grounding of Everyday Activities

We demonstrate the above model for grounding everyday activities in perceptual data obtained from RGB-D sensing. The model has been implemented within (Prolog based) constraint logic programming based on formalisations of qualitative space in CLP(QS) [Bhatt et al., 2011b]. Using the presented model it is possible to generate grounded sequences of interactions performed within the course of an activity.

The presented activity is part of a larger dataset on everyday human activities (see Table 3), including RGB and RGB-D data for from different viewpoints of human-human, and human-object interactions.

**Sample Activity: “Pass Cup of Water”** The activity of passing a cup of water is characterised with respect to the interactions between the humans and their environment, i.e. objects the human uses in the process of passing the cup. Each of these interactions is defined by its spatio-temporal characteristics, in terms of changes in the spatial arrangement in the scene (as described in Sec. 3). As a result we obtain a sequence of interactions performed within the track of the particular instance of the activity, grounded in the spatio-temporal dynamics of the scenario. As an example consider the sequence depicted in fig. 1, the interactions in this sequence can be described as follows:

Person1 reaches for the cup, picks up the cup, and moves the hand together with the cup towards Person2. Person2 grasps the cup and Person1 releases the cup.

The data we obtain from the RGB-D sensor consists of 3D positions of skeleton joints for both persons and the tabletop objects for each time-point.

---

Table 3: Exemplary Activities from the Dataset of Human Activities

<table>
<thead>
<tr>
<th>Activities</th>
<th>Interactions</th>
<th>Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Making sandwich, Making tea, Making salad, Making cereals</td>
<td>cut Cucumbers, Onions, Tomatoes, Sandwich</td>
<td>cut Cucumbers, Onions, Tomatoes, Sandwich</td>
</tr>
<tr>
<td></td>
<td>pour Dressing on the plate, Tea in the cup, Juice in the glass, Water in the glass, Coffee in the cup, Cereal in the bowl, Milk in the bowl</td>
<td>pour Dressing on the plate, Tea in the cup, Juice in the glass, Water in the glass, Coffee in the cup, Cereal in the bowl, Milk in the bowl</td>
</tr>
<tr>
<td></td>
<td>pass Cup of water / coffee / tea</td>
<td>pass Cup of water / coffee / tea</td>
</tr>
<tr>
<td></td>
<td>pick Cup from the cupboard; Slices of bread from the packet, Vegetables/Fruits from the basket, Basket from the kitchen plane, Tea bag from the box</td>
<td>pick Cup from the cupboard; Slices of bread from the packet, Vegetables/Fruits from the basket, Basket from the kitchen plane, Tea bag from the box</td>
</tr>
<tr>
<td></td>
<td>put Sugar in the cup, Tea bag in the cup</td>
<td>put Sugar in the cup, Tea bag in the cup</td>
</tr>
</tbody>
</table>

---

4 RGB-D Data (video, depth, body skeleton): We collect data using Microsoft Kinect v2 which provides RGB and depth data. The RGB stream has a resolution of 1920x1080 pixel at 30 Hz and the depth sensor has a resolution of 512x424 pixels at 30 Hz. Skeleton tracking can track up to 6 persons with 25 joints for each person. Further we use the point-cloud data to detect objects on the table using tabletop object segmentation.
Grounded Interaction Sequence

Based on the sensed body-pose data and the detected objects, a sequence of interactions can be queried from the example sequences using Prolog's interactive query answering mode.

?- grounded_interaction(occurs_in(Interaction, Interval), Grounding).

This results in all interactions identified in the example sequence and their respective grounding with respect to the spatio-temporal dynamics constituting the interaction,

```
Interactive = pick_up(person(P), object(cup)),
Interval = interval(t1, t3),
Grounding =
\[\text{occurs}_\text{at} (\text{grasp(body_part(right_hand, person(id(0))), object(cup)), timepoint(t1)},
\text{holds}_\text{in}(\text{attached(body_part(right_hand, person(id(0))), object(cup)), interval(t2,t6)},
\text{holds}_\text{in}(\text{move_up(body_part(right_hand, person(id(0))), interval(t2,t3))});
```

```
Interactive = pass_over(person(P), person(Q), object(cup)),
Interval = interval(t4, t7),
Grounding =
\[\text{holds}_\text{in}(\text{approaching(body_part(right_hand, person(id(0))), person(id(1))), interval(t4,t5)},
\text{holds}_\text{in}(\text{approaching(body_part(right_hand, person(id(1))), object(cup)), interval(t4,t5)},
\text{occurs}_\text{at} (\text{grasp(body_part(right_hand, person(id(1))), object(cup)), timepoint(t6)},
\text{occurs}_\text{at} (\text{release(body_part(right_hand, person(id(0))), object(cup)), timepoint(t7)},
\text{occurs}_\text{at}(object(id(0)), type(cup), pos_3d(point(0.667643,-0.213097,1.83488)), time_point(2385869011))].
```

This grounding of the activity may be used for interpretation and learning from the observed activities and the involved spatio-temporal dynamics, e.g., in the example above the person is passing the cup over the laptop, which is safe when the cup is empty, but in the case that the cup is filled with water one would pass it around the laptop.

5 Summary and Outlook

Deep semantics denotes the existence of declaratively grounded models — e.g., pertaining to space, time, space-time, motion, actions & events, spatio-linguistic conceptual knowledge — and systematic formalisation supporting KR-based capabilities such as abstraction, learning, reasoning, embodied simulation. Rooted in this concept of deep (visuo-spatial) semantics, this paper presents an ontological and formal representational framework aimed at grounding embodied human-object interactions in a commonsense cognitive robotics setting. The model is illustrated with select RGBD datasets corresponding to representative activities from a larger dataset of everyday activities; as preliminary application, we also
show how the formal model can be directly applied for commonsense reasoning with constraint logic programming, with a particular focus on space-time histories and motion patterns.

Immediate next steps involve expanding the scope of everyday activities from table-top or kitchen based scenarios to situations involving indoor mobility and abstractions for the representation of social interactions between humans and mobile agents. This will enable to further enhance the scope of the ontology and corresponding spatio-temporal relations. Furthermore, the demonstrated applications of the ontology of space & motion are currently preliminary; next steps here involve integration with state of the art robot control platforms such as ROS; this will be accomplished via integration into the ExpCog commonsense cognition robotics platform for experimental / simulation purposes, and within openEASE as a state of the art cognition-enabled control of robotic control platform for real robots.5

Acknowledgements. We acknowledge funding by the Germany Research Foundation (DFG) via the Collaborative Research Center (CRC) EASE – Everyday Activity Science and Engineering (http://ease-crc.org). This paper builds on, and is a condensed version of, a workshop contribution [Suchan and Bhatt, 2017] at the ICCV 2017 conference. We also acknowledge the support of Vijayanta Jain in preparation of part of the overall activity dataset; toward this, the help of Omar Moussa, Hubert Kloskoski, Thomas Hudkovic, Vyoom Kelkar as subjects is acknowledged.

Bibliography


5 ExpCog – http://www.commonsenserobots.org
openEASE – http://ease.informatik.uni-bremen.de/openease/


5 - Robust Natural Language Processing

Published in:
International Joint Conference on Artificial Intelligence (IJCAI 2016)

Citation:

This conference paper was published at the International Joint Conference on Artificial Intelligence (IJCAI 2016). The presented work was conducted in collaboration with Michael Spranger (Sony CSL, Tokyo, Japan) and focuses on dynamic spatio-temporal relations in natural language robot interactions. Here, the theory about space and motion is applied in the context of embodied natural language generation and understanding for robotic interactions. From the viewpoint of this thesis the core aspect of this work is in providing commonsense abstractions of space and motion (implemented in constraint logic programming and available via declarative query answering) to the robots to interpret and reason about spatial dynamics in the perceived scenes. In particular, declarative abstractions of space and motion are used by the robots to represent the visually perceived scenes and to facilitate reasoning about the scene dynamics, e.g., for reasoning about perspective, or for reasoning about incomplete event sequences.

The system is demonstrated and evaluated focusing on robot-robot interactions, where one robot describes an environment of moving blocks using English phrases that include dynamic and static spatial relations, and the other robot has to parse this sentence and compare it to the own perception of the scene. Empirical results show that using declarative abstractions of space and motion improves the system’s capabilities to generate and interpret scene descriptions in the presence of visual perception errors and missing information.
Robust Natural Language Processing - Combining Reasoning, Cognitive Semantics and Construction Grammar for Spatial Language

Michael Spranger\textsuperscript{1}, Jakob Suchan\textsuperscript{2} and Mehul Bhatt\textsuperscript{2}

\textsuperscript{1}Sony Computer Science Laboratories Inc., Tokyo, Japan, michael.spranger@gmail.com
\textsuperscript{2}University of Bremen, Bremen, Germany, \{jsuchan|bhatt\}@informatik.uni-bremen.de

Abstract

We present a system for generating and understanding of dynamic and static spatial relations in robotic interaction setups. Robots describe an environment of moving blocks using English phrases that include spatial relations such as “across” and “in front of”. We evaluate the system in robot-robot interactions and show that the system can robustly deal with visual perception errors, language omissions and ungrammatical utterances.

1 Introduction

Spatial language is no doubt important for robots, if they need to be able to communicate with humans. For instance, robots need to be able to understand descriptions such as the following.

(1) The block moves across the red region.

Example 1 focusses on the path of the object [Croft and Cruse, 2004]. English speakers also have other means of conceptualizing movement events. They can, for instance, focus on the source of the movement or the goal.

(2) The block moves from left of you, to right of me.

These examples include various aspects of English language syntax, semantics and pragmatics [Levinson, 2003; Svorou, 1994]. A complete language processing system for robots needs to be able to understand and also generate such utterances.

Importantly, natural language processing systems need to be robust against various sources of errors. Humans invariably make mistakes and robots need to be able to deal with missing or misunderstood words, grammatical errors etc. At the same time, visual processing of scenes is not perfect. Objects might be occluded and errors in visual tracking might impact tracked paths, and visual recognition of events. Robustness against visual and language perturbations is crucial.

In this paper, we present a complete system that allows robots to describe and understand descriptions of spatial scenes involving movement events (see Figure 1). The system is robust against perceptual errors, missing words and grammatical errors.

2 Related Work

Earliest systems for spatial language [Retz-Schmidt, 1988; Gapp, 1995; Skubic et al., 2004] showed how artificial agents can understand static spatial relations such as “front”, “back”. This work has continued. We have now various ways of modeling static spatial relations: proximity fields for proximal relations [Kelleher et al., 2006], prototypes for projective and absolute spatial relations [Spranger and Pauw, 2012]. Models of static spatial relations are interesting but they only cover relations not encoding dynamic qualities.

Recent models of dynamic spatial relations use semantic fields [Fasola and Mataric, 2013] and probabilistic graphi-
Figure 2: Spatial setup. Left: scene model extracted by the left robot. Right: scene model computed by the right robot. Estimated movement of the block (circle) is visualized through opacity. The starting point has a lower opacity (alpha). Regions are visualized by colored quadrangles. The blue square shows position, orientation of the box. Arrows are robots. robot-1 is the origin of the coordinate system, robot-2 position and orientation of the other robot.

### 3 Grounded Spatial Language Processing

Two robots interact in an environment such as the one shown in Figure 2. For the experiments discussed in this paper, we used Sony humanoid robots. The vision system of these robots fuses information from the robot’s camera (30 fps) with proprioceptive sensors distributed across the body (gyroscope, internal body model from motor position sensors), in order to single out and track various objects in the environment [Spranger et al., 2012a].

The environment features four types of objects: blocks, boxes, robots and regions. The vision system extracts the objects (as blobs) from the environment and computes a number of raw, continuous-valued features such as x, y position, width, and height and colour values (YCbCr). Objects are tracked over time and assigned unique identifiers as long as there is spatio-temporal continuity. For instance, the green block has been given the arbitrary id obj-755 by the left robot.

#### 3.1 Reasoning about Space and Motion

The robots generate qualitative representations of the spatio-temporal dynamics in the scene as perceived by their vision system. Towards this, we use a general theory of space and motion implemented based on CLP(QS) [Bhatt et al., 2011] - a declarative spatial reasoning framework, which implements declarative spatial relations in constraint logic programming within the PROLOG programming environment. We use the framework for defining events grounded in the visual observations of the robots, using qualitative spatial and temporal relations between objects in the scene, i.e. topology, orientation, and movement.

In order to reason about the perceived dynamics of scenes (for example the scene in Figure 2), we generate sequences of movement events based on the perceptual data of the robots, as depicted in Figure 3. Towards this, objects are represented using qualitative abstractions of spatial properties, e.g. position, orientation, extend in space, using primitives such as regions, points, oriented points, line segments. Perceived spatio-temporal dynamics, i.e. the movement of the block is represented by the source and the goal of the movement, and the path, on which the object moves from the source to the goal. For describing the movement and involved movement events, we use spatio-temporal relations, e.g. for representing the source and goal locations of the movement with respect to the observing robots or the characteristics of the path.

The spatial configuration of objects in the scene is represented using n-ary spatial relations $R = \{r_1, r_2, ..., r_n\}$, in particular, we use topological relations of the RCC8 fragment of the RCC calculus [Randell et al., 1992], $R_{top} \equiv \{dc, ec, po, eq, tpp, ntpp, tpp^{-1}, ntppp^{-1}\}$ and orientation relations of the $LR$ calculus [Scivos and Nebel, 2005] $R_{orient} \equiv \{l, r, i, s, e, f, b\}$. Predicates holds-at($\phi , r, t$) and holds-in($\phi , r, \delta$) are used to denote that the fluent $\phi$ has the value $t$ at time point $t$, resp. in the time interval $\delta$. Movement events are used to describe spatio-temporal dynamics of the perceived scene, i.e. how the spatial configuration of objects changes during the movement of the block. We use the predicate occurs-in($\theta , \delta$) to denote that an event $\theta$ occurred in a time interval $\delta$.

<table>
<thead>
<tr>
<th>Spatial Relations</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>dc</td>
<td>Disconnected</td>
</tr>
<tr>
<td>ec</td>
<td>Equivalent</td>
</tr>
<tr>
<td>po</td>
<td>Part of</td>
</tr>
<tr>
<td>eq</td>
<td>Equiangular</td>
</tr>
<tr>
<td>tpp</td>
<td>Temporal part of</td>
</tr>
<tr>
<td>ntpp</td>
<td>Non-temporal part of</td>
</tr>
<tr>
<td>tpp^{-1}</td>
<td>Temporal part of inverse</td>
</tr>
<tr>
<td>ntppp^{-1}</td>
<td>Non-temporal part of inverse</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Orientation Relations</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>l</td>
<td>Left</td>
</tr>
<tr>
<td>r</td>
<td>Right</td>
</tr>
<tr>
<td>i</td>
<td>Inferior</td>
</tr>
<tr>
<td>s</td>
<td>Superior</td>
</tr>
<tr>
<td>e</td>
<td>External</td>
</tr>
<tr>
<td>f</td>
<td>Vertical</td>
</tr>
<tr>
<td>b</td>
<td>Horizontal</td>
</tr>
</tbody>
</table>
In particular, movement events are defined by spatio-temporal relations holding between the involved objects and changes within these relations, happening as a part of the event, using the relations of Allen’s interval algebra [Allen, 1983] (before, after, during, contains, starts, started_by, finishes, finished_by, overlaps, overlapped_by, meets, met_by, equal) for representing temporal aspects of the event. E.g. the event moves into, representing that a block moves into a region is defined as follows.

\[
\text{occurs-in}(\text{moves}_\text{into}(o_1, a_1), \delta) \supset
\begin{align}
&\text{holds-at}(\phi_{\text{top}}(\text{position}(o_1), \text{region}(a_1)), \text{outside}, t_1) \\
&\text{holds-at}(\phi_{\text{top}}(\text{position}(o_1), \text{region}(a_2)), \text{inside}, t_2) \\
&\text{starts}(t_1, \delta) \land \text{finishes}(t_2, \delta) \land \text{meets}(t_1, t_2).
\end{align}
\]

Accordingly, movement events describing a range of perceivable spatial changes can be defined, e.g. moves to, moves across, etc. Complex interactions can be described by combining multiple movement events.

To describe the dynamics observed by one of the robots we generate a temporally-ordered sequence of movement events. E.g. the following Movement Sequence (Ψ) describes the movement in a scene (Figure 2), as observed by the robot to the left.

\[
\Psi \equiv \text{occurs-in}(\text{moves}_\text{into}(obj - 755, reg - 36), \delta_1) \land
\text{occurs-in}(\text{moves}_\text{out}_\text{of}(obj - 755, reg - 36), \delta_2) \land
\text{occurs-in}(\text{moves}_\text{across}(obj - 755, reg - 36), \delta_3) \land
\text{occurs-in}(\text{moves}_\text{into}(obj - 755, reg - 37), \delta_4) \land
\text{occurs-in}(\text{moves}_\text{out}_\text{of}(obj - 755, reg - 37), \delta_5) \land
\text{occurs-in}(\text{moves}_\text{across}(obj - 755, reg - 37), \delta_6) \land
\text{occurs-in}(\text{moves}_\text{into}(obj - 755, reg - 38), \delta_7).
\]

To reason about the possibility of a movement event to happen at a certain time point, we introduce predicates to describe in which spatial situations an event might happen, i.e. we use the predicate poss-at(θ, t), to describe the spatial preconditions of an event.

\[
\text{poss-at}(\text{moves}_\text{into}(o_1, a_1), t) \supset
\begin{align}
&\text{holds-at}(\phi_{\text{top}}(\text{position}(o_1), \text{region}(a_1)), \text{outside}, t). \quad (5)
\end{align}
\]

Further, we use the predicate causes(θ, φ, r) to describe how an event changes the spatial configuration in the scene.

\[
\text{causes}(\text{moves}_\text{into}(o_1, a_1), \phi_{\text{top}}(\text{position}(o_1), \text{region}(a_2)), \text{inside}).
\]

These predicates are used to reason about whether an event is a possible subsequent event given observed events.

### Mechanisms for Robustness

The reasoning system abstracts from the numerical values of the visual data stream, thereby generalizing observations. Consequently, small perceptual errors have less or no effect on computed movement events. Similarly, missing observations have little effect on the extracted movement sequence, as long as there is at least one observation for each qualitative state. For example, for moves_into only one observation outside the region and one observation inside the region is needed. Lastly, reasoning about the possibility of movement events increases the chances of agreement between two robots. E.g. if a robot observes a moves_into event in a particular region, the robot can reason, that the next possible event could be a moves_out_of event from that region. The possibility of a moves_out_of event together with the observed moves_into leads to the possibility of a moves_across event. If now he hears from the other robot that there was a moves_across event - he can conclude that this is a possible description (taking into account that there might have been perception errors).

### 3.2 Spatio-Temporal Semantics

We model the semantics of spatial phrases using a computational cognitive semantics system called Incremental Recruitment Language (IRL) [Spranger et al., 2012b]. The key idea in IRL is that semantics of natural language phrases can be modeled as a program (henceforth IRL-program) [Johnson-Laird, 1977]. The meaning of an utterance consists of an algorithm and data pointers that when executed by the hearer will lead him to identify the topic (i.e. some event or object). Figure 4 shows a graphical representation of the IRL-
program (i.e. meaning) underlying some part of the phrase from Example 1. The IRL-program consists of 1) cognitive operations (e.g. filter-by-class) implementing algorithms such as categorization and 2) semantic entities – the data that cognitive operations work with. Semantic entities can be prototypes, concepts and categories or more generally representations of the current context, as well as data exchanged between cognitive operations. They can be introduced explicitly in the network via bind-statements. The statement (bind dynamic-spatial-relation ?acr across) encodes the access to the agent-internal, dynamic spatial relation across which will be bound to the variable ?acr. Semantic entities are linked with particular parameters of cognitive operations via variables (starting with ?). In IRL-programs (meaning structures) many cognitive operations can be combined. Most relevant for this paper are the spatio-temporal aspects of these programs.

**Profiling** operations pick out aspects of movement events. We implemented Source-Path-Goal image schemas (known from Cognitive Semantics). The operation apply-path picks out the trajectory of an event. Other operations deal with the source position or goal locations (e.g. apply-source). In English source and goal are specifically marked using the prepositions “to” and “from”. Profiling operations work directly on predicates extracted by the reasoning system.

**Dynamic Spatial Relations** are concerned with describing aspects of the path of an event. Here we focus on the major relations such as “across”, “in to”, “out of”. The operation apply-dynamic-spatial-relations computes whether a event or set of events fulfills a trajectory condition, for example that the undergoer of the movement event moves across some region (all input parameters). This operation checks the object relations computed by CLP(QS).

**Static Spatial Relations** are for characterizing source and goal aspects of movement events. We implemented operations that take care of locating an object based on its position with respect to various landmark objects (robots and boxes), various frames of reference (absolute, relative and intrinsic) and various spatial relations (proximal, projective, absolute etc). The system integrates previous work [Spranger and Pauw, 2012].

**Conceptualization and Interpretation** IRL includes mechanisms for the autonomous construction of IRL-programs. Agents use these facilities in two ways. First, when the speaker wants to talk about a particular scene, he constructs an IRL-program for reaching that goal. Secondly, a listener trying to interpret an utterance will construct and evaluate programs, in order to find the best possible interpretation of the utterance (see conceptualization/interpretation in Figure 1). The processes are constrained by the particular goal given to the system. For instance, if the system needs to discriminate an object or event - it automatically selects features that are most dissimilar with other objects and events. If the goal is to describe, then features that are most similar to the object or event are selected without attention to other objects. Interpretation and conceptualization are implemented as heuristics-guided search processes that traverse the space of possible IRL-programs by automatic programming.

**Mechanisms for Robustness** The system features a number of mechanisms for robustness. For instance, the implementation of static spatial categories follows a lenient approach that increases tolerance for errors in perception [Spranger and Pauw, 2012]. The most important mechanism for the purpose of this paper though is the interpretation of partial networks. For instance, suppose the hearer only parsed a partial sentence (because of transmission errors) and can only recover a partial IRL-program. The system then tries to complete the network thereby generating various hypotheses that are tested with the perceptual data and the object relations available at that moment in time. This completion allows hearers to understand sentences even when there are utterance transmission errors and/or ungrammatical sentences.

### 3.3 Spatial Construction Grammar

In order to compute utterances for meaning (production) and meaning of utterances (parsing), we use a recent version of a computational construction grammar system called Fluid Construction Grammar (FCG) [Steels, 2011]. FCG allows to specify bidirectional mappings between meanings and utterances in the form of a single grammar. Robots operate a spatial grammar comprised of roughly 70 constructions (bidirectional rules) - primarily **lexical constructions** for basic concepts (e.g. block, box), events (e.g. move), spatial relations (e.g. along, across, into, out of), as well as a number of **phrasal constructions**.

**Constructions** The most important constructions are **lexical and phrasal**. Lexical constructions are bidirectional mappings between semantic entities and words. For instance, there is a lexical construction for “across” that maps (bind dynamic-spatial-relation ?acr across) to the stem “across”. Phrasal constructions take into account the larger syntactic and semantic context. An example is the adjective-noun-phrase construction, which looks for an adjective and a noun as well as a particular linkage of operations in the IRL-program and adds word order information. Similar constructions are implemented for determined noun phrases, prepositional phrases and verb phrases.

**Mechanisms for Robustness** The single most important robustness mechanism for the grammar is that the system applies as many constructions as possible. This is helpful when there are transmission errors, words can not be recognized and there are grammatical problems with word order etc. Even in such cases, the system will try to catch the lexical items that are recognizable in a phrase and they will be mapped to semantic entities, concepts etc. Moreover, fragments of utterances such as noun phrases that are recognizable will be processed as well. This information can be used by the semantics system to try and understand even phrases with errors.

### 4 Evaluation and Results

In order to evaluate the whole system we developed scenarios in which two robots interact with each other. Robots interact
on roughly 200 pre-recorded spatial scenes (similar to the one depicted in Figure 2). Scenes vary in spatial configurations of the two robots, objects, regions boxes etc.

In an interaction, one of the agents acts as the speaker, the other as the hearer. Roles are randomly assigned. The speaker picks some aspect of the scene and describes it to the hearer. For instance, the speaker might choose to describe the path of the moving object. The speaker describes the scene and the hearer tries to see if this is a possible description of the scene from his perspective. The interaction is a success if the hearer can agree with the description. The following details the interaction steps (see also Figure 1)

1. The robots perceive the scene and reason about spatio-temporal relations of objects.
2. The speaker conceptualizes a meaning comprised of dynamic or static spatial relations, and construal operations for describing the scene.
3. The speaker expresses the conceptualization using an English grammar. E.g., the speaker produces “the green block moves from left of you, across the red region, to right of me”.
4. The hearer parses the phrase using his English grammar and computes the meaning underlying the phrase.
5. When the hearer was able to parse the phrase or parts of the phrase, he examines the observed scene to find out whether the scene satisfies the conceptualization.
6. The hearer signals to the speaker whether he agrees with the description.
7. The interaction is a success if the hearer agrees with the speaker. Otherwise it is considered a failure.

There are a few important points about this setup. Most importantly, each robot sees the world from his perspective. This means that robots always deal with issues of perceptual deviation [Spranger and Pauw, 2012]. Robots have different viewpoints on the scene, which impacts on issues of egocentric spatial language. For instance, “the block to the left” can mean different objects depending on the viewpoint. But even on a more basic level, robots will estimate the world and its properties from their viewpoints. This leads to different estimations of distance and direction and in some cases can lead to dramatic differences in perception of the scene. The spatio-temporal continuity of objects can be disrupted, which means that events can be missed by some robot.

Another important aspect is that robots are not only interpreting but also speaking using the same system. Therefore, our setup allows us to quantify the impact of particular algorithms on the ability of robots to communicate.

### 4.1 General Evaluation

We evaluate the performance of the system on roughly 200 spatial scenes on which robots interact 10000 times. Each time one of the scenes is randomly drawn. Each time speaker and hearer are randomly assigned some perspective. The number of descriptions that can be generated for a scene is finite - in particular because agents can generate arbitrarily long descriptions. For the purpose of this paper though, we restrict generation to simple sentences that include just 1 preposition and 2 noun phrases, e.g. “the object moves into the red region” or “the object moves from left of you”.

The simplicity constraint allows us to compute all the meanings and utterances for descriptions of a scene from the viewpoint of any of the robots. In total we observed for this data set about 40 different utterances exchanged between robots. Each utterance was checked by 3 different English speakers. All of them were syntactically correct and intelligible.

For each interaction of two robots, we track 1) whether it was successful (SUCC), 2) how often the speaker was able to construe a meaning in production (CM), 3) how often the speaker produced an utterance (PU), 4) how often the hearer parsed a meaning (PM) and 5) how often the hearer was able to interpret the meaning in the current scene (IM). We also do one more check, which is whether the initial meaning that the speaker had in mind is part of the meanings recuperated by the hearer (overlap or OL).

<table>
<thead>
<tr>
<th>SUCC</th>
<th>CM</th>
<th>PU</th>
<th>PM</th>
<th>IM</th>
<th>OL</th>
</tr>
</thead>
<tbody>
<tr>
<td>.78</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
<td>.78</td>
<td>.79</td>
</tr>
</tbody>
</table>

Results show that in roughly 80% of interactions, the hearer can agree with the description. In case of failure it is more likely to be a failure of the listener to interpret the description of the speaker (IM), then 1) a speaker coming up with a description (CM), 2) speaker producing an utterance (PU), or 3) the hearer failing to parse the utterance (PM).

Upon further examination we observe that in cases where communication fails, there are perceptual problems. If we ask the hearer to conceptualize and produce utterances using his viewpoint on the world, we can see that the utterance of the speaker is not actually part of those descriptions produced by the hearer in 20% of the cases (.79 OL). The reason is that hearer and speaker in some scenes extract different events. For instance, the hearer might miss important parts of the trajectory and cannot agree to a description (for example “across the red region”).

This is also confirmed by examining F-scores for utterances and meanings. For this, we have the two robots (a and b) produce all utterances and all meanings for a scene. We then compare utterances and meanings. True positives are those utterances produced both by b and by a. False negatives are utterances produced by a AND not produced by b. False positives are utterances produced by b AND not by a.

<table>
<thead>
<tr>
<th>precision</th>
<th>recall</th>
<th>f-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>85.54</td>
<td>89.97</td>
<td>87.70</td>
</tr>
</tbody>
</table>

We can conclude that there are problems prior to language processing in how the scene is perceived and subsequently conceptualized, which leads to different utterances being produced and then false positives and false negative utterances subsequently.

### 4.2 Evaluation of Robustness

Results in the previous section beg the question how robust the system is. In further studies, we manipulated the two inputs to the system: visual information and language. Each of these can be individually perturbed to see when the system breaks down.
Table 1: Results visual perturbations

<table>
<thead>
<tr>
<th>drop</th>
<th>precision</th>
<th>recall</th>
<th>f-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>85.54</td>
<td>89.97</td>
<td>87.70</td>
</tr>
<tr>
<td>10%, both</td>
<td>85.28</td>
<td>89.97</td>
<td>87.56</td>
</tr>
<tr>
<td>25%, both</td>
<td>85.17</td>
<td>89.87</td>
<td>87.45</td>
</tr>
<tr>
<td>40%, both</td>
<td>83.97</td>
<td>89.47</td>
<td>86.63</td>
</tr>
<tr>
<td>50%, both</td>
<td>83.99</td>
<td>89.18</td>
<td>86.50</td>
</tr>
<tr>
<td>75%, both</td>
<td>77.04</td>
<td>81.47</td>
<td>79.19</td>
</tr>
<tr>
<td>10%, only-a</td>
<td>85.35</td>
<td>89.95</td>
<td>87.59</td>
</tr>
<tr>
<td>25%, only-a</td>
<td>85.12</td>
<td>90.23</td>
<td>87.60</td>
</tr>
<tr>
<td>40%, only-a</td>
<td>84.20</td>
<td>89.96</td>
<td>86.98</td>
</tr>
<tr>
<td>50%, only-a</td>
<td>83.14</td>
<td>90.25</td>
<td>86.55</td>
</tr>
<tr>
<td>75%, only-a</td>
<td>68.19</td>
<td>91.51</td>
<td>78.15</td>
</tr>
<tr>
<td>10%, only-b</td>
<td>85.51</td>
<td>89.77</td>
<td>87.59</td>
</tr>
<tr>
<td>25%, only-b</td>
<td>85.77</td>
<td>89.63</td>
<td>87.66</td>
</tr>
<tr>
<td>40%, only-b</td>
<td>86.22</td>
<td>89.72</td>
<td>87.94</td>
</tr>
<tr>
<td>50%, only-b</td>
<td>86.21</td>
<td>89.77</td>
<td>87.09</td>
</tr>
<tr>
<td>75%, only-b</td>
<td>88.87</td>
<td>73.10</td>
<td>80.22</td>
</tr>
</tbody>
</table>

Reasoning boosts success in roughly 10% of the cases and helps establish agreement in description.

**Language Perturbations** We were also interested in impact of perturbations of utterances computed by the speaker on the overall success. We looked at two manipulations: word order and missing words.

The first manipulation is to drop random words from the string the speaker has uttered (this is similar to non-understood words). So for instance, when the speaker said “the block moves into the red region”, the hearer will only see “the moves into the red region”. The second manipulation is to permute words. A sentence such as “the block moves into the red region” might be passed to the hearer as “the red moves block region the into”.

The following table shows results for 0 to 3 dropped words (d=0 to d=3) and permutations of words (p=T - permutation; p=F - no permutation).

<table>
<thead>
<tr>
<th>COND</th>
<th>SUCC</th>
<th>CM</th>
<th>PU</th>
<th>PM</th>
<th>IM</th>
<th>OL</th>
</tr>
</thead>
<tbody>
<tr>
<td>d=0, p=F</td>
<td>.78</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
<td>.78</td>
<td>.79</td>
</tr>
<tr>
<td>d=1, p=F</td>
<td>.89</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
<td>.89</td>
<td>.79</td>
</tr>
<tr>
<td>d=2, p=F</td>
<td>.78</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
<td>.78</td>
<td>.69</td>
</tr>
<tr>
<td>d=3, p=F</td>
<td>.82</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
<td>.82</td>
<td>.70</td>
</tr>
<tr>
<td>d=0, p=T</td>
<td>.70</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
<td>.70</td>
<td>.60</td>
</tr>
<tr>
<td>d=1, p=T</td>
<td>.74</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
<td>.74</td>
<td>.60</td>
</tr>
<tr>
<td>d=2, p=T</td>
<td>.78</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
<td>.78</td>
<td>.65</td>
</tr>
<tr>
<td>d=3, p=T</td>
<td>.85</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
<td>.83</td>
<td>.70</td>
</tr>
</tbody>
</table>

Results suggest that agents are well capable of dealing with language perturbations. If anything communicative success improves because the hearer can rearrange the words in such a way or imagine missing words so as to make the sentence fit his observation of the scene.

5 Discussion

The system presented in this paper is a fully working system able to interpret and produce natural language phrases with dynamic and static spatial relations. Such a system is useful for human-robot interaction about aspects of the environment. For instance, components of these phrases can be used in question-answer scenarios or in command-driven human-robot interfaces. Robots can understand the need to move to a certain location. Description of regions path, source and goal can be used to drive behavior and action planning systems. Part of our ongoing work is to test the system for command language with human subjects.

This paper reviewed the proposed system primarily with respect to perturbations in visual processing and language transmission. We believe that this is a fruitful way of analyzing the robustness of Natural Language systems, something that is often not done in the AI/Robotics community. Importantly, we found that robustness is primarily a function of integration of various cues from vision, reasoning, semantics and syntax. Only if each part of the system has some notion of dealing with perturbations can the system as a whole cope with various robustness issues.
References


Chapter 4

Visuospatial Abduction

This chapter presents a general method for hypothesising explanations of visuospatial phenomena based on abductive reasoning with space-time objects. We present the formal framework and its general implementation as a robust declarative method in Answer Set Programming (ASP). Furthermore, the visual abduction method is also developed into an online visual sensemaking framework directly within ASP. In essence, the resulting framework supports (optimised) hypothesis generation and joint abduction of high-level scene dynamics and corresponding object-level motion both for offline as well as real-time / online problem contexts. As one application of the method, we demonstrate and evaluated for semantic interpretation and decision support in autonomous driving in the context of safety-critical situations.

Included Publications:


4.1. Abducing Visual Explanations

Abductive reasoning denotes reasoning about possible explanations for a set of observations. In (human) visual sensemaking the process of generating high-level conceptual explanations for observed visual phenomena can be viewed as an abductive problem (Magnani, 2015; Moriarty, 1996), involving tight linkages between low-level subsymbolic visual processing on the one hand, and high-level object and event-based segmentation and inference involving concepts and relations on the other.

Within this thesis, we develop a general method for computing visual explanations, implemented based on Answer Set Programming (ASP) (Brewka, Eiter, and Truszczyński, 2011; Gebser et al., 2012), tightly integrating visual processing for object detection and tracking with declarative reasoning about events, objects, and spatial dynamics. The method is based on (computational) abductive reasoning as a mechanism for generating hypotheses explaining visuospatial scene dynamics, and builds on formal characterisations of logical abduction as it has evolved in KR research (J. R. Josephson and S. G. Josephson, 1994). In essence, visuospatial abduction, as developed in this thesis, denotes the process of reasoning from visual observations to explanations, consisting of spatio-temporal belief (states) and events, grounded in perceived low-level (object) motion (e.g., as illustrated in Example 5). In the following we describe the main elements of this method, namely the abductive process for hypotheses formation and the optimization step to find the best (most optimal) explanation.

▶ Hypotheses Formation At the core of the method for computing visuospatial explanations is an abductive process to generate hypotheses explaining visually observed scene dynamics, given as a sequence of visual observations $VO$, e.g., motion tracks of objects in the scene obtained from (deep learning based) visual processing. The task of
Example 5. ABDUCING OCCLUSION EVENTS

Consider the situation depicted in Figure 4.1, where a cyclist riding on the road gets occluded by a car turning right. This situation may be described in natural language as follows:

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.

Abduction aims at explaining the perceived spatio-temporal changes in terms of high-level object dynamics, i.e., finding a set of events that explain the sudden disappearance of the cyclist. For the example situation we can hypothesise that the cyclist got occluded by the car in-front, based on the relational spatio-temporal structure of the scene. This abducted knowledge in turn may be used in decision-making to reason about appropriate control actions, i.e., in this case slowing down to give space to the cyclist.

Visuospatial abduction is to find a set of logically consistent hypotheses \( \mathcal{H} \) consisting of high-level object interactions, such that \( \mathcal{H} \) is consistent with the background knowledge, i.e., domain characteristics defined in the backdrop of the theory of space and motion (\( \Sigma \)), and entails the perceived visual observations \( \mathcal{VO} \), e.g., the motion tracks of scene objects, geometrically represented as space-time histories of objects bounding boxes.

Formally speaking:

\[
\Sigma \land \mathcal{H} \models \mathcal{VO}
\]

Where the computed hypotheses \( \mathcal{H} \) are based on abducibles constituting primitive events (\( \Theta \)) and beliefs (\( \Phi \)), i.e., \( \mathcal{H} \equiv \mathcal{H}_{Events} \land \mathcal{H}_{Beliefs} \).

Optimizing Event Sequences and Motion Characteristics. For finding a preferred explanation, the generated hypotheses are ranked based on the involved events and the motion characteristics of the object tracks. For the purpose of the examples in this thesis, we optimise the hypotheses with respect to the following two objectives:

1. **High-level event sequences.** We assign costs to each event and to the length of the event. Costs for event length are specific to the event, e.g., missing detections as a result of an occlusion may be longer than missing detections coming from noisy observations.

2. **Low-level motion characteristics.** Assuming constant motion, we are minimising changes in the movement, i.e., in direction, velocity, and size. In this way, we are prioritising motion tracks with only little “unexpected” motion.
This optimisation is implemented in ASP using *weak constraints* (Gebser et al., 2012), i.e., constraints whose violation has a predefined cost. When solving an ASP program with weak constraints, a search for an answer set with a minimal cost of violated constraints is performed, resulting in an (optimal) minimal-cost answer set.

**Integrating Abduction and Optimization.** Based on the above characterisation, visual abduction is implemented as a three step approach, consisting of:

**Step 1.** low-level visual processing, e.g., for detecting and tracking objects and estimate motion etc., resulting in visual observations ($\mathcal{VO}$), e.g., people detections, motion tracks, scene motion, etc;

**Step 2.** abducing hypotheses ($\mathcal{H}$) on events and belief states, consistent with $\Sigma$, and explaining $\mathcal{VO}$; and

**Step 3.** hypotheses optimisation, using the built in optimisation functionality of ASP, where weak constraints are used to set preferences among hypothesised explanations to find preferred (optimal) solutions.

The final movement tracks are then generated based on resulting explanations together with the motion tracks in $\mathcal{VO}$ by predicting the motion of the object for each hypothesised event connecting scene elements, using linear interpolation. Thus, the abduced object movement represents the low-level instantiation of the abduced high-level event sequence.

The method is demonstrated and evaluated in the context of the Movie Dataset and the MOTChallenge (Section 2.3.3, D1. and D4.) for tracking and explaining scene dynamics. Results show that high-level abduction can be used to abduce semantics and help to reduce common tracking errors, e.g., identity switches, missing detections, etc.

**INCLUDED PUBLICATIONS:** The general concept of visuospatial abduction is introduced in the following publication:


*A copy of the above stated publication follows in the next section.*
6 - Visual Explanation by High-Level Abduction

Published in: 
AAAI Conference on Artificial Intelligence (AAAI 2018)

Citation: 

This conference paper was published at the AAAI conference on Artificial Intelligence (AAAI 2018). The paper introduces the concept of visual abduction and presents a general method for systematically generating visual explanations driven by ASP encompassing hypothesis formation, belief revision, and default reasoning with video data. A particular focus of the paper is on the integration of high-level abductive reasoning with low-level visual processing for object tracking, facilitating abductive hypotheses generation and optimisation based evaluation of these hypotheses, with respect to involved event sequences and motion characteristics.

The method is demonstrated using select examples from the Movie Dataset (Section 2.3.3, D1.), showing how the proposed system generates explanations in the context of visual phenomena, such as object persistence, occlusion, or attachment. Performance of the visual abduction is empirically evaluated using the community established benchmark dataset for Multi-Object Tracking (MOT) (Section 2.3.3, D4.), showing that abduced explanations of visuospatial phenomena can help to reduce common errors in object tracking, e.g., identity switches, missing detections, etc.
Visual Explanation by High-Level Abduction: On Answer-Set Programming Driven Reasoning about Moving Objects

Jakob Suchan¹, Mehul Bhatt¹,², Przemysław Wałęga³, and Carl Schultz⁴

Cognitive Vision – www.cognitive-vision.org
¹HCC Lab., University of Bremen, Germany, ²MPI Lab., Örebro University, Sweden
³University of Warsaw, Poland, and ⁴Aarhus University, Denmark

Abstract

We propose a hybrid architecture for systematically computing robust visual explanation(s) encompassing hypothesis formation, belief revision, and default reasoning with video data. The architecture consists of two tightly integrated synergistic components: (1) (functional) answer set programming based abductive reasoning with SPACE-TIME TRACKLETS as native entities; and (2) a visual processing pipeline for detection based object tracking and motion analysis.

We present the formal framework, its general implementation as a (declarative) method in answer set programming, and an example application and evaluation based on two diverse video datasets: the MOTChallenge benchmark developed by the vision community, and a recently developed Movie Dataset.

Introduction

A range of empirical research areas such as cognitive psychology and visual perception articulate human visual sense-making as an inherently abductive (reasoning) process (Moriarty 1996; Magnani 2015) involving tight linkages between low-level sub-symbolic processes on the one hand, and high-level object and event-based segmentation and inference involving concepts and relations on the other. In spite of the state of the art in artificial intelligence and computer vision, and most recent advances in neural visual processing, generalised explainable visual perception with conceptual categories in the context of dynamic visuo-spatial imagery remains an exceptionally challenging problem presenting many research opportunities at the interface of Logic, Language, and Computer Vision.

Explainable Visual Perception

We define explainable visual perception from a human-centred, and commonsense reasoning viewpoint. In this paper, it denotes the ability to declaratively:

\( \forall XP_1 \): hypothesise spatio-temporal belief (states) and events; events may be both primitive or temporally-ordered aggregates; from a more foundational viewpoint, what is alluded to here is a robust mechanism for counterfactual reasoning.

\( \forall XP_2 \): revise spatio-temporal beliefs, e.g., by non-monotonically updating conflicting knowledge, to fix inherently incompatible configurations in space-time-defying geometric constraints and commonsense laws of naïve physics, e.g., pertaining to physical (un)realisability, spatio-temporal continuity.

\( \forall XP_3 \): make default assumptions, e.g., about spatio-temporal property persistence concerning occupancy or position of objects; identity of tracked objects in space-time. Explanatory reasoning in general is one of the hallmarks of general human reasoning ability; robust explainable visual perception particularly stands out as a foundational functional capability within the human visuo-spatial perception faculty. In this respect, the following considerations — establishing the scope of this paper— are important wrt. \( \forall XP_{1-3} \):

- our notion of explainability is driven by the ability to support commonsense, semantic question-answering over dynamic visuo-spatial imagery within a declarative KR setting;
- the features alluded to in \( \forall XP_{1-3} \) are not exhaustive; we focus on those aspects that we deem most essential for the particular case of movement tracking.

A Hybrid Architecture for Visual Explanation

This paper is driven by the development of a visual explanation component within a large-scale computational vision & perception system targeted at a range of cognitive interaction technologies and autonomous systems where dynamic visuo-spatial imagery is inherent.

The key contribution is a hybrid visual explanation method based on the integration of high-level abductive reasoning within Answer Set Programming (ASP) ((Brewka, Eiter, and Truszczyński 2011)) on the one hand, and low-level visual processing for object tracking on the other. The core focus of the paper is on the theory, implementation, and applied evaluation of the visual explanation method. We particularly emphasise the closely-knit nature of two key sub-components representing abductive explanation (\( \Sigma_{abd} \)) and low-level motion tracking (\( \Sigma_{trk} \)) modules respectively:

\( \Sigma_{abd} \): ASP-based abductive reasoning with abstract visuo-spatial concepts — such as \textsc{objects}, \textsc{events}, \textsc{space-time tracklets} — as native objects within ASP.

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
Ω_{trk}. Low-level visual processing pipeline for motion tracking, consisting of detection-based object tracking and optical-flow based (scene-level) movement tracking.

The abductive component Ω_{abd} is suited for a wide-range of dynamic visuo-spatial imagery; however, we only focus on video in this paper. As an application, we focus on scene interpretation from video with two datasets: a Movie Dataset (Suchan and Bhatt 2016a) and the MOT16 Dataset, a benchmark dataset released as part of The Multiple Object Tracking Challenge (Milan et al. 2016).

**Visual Explanation: A Hybrid Architecture**

We present a general theory for explaining visuo-spatial observations by integrating low-level visual processing and high-level abductive reasoning (Fig. 1). As such, we consider visual abduction as reasoning from visual observations to explanations consisting of high-level events grounded in low-level motion tracks. The resulting set of hypotheses is optimised based on the abduced events and the corresponding object movement.

**Ontology: Space, Time, Objects, Events**

The framework for abducting visual explanations is based on visuo-spatial domain objects representing the visual elements in the scene. The domain objects are associated with spatio-temporal objects describing motion tracks obtained from Ω_{trk}, which form the basis for qualitative spatio-temporal abstractions facilitating high-level reasoning about visuo-spatial dynamics.

The **Qualitative Spatio-Temporal Domain (QS)** is characterised by the basic spatial and temporal entities (E) that can be used as abstract representations of domain-objects and the relational spatio-temporal structure (R) that characterises the qualitative spatio-temporal relationships amongst the supported entities in (E). For this paper, we restrict the basic spatial entities to:

- points are a pair of reals x, y,
- axis-aligned rectangles are a point p and its width and height w, h,

and the temporal entities to:

- time-points are a real t

Visuo-spatial domain objects \( \mathcal{O} = \{o_1, o_2, ..., o_n\} \) are described as spatio-temporal objects by a set of spatial entities, i.e., points, and axis-aligned rectangles, in time. Towards this, \( \mathcal{MT} \) contains all object tracks obtained form \( \Omega_{trk} \). The track of a single object \( o_i \) is represented by \( \mathcal{MT}_{o_i} = (t_1, ..., t_k) \), where \( t_s \) and \( t_e \) denote the start and end frame of the object, and \( t_s \) to \( t_e \) denotes a spatial primitive representing the object \( o_i \) at the time points \( t_s \) to \( t_e \), e.g., the axis aligned bounding box of the object.

For reasoning about visuo-spatial phenomena of object tracks, spatio-temporal relationships (R) between the basic entities in \( \mathcal{E} \) may be characterised with respect to arbitrary spatial and spatio-temporal domains such as mereotopology, orientation, distance, size, motion. From the viewpoint of the examples of this paper, it suffices to focus on the language of the mereotopological system of the Region Connection Calculus (RCC8) (Randell, Cui, and Cohn 1992) consisting of the following jointly exhaustive and pair-wise disjoint relations: disconnected (dc), externally connected (ec), partially overlapping (po), equal (eq), (non-) tangential proper part ((n)tp), and their inverse ((n)tpp).

Abducable events (Θ) and beliefs (Φ) are defined by their (spatio-temporal) preconditions and observer effects, i.e., for each event \( \theta \in \Theta \) we define which properties of the scene have to be true for the event to be possible, and what the (visible) effects of the event are. In the case of visual abduction, properties of the scene are determined by the visually observed object tracks and represent qualitative relations between tracks, i.e., spatial relation \( r \in \mathcal{R} \) holding between basic spatial entities \( e \) of a motion track. Complex events are defined by combining multiple events and beliefs, e.g., an event of an object \( o_i \) passing behind another object \( o_j \) can be defined based on the events of \( o_i \) being occluded by \( o_j \) and \( o_i \) or \( o_j \) changing sides.

**Abducting Visual Explanations**

We implement the theory for visual explanations combining visual processing for object detection and tracking, and estimating movements in the scene, with ASP based reasoning about events, objects, and spatial-dynamics (Fig. 1). The main components of the overall tightly-integrated system comprising of low-level motion tracking with high-level explanation is as follows:
I. Visuo-Spatial Observations (VO) – low-level visual processing consisting of detection based tracking of object and people movements.

II. Hypotheses (H) – abducting hypotheses including belief states, events, and default assumptions given a set of visuo-spatial observations (VO).

III. Hypotheses to Explanations — as encompassed in \( VXP \) — are generated by evaluating abduced hypothesis (H) based on high-level optimisation of event sequences and low-level cost minimisation of corresponding motion tracks.

I. Visuo-Spatial Observations (VO) Visual observations are based on observations obtained from visuo-spatial imagery, e.g., video, RGB-D. For the examples in this paper, we focus on detection and tracking of people and objects for estimating motion trajectories of semantic entities in the scene. However, the presented approach is also capable of incorporating other kinds of motion, e.g., optical flow based low-level movement analysis using long term observations (Ochs, Malik, and Brox 2014), or dense motion-tracks (Gaidon, Harchaoui, and Schmid 2014) for estimating pixel level motion, corresponding to camera movement, or fine grained object motion, etc. This may be used to abduce fine grained interactions and the interplay of different movements, e.g. people movement in the presence of camera movement, by combining motion trajectories of semantic entities with pixel movements.

Movement of people and objects is estimated following the tracking by detection paradigm, which is based on object detections for each frame and association of the detections across frames. Object detections can in principle be obtained using any state of the art (deep learning based) detector (e.g., faster RCNN (Ren et al. 2015), YOLO (Redmon et al. 2016)), or deformable part models (DPM) (Felzenszwalb et al. 2010). For the examples in this paper we are using faster RCNN in the movie examples and DPM detections for the MOT dataset (which come as part of the dataset). For association of detections we apply the well established approach of combining min cost assignment and kalman filters for finding optimal tracklets, where the cost for assigning a detection to a track is calculated by the distance between the prediction for a track and the detection.

- Prediction for each track Kalman filters are used to predict the next position of the track and the costs for each detection is calculated based on the distance between the prediction and the detection.
- Assignment detections are assigned to a track using min cost assignment which calculates the best assignment of detection to tracks based on the costs calculated in the prediction step. If no assignment is possible for a detection a new track is started.

The resulting object tracks \( MT \) form the basis for abducing explanations on movement events occurring in the input data.

II. Hypotheses (H) Explanations for visual observations are abduced based on a sequence of visual observations obtained from the video data. For abducing visual explanations from VO, given,

- set \( VO \) consisting of visuo-spatial observations obtained from \( \Sigma_{\text{track}} \),
- domain independent theory of space and time (\( \Sigma_{\text{space}} \)) based on the spatio-temporal ontology (QS)
- observable events (\( \Sigma_{\text{events}} \))
- domain dependent background knowledge, describing properties of the domain (\( \Sigma_{\text{domain}} \))

the task of visual abduction is to find a set of logically consistent hypotheses \( H \) consisting of high-level events and beliefs grounded in low-level motion tracks, such that:

\[
\Sigma_{\text{space}} \wedge \Sigma_{\text{events}} \wedge \Sigma_{\text{domain}} \wedge H \models VO
\]

The computed hypotheses (H) are based on abducibles constituting primitive events and beliefs: \( H \subseteq H_{\text{Events}} \wedge H_{\text{Belief}} \); these hypotheses in turn are directly usable for inducing motion tracks:

\[
MT_{VXP} \leftarrow H_{\text{event}} \wedge H_{\text{belief}} \wedge MT
\]

The resulting motion tracks \( MT_{VXP} \) represent the low-level instantiation of the abduced high-level event sequence.

III. Hypotheses to Explanations Hypotheses for visual observations (VO) may be ranked based on the abduced event sequences and cost minimisation of corresponding motion trajectories, i.e., the costs for connecting motion tracks in the hypothesised movements, e.g. considering changes in velocity, size, and length of missing detections. As such, hypothesised explanations are ranked using the built in optimisation functionality of ASP\(^1\). In particular, we use minimisation by assigning preferences to the abducibles events and beliefs and optimise towards minimising the costs of events and beliefs in the answer. E.g., by minimising the duration of missing detections for a particular object, or minimising assigning the property noise to a track to explain its observation.

- High-level event sequences the cost for high-level events is estimated by assigning a cost for each event. Additionally, for events having a duration there is also a cost assigned to the length of the event, e.g., to abduce that a track is noise is more likely, when it is a very short track. These costs are weighted based on the abduced event that caused the missing detections, e.g., missing detections caused by an occlusion are more likely to be longer (and therefore have a lower cost), than missing detections caused by the detector.

\(^1\)For optimisation we use ASP with the so-called weak constraints (Gebser et al. 2012), i.e., constraints whose violation has a predefined cost. When solving an ASP program with weak constraints, a search for an answer set with a minimal cost of violated constraints is performed. Each such minimal-cost answer set is called optimal. The mechanism involving weak constraints enables us to set preferences among hypothesised explanations and search for the ones that are most preferred (optimal). Importantly, the approach enables us to exhaustively search for all optimal explanations. As a result, we can subsequently use other (more fine-graded) evaluation techniques to choose the most preferred explanations.
EVENTS

<table>
<thead>
<tr>
<th>EVENTS</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>enters(Border, Trk, T)</td>
<td>The object corresponding to track Trk enters the scene at time point T.</td>
</tr>
<tr>
<td>exits(Border, Trk, T)</td>
<td>The object corresponding to track Trk exits the scene at time point T.</td>
</tr>
<tr>
<td>occludes(Trk1, Trk2, T1, T2)</td>
<td>The object corresponding to track Trk1 and track Trk2 is occluded by the object corresponding to track Trk3 between time points T1 and T2.</td>
</tr>
<tr>
<td>missing_det(Trk1, Trk2, T1, T2)</td>
<td>Missing detections for the object corresponding to the tracks Trk1 and Trk2 between time points T1 and T2.</td>
</tr>
</tbody>
</table>

COMPLEX EVENTS

<table>
<thead>
<tr>
<th>EVENTS</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>passing_behind(O1, O2, T1, T2)</td>
<td>Object O1 is passing behind object O2 between time points T1 and T2.</td>
</tr>
<tr>
<td>moving_together(O1, O2, T1, T2)</td>
<td>Objects O1 and O2 are moving together between time points T1 and T2.</td>
</tr>
</tbody>
</table>

BELIEFS

<table>
<thead>
<tr>
<th>EVENTS</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>same_object(Trk1, Trk2)</td>
<td>The tracks Trk1 and Trk2 belong to the same object.</td>
</tr>
<tr>
<td>belongs_to(Trk1, Trk2)</td>
<td>The object corresponding to track Trk1 is a part of the object corresponding to track Trk2.</td>
</tr>
</tbody>
</table>


Table 1: Abducibles: Events and Beliefs for Explaining Observed Object Tracks.

> **Low-level motion characteristics** the cost of the motion tracks $MT$ is estimated based on the characteristics of the abduced movement. Towards this we consider changes in *velocity*, for each abduced event that connects two object tracks. For the examples in this paper we use a constant velocity model to minimize changes in velocity of an abduced object track.

The best explanation is selected by *minimising* the costs of the hypothesised answer set based on the motion and the high-level event sequence. The final movement tracks for the optimal explanation $MT_{\mathcal{A},p}$ are then generated by predicting the motion of the object for each hypothesised event associating two tracks, using linear interpolation.

**Visuo-Spatial Phenomena**

The framework may be used for abducing explanations by modelling visuo-spatial phenomena including but not limited to:

- **Object Persistence** objects can not appear and disappear without a cause, e.g. getting occluded, leaving the field of view of the camera, etc.
- **Occlusion** objects may disappear or re-appear as a result of occlusion between two non-opaque objects.
- **Linkage** objects linked to each other, such that movement of one object influences movement of the other object, e.g. a face belonging to a person.
- **Sensor Noise** observations that are based on faulty data, e.g. missing information, miss-detections, etc.

**Event Semantics as Spatial Constraints** For explaining perceived visuo-spatial dynamics of objects in the scene, we define the basic events listed in Table 1 to assure spatio-temporal consistency, e.g. object persistence, or occlusion. The focus is on explaining appearance and disappearance of objects in the scene.2

> **Entering and Leaving** Objects can only enter or exit the scene by leaving the screen at one of its borders. For these events to happen the object has to be overlapping with the border of the screen while appearing or disappearing.

<table>
<thead>
<tr>
<th>EVENTS</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>enters:</td>
<td>topology(po, TRbox, left_border) :- enters(from_left, TR, T), track(TR, TRbox, T).</td>
</tr>
<tr>
<td></td>
<td>topology(po, TRbox, right_border) :- enters(from_right, TR, T), track(TR, TRbox, T).</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EVENTS</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>exits:</td>
<td>topology(po, TRbox, left_border) :-</td>
</tr>
<tr>
<td></td>
<td>exits(to_left, TR, T), track(TR, TRbox, T).</td>
</tr>
<tr>
<td></td>
<td>topology(po, TRbox, right_border) :-</td>
</tr>
<tr>
<td></td>
<td>exits(to_right, TR, T), track(TR, TRbox, T).</td>
</tr>
</tbody>
</table>

> **Missing Detections and Occlusion** Appearance and disappearance of tracks in the middle of the screen can be either caused by a missing detection or by an occlusion from some other object. The event that an object gets occluded by some other object may be possible, when the object disappears while overlapping with the other object.

<table>
<thead>
<tr>
<th>EVENTS</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>occludes:</td>
<td>topology(po, TRbox1, TRbox2) :-</td>
</tr>
<tr>
<td></td>
<td>occludes(TR1, TR2a, TR2b, T1, T2), track(TR1, TRbox1, T1), track(TR2a, TRbox2, T2).</td>
</tr>
<tr>
<td></td>
<td>occludes(TR1, TR2a, TR2b, T1, T2), track(TR1, TRbox1, T1), track(TR2a, TRbox2, T2).</td>
</tr>
</tbody>
</table>

**Generating Hypotheses on Events** We generate hypotheses explaining the observation of a track starting and ending based on the defined events, such that the spatial constraints defined above are satisfied.

<table>
<thead>
<tr>
<th>EVENTS</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>starts:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>noise(TR); enters(from_left, TR, T); enters(from_right, TR, T);</td>
</tr>
<tr>
<td></td>
<td>occludes(TR1, TR2a, TR2b, T1, T2), type(TR1, Type), Type=border, starts(TR1, T11),</td>
</tr>
<tr>
<td></td>
<td>ends(TR1, T12), ends(TR1, T11), T11&lt;=T, T&lt;=T12</td>
</tr>
<tr>
<td></td>
<td>:- starts(TR2, T).</td>
</tr>
</tbody>
</table>

---

2The semantics of the underlying spatial and temporal relations with $(QS)$ is founded on the geometric and spatial reasoning capability provided by the ASPMT(QS) spatial reasoning system (Walęga, Bhatt, and Schultz 2015); the system, implemented within ASPMT (Lee and Meng 2013), is directly available to be used as a black-box within our visual explanation framework.
Beliefs as (Spatial) Constraints  Beliefs about objects in the scene are stated as constraints in ASP.

- **Part-Whole Relations** E.g. the fact that every face belongs to exactly one person is stated as follows.

  ```prolog
  :- belongs_to(Face1,Person),
  belongs_to(Face2,Person),
  Face1=Face2.
  ```

Further we define that the face of a person has to stay together with the person it belongs to, using spatial constraints, i.e. the face track is a non-tangential proper part of the person track.

```prolog
topology(ftp,Box1,Box2) :-
  belongs_to(Face, Person),
  track(Person, Box1, Box2),
  track(Face, Box1, Box2).
```

Generating Hypotheses on Beliefs  Hypotheses on faces belonging to persons are generated by stating that for each detected face, there has to be a corresponding person, such that the spatial constraint is satisfied.

```prolog
{ belongs_to(Face, Person) :- type(Person, person) } 1
:- type(Face, face).
```

Costs of Hypotheses using Optimization  Costs for abduction visual explanations are minimized using ASP based optimization, e.g., the cost for missing detections are based on their length.

```prolog
#minimize { (T2_start-T1_end)*ALPHA, TR1, TR2 :
  missing_det(TR1,TR2,T1_end,T1_start),
  weight(missing_det, ALPHA)}.
```

Further, the characteristics of the underlying motion is taken into account, assuming constant velocity, by taking differences in velocity between the two tracks and the interpolated segment in between.

```prolog
#minimize { ((X_vel_prev-2*X_vel_during*X_vel_next)**2 +
  (Y_vel_prev-2*Y_vel_during*Y_vel_next)**2)*ALPHA,
  TR1, TR2 :
  missing_det(TR1,TR2,T1,T2),
  track(TR1, Box1, T1), box(Box1, X_e, Y_e, ...),
  [... long ...]
  X_vel_during = (X_e - X_s) / (T2-T1),
  Y_vel_during = (Y_e - Y_s) / (T2-T1),
  [... long ...]
  weight(missing_det_vel, ALPHA)}.
```

**Application and Evaluation:**

**Scene Interpretation with Moving Objects**

We demonstrate the proposed theory of visual abdiction by applying it in the context of scene interpretation focussing on generating visual explanations on perceived motion. In particular, the emphasis is on spatio-temporal consistency of abduced explanations with respect to the underlying motion tracks.

**Movie Dataset** (Suchan and Bhatt 2016a; 2016b). We use the video part of the Movie Dataset consisting of 16 select scenes from 12 films, with each scene ranging between 0 : 38 minute to max. of 9 : 44 minutes in duration. Most of the scenes involve multiple moving objects, and moving camera(s). Object detection with the movie dataset is performed using faster RCNN (Ren et al. 2015) with the pretrained VGG16 model for detection of people and objects in the scene.

**Visual Explanation of Object Movement** As an example consider the scene from the movie The Grand Budapest Hotel (2014) by Wes Anderson (Figure 2). Here we abduce the movement of the two main characters walking down the hallway of the hotel. The set of visual observations consist of 11 tracks for the detected people in the scene. The abduced events explain occurring missing detections, occlusion and re-appearance, as well as entering, and leaving the scene.

```prolog
exits(to_right, trk10, 1511) enters(from_left, trk9, 1500)
occludes(trk4, trk2, trk6, 1490, 1495)
occludes(trk3, trk0, trk7, 1490, 1496)
...
nnoise(trk4) nnoise(trk8) ending(trk1, 1513)
```
Similarly we can abduce complex events based on movement events and beliefs, e.g., based on the following sequence of movement events from the movie “The Bad Sleep Well” (1960) by Akira Kurosawa (depicted in Figure 3), we can abduce the occurrence of the complex event passing behind between two objects in the scene.

Similarly we can abduce complex events based on movement events and beliefs, e.g., based on the following sequence of movement events from the movie “The Bad Sleep Well” (1960) by Akira Kurosawa (depicted in Figure 3), we can abduce the occurrence of the complex event passing behind between two objects in the scene.

MOT16 Benchmark Video Dataset We use the MOT16 (Milan et al. 2016) dataset consisting of highly accurate and consistent annotation protocols. MOT16 is a benchmark dataset released as part of The Multiple Object Tracking Challenge (MOTChallenge). It consists of 14 complex video sequences in highly unconstrained environments filmed with both static and moving cameras. We use the detections (provided by the dataset) based on deformable part models (DPM); these are noisy and include numerous miss detections, i.e., false positives and false negatives. We focus on abducting people motion and on generating concise explanations for the perceived movements, i.e., under consideration of occlusion and appearance / disappearance of characters as per the abducible events in Table 1. As a result of the noisy detections and the complexity of the movements in the dataset the obtained motion tracks include a high amount of errors, e.g., identity switches, missing detections, etc. (Figure 4). For the sample scene we abduced the following events:

```
... missing_det(trk1, trk2, 55, 63) occlusion(trk3, trk4, 62, 59) missing_det(trk4, trk5, 62, 63)
... passing_behind(obj1, obj2, 56, 62)
...
```

Evaluating Visual Explanations

We evaluate the generated visual explanations based on their ability to generate low-level object tracks. Towards this we compare the accuracy and precision of the movement tracks the hypothesised event sequences are grounded in.

Multi-Object Tracking For evaluating the precision and accuracy of the abduced object tracks we follow the ClearMOT evaluation schema for evaluating multi-object tracking performance as described in (Bernardin and Stiefelhagen 2008).

- **MOTA** describes the accuracy of the tracking, taking into account the number of missed objects / false negatives (FN), the number of false positives (FP), and the number of miss-matches (MM).
- **MOTP** describes the precision of the tracking based on the distance of the hypothesised track to the ground truth of the object it is associated to.

These metrics are used to assess how well the generated visual explanations describe the low-level motion in the scene.

Results & Discussion

We present results of the presented approach for abducting visual explanations (VXML) on improving multi-object tracking performance using selected scenes from the Movie Dataset and the MOT 2016 Dataset. Overall the results show that using our proposed method can increase accuracy (MOTA) of the tracking. However, the precision (MOTP) of the tracking is dropping a little, which is a result of the interpolation, which is not as precise as the detections.

**Movie Dataset** The scenes in the Movie Dataset contain relatively controlled scenes with few targets. Results on these scenes show that the presented approach can abduce correct event sequences and is capable of correcting many of the errors normally occurring in multi-object tracking tasks, e.g., fragmented object tracks, id-switches, etc. i.e., the object tracks obtained from the high-level event sequences improve the accuracy (MOTA) of the tracking (see Table 2).

**MOT Dataset** The results for the visual tracking on the Venice-2 file from the MOT2016 dataset (see Table 2) show that our approach is capable of dealing with complex data in challenging settings. For comparability we use the DPM
Table 2: Evaluation of Tracking Performance: false positives (FP), misses (M), miss-matches (MM), non-recoverable miss-matches (non-r. MM), recoverable miss-matches (r. MM), track precision (TP), track recall (TR)

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Tracking</th>
<th>MOTA</th>
<th>MOTP</th>
<th>FP</th>
<th>M</th>
<th>MM</th>
<th>non-r. MM</th>
<th>r. MM</th>
<th>TP</th>
<th>TR</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Bad Sleep Well</td>
<td>without VXP</td>
<td>58.5</td>
<td>80.8</td>
<td>1</td>
<td>86</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>0.875</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>with VXP</td>
<td>100.0</td>
<td>69.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>The Drive</td>
<td>without VXP</td>
<td>59.8</td>
<td>76.7</td>
<td>0</td>
<td>345</td>
<td>18</td>
<td>0</td>
<td>18</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>with VXP</td>
<td>79.7</td>
<td>76.6</td>
<td>0</td>
<td>182</td>
<td>18</td>
<td>0</td>
<td>18</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>MOT2016 - Venice-2</td>
<td>without VXP</td>
<td>6.4</td>
<td>69.9</td>
<td>47</td>
<td>2537</td>
<td>153</td>
<td>4</td>
<td>150</td>
<td>0.987</td>
<td>0.241</td>
</tr>
<tr>
<td></td>
<td>with VXP</td>
<td>8.1</td>
<td>65.2</td>
<td>216</td>
<td>26535</td>
<td>86</td>
<td>27</td>
<td>77</td>
<td>0.946</td>
<td>0.486</td>
</tr>
</tbody>
</table>

Based on basic tracking by detection, we suppose that ASP-based visual explanations can also be used to improve multi-object tracking using more elaborate tracking approaches, e.g., based on continuous energy minimization (Milan, Schindler, and Roth 2016) or minimum cost multi-cuts (Tang et al. 2017).

Related Work

Answer Set Programming (ASP) has become a widely used tool for abductive reasoning and non-monotonic reasoning in general. The paper presented in this paper aims at bridging the gap between high-level formalisms for logical abduction and low level visual processing, by tightly integrating qualitative abstractions of space and time with the underlying numerical representations of spatial change. The significance of abducing high-level explanations in a range of contexts has been well established in AI and KR, e.g., in planning and process recognition (Kautz and Allen 1986; Kautz 1991), vision and abduction (Shanahan 2005), probabilistic abduction (Blythe et al. 2011) etc. Within KR, reasoning about spatio-temporal dynamics on the basis of an integrated theory of space, time, objects, and position (Galtung 2000) or defined continuous change using 4-dimensional regions in space-time has also received significant theoretical interest (Muller 1998; Hazarika and Cohn 2002). Dubba et al. (2015) uses abductive reasoning for improving learning of events in an inductive-abductive loop, using inductive logic programming (ILP). The role of visual commonsense in general, and answer set programming in particular, has been used in conjunction with computer vision to formalise general rules for image interpretation in the recent works of Aditya et al. (2015). From the viewpoint of computer vision research there has been an interest to synergise with cognitively motivated methods (Aloimonos and Fermüller 2015); in particular the research on semantic interpretation of visual imagery is relevant to this paper, e.g., for combining information from video analysis with textual information for understanding events and answering queries about video data (Tu et al. 2014), and perceptual grounding and inference (Yu et al. 2015).

Summary and Outlook

The paper presents a robust, declarative, and generally usable hybrid architecture for computing visual explanations with video data. With a focus on abductive reasoning in the context of motion tracking, the architecture has been formalised, fully implemented, evaluated with two diverse datasets: firstly, the benchmark MOTChallenge (evaluation focus), and secondly a Movie Dataset (demonstration focus). The overall agenda of the work in this paper is driven by a tighter integration of methods in KR and Computer Vision on the one hand, and the twin concepts of “deep semantics” & “explainability” on the other. $VXP$ is rooted in state of the art methods in knowledge representation and reasoning (i.e., answer set programming), and computer vision (detection based object tracking, optical flows, RCNN). The overall system is designed to be a part of a larger perception module within autonomous systems, and cognitive interaction systems. The scope of $VXP$ may be further expanded, e.g., for visuo-spatial learning (with inductive logic programming), ontological reasoning (with description logics), are achievable depending on the scope and complexity of the low-level visual signal processing pipeline, and chosen high-level commonsense knowledge representation and reasoning method(s) at hand.

Acknowledgements

We acknowledge funding as part of the German Research Foundation (DFG) CRC 1320 “EASE – Everyday Activity Science and Engineering” (http://www.ease-crc.org/) Project P3: “Spatial Reasoning in Everyday Activity”. We also acknowledge the Polish National Science Centre project 2016/23/N/HS1/02168.
References


Bhatt, M. 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.


4.2. Integrated Abduction and Object-Level Motion

The visuospatial abduction method presented in Section 4.1 is integrated within a visual perception loop to facilitate online visual sensemaking applicable in artificially intelligent systems concerned with real-time execution and involving commonsense semantic question-answering and belief maintenance with dynamic visuospatial imagery. We integrated ASP based abduction with motion tracking, following a tracking by detection approach, where tracking of object motion is solved by finding the best assignment for detections to object tracks based on visuospatial matching criteria. The developed method jointly solves the problem of assigning detections to tracks and explaining overall scene dynamics (e.g. appearance, disappearance) in terms of high-level events (illustrated in Example 6). Instead of solving the abductive problem for the whole scene at once (as it is done in Section 4.1), the problem is defined incrementally and solved at each time point individually, continuously optimising the generated hypotheses based on motion characteristics and involved events. This procedure enables online execution (with real-time performance on real-world examples from the community established KITTI and MOTChallenge benchmark datasets), and thus also supports processing of long scenes (which is not possible with the foundational offline method of Section 4.1).

► Joint Abduction and Motion Tracking. At the core of the presented online abduction method is the abductive process as formulated in Section 4.1, which is used for generating hypotheses on motion events, together with the corresponding motion tracks of involved scene elements. We build on this abductive step by integrating it into an iterative method for online visuospatial abduction, performing Steps 1–3 at each time point \( t \) (A detailed description of these steps together with the corresponding algorithm can be found in the included publication [7, Pg. 113]):

**Step 1.** Formulating the ASP problem specification \(<\mathcal{V}_t, \mathcal{P}_t, \mathcal{M}L_t>\) consisting of:

- **visual observations** \( \mathcal{V}_t \) consisting of detections of scene elements at time point \( t \),
- **predictions** \( \mathcal{P}_t \) of the position and size of each element at time point \( t \), based on movement of previously tracked scene elements represented by their motion tracks \( \mathcal{M}T \), and
- the **matching likelihood** \( \mathcal{M}L_t \) between the observations and motion tracks, calculated based on \( \mathcal{V}_t \) and \( \mathcal{P}_t \) (for the purpose of the examples in this thesis we are using the Intersection over Union (IoU)).

**Step 2.** Associating **visual observations** \( \mathcal{V}_t \) to **motion tracks** \( \mathcal{M}T \), by jointly abducting matchings between the elements in \( \mathcal{V}_t \) and \( \mathcal{M}T \), together with the high-level events explaining these associations.

- finding an association \( \mathcal{H}^{\text{assign}}_t \) of observed scene elements in \( \mathcal{V}_t \) to the motion tracks in \( \mathcal{M}T \) given by their predictions in \( \mathcal{P}_t \),
Example 6. Joint Abduction of Scene Dynamics and Object Motion

Looking once again at the example in Figure 4.1, visual abduction implemented in ASP facilitates reasoning about and explaining these kinds of occlusion events. Towards this we define actions that control how observations are assigned to tracks, i.e., for the purpose of this example we only consider the case that an observation is assigned to a track, and the case where no observation can be assigned to the track and the track is halted.

\[
\begin{align*}
\text{assign}(Trk, Det): & \quad \text{det}(Det, _, _) \quad \text{halt}(Trk); \quad \ldots \\
\text{trk}(Trk, _).
\end{align*}
\]

Further, we define a choice rule stating that a track \(Trk\) can get halted if the track hides behind another track \(Trk_2\).

\[
\begin{align*}
\text{occurs_at}(\text{hides_behind}(Trk, Trk_2), \text{curr_time}): & \quad \text{trk}(Trk_2, _); \quad \ldots \\
\text{halt}(Trk).
\end{align*}
\]

Joint abduction of scene dynamics and object motion then results in associations of detections to motion tracks and corresponding events explaining these associations:

\[
\begin{align*}
\text{halt}(trk_{13}) & \quad \text{assign}(trk_{15}, det_0) \quad \text{assign}(trk_{12}, det_1) \quad \text{assign}(trk_{8}, det_2) \\
\text{assign}(trk_{3}, det_3) & \quad \text{assign}(trk_{7}, det_4) \\
\text{occurs_at}(\text{hides_behind}(trk_{13}, trk_{12}), 235)
\end{align*}
\]

Events and corresponding event dynamics may be defined in standard KR based languages for reasoning about action and change. The conditions for occurrence of the event \(\text{hides_behind}\) for instance, can be defined using the axioms of event calculus as demonstrated in [8, Pg. 149].

- generating corresponding high-level explanations \((\mathcal{H}^{events}_t)\), such that hypotheses \(\mathcal{H}^{assign}_t \land \mathcal{H}^{events}_t\) are consistent with the spatio-temporal background knowledge in \(\Sigma\) and the previously abduced event sequence \(\mathcal{H}^{events}_t\), and entails the perceived scene given by the problem specification \(<\mathcal{VO}_t, \mathcal{P}_t, \mathcal{ML}_t>\).

This process is formalised as an abductive problem as follows:

\[
\Sigma \land \mathcal{H}^{events}_t \land [\mathcal{H}^{assign}_t \land \mathcal{H}^{events}_t] \models \mathcal{VO}_t \land \mathcal{P}_t \land \mathcal{ML}_t
\]

Step 3. Finding event hypotheses and corresponding associations \((\mathcal{H}^{assign}_t \land \mathcal{H}^{events}_t)\) best explaining the visual observations \(\mathcal{VO}\) is based on optimisation within ASP as de-
Example 7. Anticipating Driving Dynamics

Consider the situation when a car changes to our lane (Figure 4.2), where there is a car on the lane next to ours and this lane is blocked, e.g., by a slower vehicle.

Car $(c_1)$ is on the right lane; during this time, there is another car $(c_2)$ on the right lane in-front of $c_1$, which is slower than $c_1$.

In this situation, we have to take into account that the car might change to our lane, i.e., we may predict:

$c_1$ moves left, to change to the left lane, and overtake $c_2$.

Based on such a prediction we can then decide on appropriate control decisions to ensure safety, e.g., changing to another lane or reducing speed.

scribed in Section 4.1. In principle, optimisation may be driven by diverse criteria involving visuospatial dynamics, however, for the purpose of the examples in this thesis we are maximizing the matching likelihood ($ML$) and minimizing event costs. The resulting hypotheses, accumulated over the entire sequence, serve as explanations of the perceived scene dynamics, and is given by the event sequences in $H_{events}$ and the motion tracks $MT$, generated based on the associations at each time point ($H_{assign}$) and consistent with the abduced events.

These abduced explanations, together with the imposed constraints and event likelihoods defined by the spatio-temporal background knowledge in $\Sigma$, directly guide the motion tracking as association of observations to motion tracks is only possible when there is a corresponding explanation in $H_{events}$. Furthermore, the high-level events may be directly integrated with formalisms for reasoning about action and change, e.g., in the event calculus (Kowalski and Sergot, 1989), facilitating rich reasoning about event dynamics at runtime (also influencing low-level motion tracking).

Real-Time Visual Sensemaking in Autonomous Driving. The presented approach is applied for real-time visual sensemaking and decision support in the con-
text of safety-critical situations in autonomous driving, focusing on human-centred aspects, such as interpreting and anticipating movements of other traffic participants, (non-verbal) interactions with other pedestrians and cyclists, ethical and legal considerations involving standardisation and reporting, etc. A developmental focus is on declarative sensemaking with commonsense semantics of visuospatial phenomena, e.g., object persistence, occlusion, attachment, etc., demonstrated in the context of reasoning about occluded entities in typical driving situations. For instance, consider the situation in the introductory Example 5, here conceptual knowledge about occlusion of objects is used to maintain beliefs about the existence of the occluded entity and anticipate its position and reappearance. Such reasoning is essential in the context of anticipatory driving and necessary to avoid dangerous situations in cluttered and confusing driving situations. Another example involves reasoning about intention (Example 7), where it is essential to anticipate the lane change of the car in front, in order to maintain an appropriate safety distance. Online visual abduction is demonstrated in the above context and evaluated with community established benchmark datasets on autonomous driving and multi-target tracking, i.e., KITTI and the MOTChallenge (Section 2.3.3, D3, and D4). In this context empirical evaluation has shown that abduction can improve overall tracking performance by anticipating and interpolating motion of occluded objects.

INCLUDED PUBLICATIONS: Online visuospatial abduction as described in this section was published in the following two publications:

- A journal publication in the Journal of Artificial Intelligence (AIJ) (Suchan, Bhatt, and Varadarajan, 2021, 7), presenting the general method for online visuospatial abduction implemented in ASP and demonstrating visual sensemaking for safety-critical situations in autonomous driving; and

- A conference contribution published at the International Joint Conference on Artificial Intelligence (IJCAI 2019) (Suchan, Bhatt, and Varadarajan, 2019, 8), which was recognised by the distinguished paper award honorable mention. The paper introduces visuospatial abduction, integrating ASP based abduction and object-level motion for online, real-time visual sensemaking in autonomous driving.

Copies of the above stated publications in the given order follow in the next sections.
4.2. Integrated Abduction and Object-Level Motion

7 - Commonsense Visual Sensemaking for Autonomous Driving

Published in:
The Artificial Intelligence Journal (AIJ)

Citation:

This journal article was published in the Journal of Artificial Intelligence (AIJ) as an extended version of the work presented in (Suchan, Bhatt, and Varadarajan, 2019) (Pg. 149). It presents a general method for online visual sensemaking in ASP, based on integrated abduction of high-level scene dynamics and low-level motion tracking. Towards this, it builds on the visual abduction capabilities developed in (Suchan, Bhatt, Wałęga, et al., 2018) and extends them to jointly track object motion and generate high-level declarative models explaining the motion tracks in terms of motion patterns and object interactions. Furthermore, the paper utilises incremental solving of the abduction problem to support online execution, providing declaratively grounded models of object interaction on runtime to support decision-making and control within real-world applications.

As a use-case, the paper focuses on the significance of human-centred visual sensemaking —e.g., semantic representation and explainability, question-answering, commonsense interpolation—in safety-critical autonomous driving situations. In this context, the overall system is empirically evaluated and demonstrated with community established real-world datasets and benchmarks in autonomous driving, KITTI and MOT (Section 2.3.3, D3 and D4). The results show a consistent increase in tracking accuracy of approx. 5% when compared to the baseline object tracking not using abduction.
Commonsense visual sensemaking for autonomous driving – On generalised neurosymbolic online abduction integrating vision and semantics

Jakob Suchana,*, Mehul Bhattacharyya, Srikrishna Varadarajanb

a University of Bremen, Germany
b Örebro University, Sweden
c CoDesign Lab (EU), Cognitive Vision, Sweden

A R T I C L E   I N F O

Article history:
Received 19 July 2020
Received in revised form 29 April 2021
Accepted 30 April 2021
Available online 7 May 2021

Keywords:
Cognitive vision
Deep semantics
Declarative spatial reasoning
Knowledge representation and reasoning
Commonsense reasoning
Visual abduction
Answer set programming
Autonomous driving
Human-centred computing and design
Standardisation in driving technology
Spatial cognition and AI

A B S T R A C T

We demonstrate the need and potential of systematically integrated vision and semantics solutions for visual sensemaking in the backdrop of autonomous driving. A general neurosymbolic method for online visual sensemaking using answer set programming (ASP) is systematically formalised and fully implemented. The method integrates state of the art in visual computing, and is developed as a modular framework that is generally usable within hybrid architectures for real-time perception and control. We evaluate and demonstrate with community established benchmarks KITTI/3D, MOT-2017, and MOT-2020. As use-case, we focus on the significance of human-centred visual sensemaking – e.g., involving semantic representation and explainability, question-answering, commonsense interpolation – in safety-critical autonomous driving situations. The developed neurosymbolic framework is domain-independent, with the case of autonomous driving designed to serve as an exemplar for online visual sensemaking in diverse cognitive interaction settings in the backdrop of select human-centred AI technology design considerations.

© 2021 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Motivation

Autonomous driving research has received enormous academic & industrial interest in recent years (Sec 5). This surge has coincided with (and been driven by) advances in deep learning based computer vision research. Although end-to-end deep learning based vision & control has (arguably) been successful for self-driving vehicles, we posit that there is a clear need and tremendous potential for hybrid visual sensemaking solutions that integrate vision and semantics towards fulfilling essential legal and ethical responsibilities involving explainability, human-centred AI (Artificial Intelligence), and industrial standardisation (e.g., pertaining to representation, realisation of rules and norms, fulfilling statutory obligations).

* This paper is an invited contribution based on a distinguished paper nomination at the 2019 International Joint Conference on Artificial Intelligence (IJCAI-19).

E-mail address: info@codesign-lab.org (M. Bhatt).
URL: http://www.codesign-lab.org (M. Bhatt).

https://doi.org/10.1016/j.artint.2021.103522
0004-3702/© 2021 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Autonomous Vehicles: “Standardisation and Regulation”

As the self-driving vehicle industry develops further, it will be necessary to have an articulation and community consensus on aspects such as representation, interoperability, human-centred performance benchmarks, and data archival & retrieval mechanisms. 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 (BMVI) taking a lead in eliciting 20 key propositions1 (with legal implications) for the fulfilment of ethical commitments for automated and connected driving systems [20]. In spite of major investments in self-driving vehicle research, issues related to human-centredness, human collaboration, and standardisation have been barely addressed, with the current focus in driving research primarily being on two basic considerations: how fast to drive, and which way and how much to steer. This is necessary, but inadequate if autonomous vehicles are to become commonplace and function with humans [4,24]. Ethically driven standardisation and regulation will require addressing challenges in foundational human-centred AI technology design, e.g., pertaining to semantic visual interpretation, natural / multimodal human-machine interaction, high-level data analytics (e.g., for post hoc diagnostics, dispute settlement). This will necessitate —amongst other things— human-centred qualitative benchmarks and design & evaluation of multifaceted hybrid solutions integrating diverse methodologies in Artificial Intelligence, Machine Learning, Cognitive Science, Design Science etc.

Neurosymbolism: Visual Sensemaking Needs Both “Vision and Semantics”

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 [18]. Towards this, this research demonstrates the significance of semantically-driven methods 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 is done in the backdrop of the autonomous driving domain; as an example, consider the occlusion scenario in Fig. 1:

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.

The occlusion scenario of Fig. 1 is one of range of (seemingly) mundane safety-critical moments that one may regularly experience while driving a vehicle (Fig. 4, Fig. 7-8 and Table 6 include additional examples). This scenario is sufficiently indicative of several challenges concerning epistemological and phenomenological aspects relevant to a wide range of dynamic spatial systems [11,14,10]:

1 The 20 key propositions elicited by the German federal ministry BMVI highlight a range of factors pertaining to safety, utilitarian considerations, human rights, statutory liability, technological transparency, data management and privacy etc [20].
• **projection and interpolation** of missing information, e.g., what could be hypothesised about bicyclist \(< p, b >\) when it is occluded; 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**, in a manner akin to human cognitive ability to perform mental simulation for purposes of introspection, performing “what-if” reasoning tasks etc

Addressing such challenges —be it realtime or post-hoc— in view of human-centred AI concerns pertaining to representations rooted to natural language, explainability, ethics and regulation requires a systematic (neurosymbolic) 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.

**Deep Semantics: (Systematically) “Integrating AI and Vision”**

The development of domain-independent computational models of perceptual sensemaking —e.g., encompassing capabilities such as visuospatial Q/A, spatio-temporal relational learning, visuospatial abduction— with multimodal human behavioural 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** [12] within a declarative AI setting 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 2)— 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 integrated “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).

**Key Contributions**

This research is situated within the broader auspices of the scientific agenda of cognitive vision and perception, which addresses visual, visuospatial and visuo-locomotive perception and interaction from the viewpoints of language, logic, spatial cognition and artificial intelligence [12] (Sec 5). The key contribution of this paper is to develop a general and systematic declarative visual sensemaking method capable of **online abduction**: realtime, incremental, commonsense question-answering and belief maintenance over dynamic visuospatial imagery. Supported are (1–3):

1. **Human-Centred Representation for Space and Motion**

   Declaratively modelled ontological characterisation of human-centric relational representations that are semantically rooted to commonsense spatio-linguistic primitives pertaining to space and motion as they occur in natural language [16,57].

2. **Systematic High-level Abductive Reasoning**

   Driven by Answer Set Programming (ASP) [25], the ability to abductively compute commonsense interpretations and explanations in a range of (a)typical everyday driving situations, e.g., concerning safety-critical decision-making; the declarative model of space and motion, in addition to supporting abductive reasoning about space and change, is also naturally amenable to high-level semantic interpretation (e.g., by question answering) for post-hoc analytical purposes (e.g., as might be relevant in situations requiring diagnosis et al. for litigation, insurance claims).

3. **Online Performance of Modularly Integrated Vision and Semantics**

   Online performance —in an “active vision” context— of the overall framework modularly integrating high-level commonsense reasoning component with state of the art low-level (deep learning based) visual computing for practical application in real world settings (with autonomous driving serving as a solid demonstration platform).

**Organisation of the Paper**

The rest of the article is organised as follows:

- **Section 2** presents the ontological and formal representational foundations of the developed visual sensemaking framework; main focus is on the commonsense representation aspects pertaining to the modelling of space, space-time, motion, events, and other aspects relevant to modelling and reasoning about spatio-temporal dynamics.
Section 3 presents the overall visual sensemaking framework and its technical implementation with a central focus on the general answer set programming based method for online abduction; we elaborate on the declarative model directly vis-à-vis the ASP implementation.

Section 4 demonstrates & empirically evaluates the core online abduction component with community established real-world datasets and benchmarks, namely: KITTI-MOD [41], MOT-17 [58], and MOT-20 [33].

Section 5 discusses related works primarily from the viewpoints of knowledge representation, and visual computing as pursued in computer vision research.

Section 6 concludes with a brief summary of our work, together with pointers to immediate research questions for follow-up, as well as more broad-based directions that this work aims to open up.

Appendices A–D. Appendix A provides a general overview of the Answer Set Programming paradigm in a manner that is independent of the rest of the paper; Appendix B provides annotations of select Answer Set Programming source code relevant to the declarative model presented in Section 3. Appendix C presents additional examples chosen from community benchmark datasets together with sample data; it also includes an elaborated version of a running example used in the paper. Appendix D provide a succinct view of (select) data corresponding to (select) scenes.

2. Commonsense – space – motion: ontological and representational aspects

We present the ontological and formal representational foundations of the developed visual sensemaking framework while focussing on the commonsense representational aspects pertaining to the modelling of space, space-time, motion, events, and other aspects relevant to modelling and reasoning about spatio-temporal dynamics. Towards this, Table 1 summarises the individual constituents of \( \Sigma_{st} \) (spatiotemporal primitives) and \( \Sigma_{sp} \) (spatiotemporal dynamics), and Table 2 elaborates the supported commonsense relations for the abstraction of space, motion, and (inter)action. Fig. 3 is a (non-exhaustive) collection of generic / domain-neutral spatiotemporal motion patterns supported; Figs. 2 and 4 include concrete instance of such generic motion patterns: Fig. 2 illustrates motion patterns for approach, occlusion, and connected motion; and Fig. 4 illustrates the motion patterns underlying a security-critical scenario involved an elaborate lane changing episode.

2.1. Commonsense abstractions for space and motion

Commonsense spatio-temporal relations and patterns (e.g., left, touching, part of, during, collision) offer a human-centered and cognitively adequate formalism for semantic grounding and automated reasoning for everyday (embodied) multimodal interactions [16,57,73,72]. Qualitative, multi-domain\(^2\) representations of spatial, temporal, and spatio-temporal relations and motion patterns (e.g., Fig. 2–3), and their mutual transitions can provide a mapping between high-level semantic models of actions and events on one hand, and low-level / quantitative trajectory data emanating from visual computing algorithms on the other. For instance, by spatio-linguistically grounding complex trajectory data – e.g., pertaining to on-road moving objects– to a formal framework of space and motion, generalized (activity-based) commonsense reasoning about dynamic scenes, spatial relations, and motion trajectories denoting single and multi-object path & motion predicates can be supported [15,85]. For instance, such predicates can be abstracted within a region-based 4D space-time framework [67,42,6], object interactions [29,30], or even spatio-temporal narrative knowledge [13,17,74,71]. An adequate commonsense spatio-temporal representation can, therefore, connect with low-level quantitative data, and also help to ground symbolic descriptions of actions and objects to be queried, reasoned about, or even manipulated in the real world.

\(^2\) Multi-domain refers to more than one aspect of space, e.g., topology, orientation, direction, distance, shape; this requires a mixed domain ontology involving points, line-segments, polygons, and regions of space, time, and space-time [84,67,42].
Table 1
Commonsense – Space – Motion: Ontological and Representational Aspects.

<table>
<thead>
<tr>
<th>Spatio-temporal Ontology ( (\Sigma_{st}) )</th>
<th>REPRESENTATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain Objects ( D = {o_1, ..., o_n} )</td>
<td>e.g., cars, people, cyclists</td>
</tr>
<tr>
<td>Spatial Entities ( E = {e_1, ..., e_k} )</td>
<td>points, line-segments, rectangles</td>
</tr>
<tr>
<td>Time ( T = {t_1, ..., t_p} )</td>
<td>time-points, time-intervals</td>
</tr>
<tr>
<td>Motion ( MT = {e_{ts}, ..., e_{te}} )</td>
<td>motion tracks / space-time histories</td>
</tr>
</tbody>
</table>

Spatio-Temporal Dynamics \( (\Sigma_{dynt}) \)

- Fluents \( \Phi = \{\phi_1, ..., \phi_l\} \) e.g., visibility, hidden_by, clipped
- Events \( \Theta = \{\theta_1, ..., \theta_m\} \) e.g., hides_behind, missing_detections

Problem Specification

- Visual Observations \( VO_1 = \{obs_1, ..., obs_m\} \) e.g. \( E \) corresponding to object detections
- Predictions \( P_1 = \{P_{e_{ts}}, ..., P_{e_{te}}\} \) e.g. \( E \) for predicted track
- Matching Likelihood \( ML_1 = \{ml_{trk_1, obs_1}, ..., ml_{trk_n, obs_m}\} \) e.g., IoU between tracks and detections

Hypothesis

- Assignments \( H_{events} \) abduced assignments
- Explanations \( \mathcal{E}, \mathcal{X}, \mathcal{P} \leftarrow < H_{events}, MT > \) scene dynamics; abduced events and corresponding motion tracks

Table 2
Commonsense Relations for Abstract Representation of Space, Motion, Interaction.

<table>
<thead>
<tr>
<th>SPATIO-TEMPORAL DOMAIN ( (Q,S) )</th>
<th>Spatial, Time, Motion Relations ( (R) )</th>
<th>Entities ( (E) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mereotopology</td>
<td>disconnected (dc), external contact (ec), partial overlap (po), tangential proper part (tp), non-tangential proper part (npp), proper part (pp), part of (p), discrete (d), overlap (o), contact (c)</td>
<td>arbitrary rectangles, circles, polygons, cuboids, spheres</td>
</tr>
<tr>
<td>Incidence</td>
<td>interior, on boundary, exterior, discrete, intersects</td>
<td>2D point with rectangles, circles, polygones</td>
</tr>
<tr>
<td>Orientation</td>
<td>left, right, colinear, front, back, on, facing towards, facing away, same direction, opposite direction</td>
<td>2D point, circle, polygon with 2D line</td>
</tr>
<tr>
<td>Distance, Size</td>
<td>adjacent, near, far, smaller, equi-sized, larger</td>
<td>rectangles, circles, polygons, cuboids, spheres</td>
</tr>
<tr>
<td>Motion</td>
<td>moving: towards, away, parallel; growing / shrinking: vertically, horizontally; splitting / merging; rotation: left, right, up, down, clockwise, counter-clockwise</td>
<td>rectangles, circles, polygons, cuboids, spheres</td>
</tr>
<tr>
<td>Time</td>
<td>before, after, meets, overlaps, starts, during, finishes, equals</td>
<td>time-points, time intervals</td>
</tr>
</tbody>
</table>

Fig. 3. Commonsense Spatial Reasoning with Spatio-Temporal Entities. Illustrated are: Space-Time Histories for Spatio-temporal Patterns and Events.

2.2. Space, motion, objects, events, change: ontology and formal model

Reasoning about spatio-temporal dynamics is based on high-level representations of objects, and their respective motion & mutual interactions in spacetime. Foundational ontological primitives for commonsense representation and reasoning about spatio-temporal dynamics are:
Fig. 4. Space-Time Histories of Moving Objects: Safety-Criticality Case of a Close Encounter / Car(1) is moving towards car(2) on the right lane, and changes to the left lane to perform an overtaking action; subsequently, car(2) also changes to left lane to overtake car(3) that stopped and is blocking the right lane. To avoid a collision car(1) performs an emergency break and leaves the left lane to the left, entering the lane for the oncoming traffic. (For interpretation of the colours in the figure(s), the reader is referred to the web version of this article.)

- $\Sigma_{st}$ corresponds to primitives for representing space, time, motion and scene-level relational spatiotemporal structure
- $\Sigma_{dyn}$ corresponds to the domain-independent commonsense theory for representing and reasoning about change.

Let $\Sigma \equiv \Sigma_{st} < \Omega, \varepsilon, T, M(T), R> \cup \Sigma_{dyn} < \Phi, \Theta >$ be as follows:  
(Refer Tables 1-2)

- **Domain Objects** ($\Omega$). The high-level, domain-dependent visual elements in the scene, e.g., road-side stakeholders such as people, cars, cyclists, constitute domain objects. Domain objects are denoted by $\Omega = \{o_1, \ldots, o_n\}$; elements in $\Omega$ are geometrically interpreted as spatial entities.

- **Spatial Entities** ($\varepsilon$). Spatial entities correspond to abstractions of domain objects by way of points, line-segments or (axis-aligned) rectangles based on their spatial properties (and a particular reasoning task at hand). Spatial entities are denoted by $\varepsilon = \{e_1, \ldots, e_n\}$.

- **Time** ($T$). The temporal dimension is represented by time points, denoted as $T = \{t_1, \ldots, t_n\}$.

- **Motion Tracks** ($M(T)$). Motion-tracks represent the spacetime motion trajectories (e.g., Fig. 2) of abstract spatial entities ($\varepsilon$) corresponding to domain object ($\Omega$) of interest. $M(T) = (e_{t_1}, \ldots, e_{t_n})$ represents the motion track of a single object $o_i$, where $t_s$ and $t_e$ denote the start and end time of the track and $e_{t_s}$ to $e_{t_e}$ denotes the spatial entity ($\varepsilon$) —e.g., the axis-aligned bounding box—corresponding to the object $o_i$ at time points $t_s$ to $t_e$. Whereas Figs. 2 and 4 presents one example of a space-time trajectory, Fig. 3 is a general (but non-exhaustive) set of patterns supported by our reasoning framework.

- **Spatio-Temporal Relationships** ($R$). The spatial configuration of the scene and changes thereof are characterised based on the spatio-temporal relationships ($R$; Table 2) between abstract representations ($\varepsilon$) of the domain objects ($\Omega$). For the running and demo examples of this paper, positional relations on axis-aligned rectangles based on the Rectangle Algebra (RA) [5] suffice: RA uses the relations of Interval Algebra (IA) [2] $R_{IA} \equiv \{\text{before, after, during, contains, starts, started_by, finishes, finished_by, overlaps, overlapped_by, meets, met_by, equal} \}$ to relate two objects by the interval relations projected along each modelled dimension separately (e.g., horizontal and vertical dimensions).

- **Dynamics / Fluent and Events**. The set of **fluent** $\Phi = \{\phi_1, \ldots, \phi_n\}$ and **events** $\Theta = \{\theta_1, \ldots, \theta_n\}$ respectively characterise the dynamic properties of the objects in the scene and high-level abducibles (e.g., Tables 4 and 5). For reasoning about dynamics (with $< \Phi, \Theta >$), we use the epistemic generalisation of the event calculus [49] as per the formalisation in [55,59]; in particular, for examples of this paper, the Functional Event Calculus (FEC) fragment of Ma et al. [55] suffices.  

\[^3\text{Main axioms relevant for this paper pertain to occurs-at(\theta, t) denoting that an event occurred at time t and holds-at(\phi, v, t) denoting that v holds for a fluent \phi at time t. It is worth noting that in so far as the approach to reason about changes is concerned, our modular framework is by no means limited to the specific approach being utilised. In principle, any method capable of modelling dynamic spatial systems [11] encompassing space, actions, and change [16,14] is usable; basic considerations guiding choice of an action theory pertain to expressivity, modular elaboration tolerance, and support for basic epistemological aspects such as frame and ramification [68]. For instance, other epistemic settings for abductive inference with ASP too may be utilised [35,36].}\]
Problem Specification and Hypothesis.

- **Problem Specification** \( < \mathcal{V} \mathcal{O}, \mathcal{P}, \mathcal{M} \mathcal{L} > \). The abduction for each time point is given by the visual observations \( \mathcal{V} \mathcal{O} \) consisting of spatial entities \( \mathcal{E} \), i.e., bounding boxes for the detected objects, spatial entities \( \mathcal{E} \) of object detections; the predicted locations \( \mathcal{P} \) for each track at time point \( t \) given as spatial entities \( \mathcal{E} \); and the matching likelihood \( \mathcal{M} \mathcal{L} \), i.e., based on the Intersection over Union (IoU) between detected objects and tracks, providing an estimate of how likely a detection belongs to a track.

- **Hypothesis.** Abduced hypothesis consist of assignments \( \mathcal{H}^{assign} \) of detections to tracks and high-level events \( \mathcal{H}^{events} \) explaining object motion, e.g., occlusion of an object, caused by the object passing behind another object. The online abduction results in abduced visuo-spatial dynamics \( \mathcal{E} \mathcal{X} \mathcal{P} \) consisting of motion tracks \( \mathcal{M} \mathcal{T} \) (generated using the abduced assignments in \( \mathcal{H}^{assign} \)) and the events \( \mathcal{H}^{events} \) explaining the motion tracks.

3. Visual sensemaking: a general method driven by answer set programming

Rooted in answer set programming, the developed framework is general, modular, and designed for integration as a reasoning engine within (hybrid) architectures designed for real-time decision-making and control where visual perception is needed as one of the several components. In such large scale AI systems the declarative model of the scene dynamics resulting from the presented framework can be used for semantic question-answering (Q/A), inference etc to support decision-making.

3.1. Tracking as abduction

Our proposed framework, in essence, jointly solves the problem of assignment of detections to tracks and explaining overall scene dynamics (e.g. appearance, disappearance) in terms of high-level events within an online integrated low-level visual computing and high-level abductive reasoning framework (Fig. 5).

Scene dynamics are tracked using a detect and track approach: we tightly integrate low-level visual computing (for detecting scene elements) with high-level ASP-based abduction to solve the assignment of observations to object tracks in an incremental manner. For each time point \( t \) we generate a problem specification consisting of the object tracks and visual observations and use ASP to abductively solve the corresponding assignment problem incorporating the ontological structure of the domain / data (abstracted with \( \Sigma \)).

**Steps 1–3** (Algorithm 1 & Table 3) consist of:
1) Formulating the ASP problem specification consisting of the visual observations, prediction of motion of each object, and a measure for the likelihood that a detection is associated with a track. Further the problem specification contains the state of the world, given by the sequence of events \( \mathcal{H}^{events} \) before time point \( t \).
2) Associating detections to tracks, by jointly abducting matchings between object detections and tracks, together with the high-level events explaining these matches.
Algorithm 1: Online_Abduction(V, Σ)  (Also see Table 3).

Data: Visual imagery (V), and background knowledge Σ \equiv_{def} Σ_{dyn} \cup Σ_{st}

Result: Visual Explanations (E XP)

1. Formulating the ASP Specification The ASP\(^4\) problem specification for each time point \(t\) is given by the tuple \(<V_0, \mathcal{T}, \mathcal{L}_t, >\) and the sequence of events (\(H_{events}^t\)) before time point \(t\).

2. Visual Observations Scene elements derived directly from the visual input data are represented as spatial entities \(E\), i.e., \(V_0 = \{o_{obs1}, ..., o_{obsn}\}\) is the set of observations at time \(t\) (Table 3). For the examples and empirical evaluation in this paper (Sec. 4) we focus on Obstacle / Object Detections – detecting cars, pedestrians, cyclists, traffic lights etc using YOLOv3 [64]. Further we generate scene context using Semantic Segmentation – segmenting the road, sidewalk, buildings, cars, people, trees, etc. using DeepLabv3+ [26], and Lane Detection – estimating lane markings, to detect lanes on the road, using SCNN [61]. Type and confidence score for each observation is given by type \(o_{obsi}\) and conf \(o_{obsi}\).

3. Movement Prediction For each track \(trk\) changes in position and size are predicted using kalman filters; this results in an estimate of the spatial entity \(e\) for the next time-point \(t\) of each motion track \(\mathcal{T}_t = \{e_{trk1}, ..., e_{trkn}\}\).

4. Matching Likelihood For each pair of tracks and observations \(e_{trk}\) and \(e_{obsj}\), where \(e_{trk}\) \(\in\) \(\mathcal{T}_t\) and \(e_{obsj}\) \(\in\) \(V_0\), we compute the likelihood \(\mathcal{L}_t = \{m_{trk,obs1}, ..., m_{trk,obsn}\}\) that \(e_{obsj}\) belongs to \(e_{trk}\). The intersection over union (IoU) provides a measure for the amount of overlap between the spatial entities \(e_{obsj}\) and \(e_{trk}\).

5. Abduction based Association Following perception as logical abduction most directly in the sense of Shanahan [69], we define the task of abducing visual explanations as finding an association (\(H_{assign}^t\)) of observed scene elements (\(V_0\)) to the motion tracks of objects (\(\mathcal{T}\)) given by the predictions \(\mathcal{T}_t\), together with a high-level explanation (\(H_{events}^t\)), such that \(\mathcal{H}_{assign}^t \land H_{events}^t\) is consistent with the background knowledge and the previously abduced event sequence \(H_{events}^t\), and entails the perceived scene given by \(<V_0, \mathcal{T}, \mathcal{L}_t, >\):

\[
\Sigma \land H_{events}^t \land [H_{assign}^t \land H_{events}^t] \models V_0 \land \mathcal{T} \land \mathcal{L}_t
\]

where \(H_{assign}^t\) consists of the assignment of detections to object tracks, and \(H_{events}^t\) consists of the high-level events \(\Theta\) explaining the assignments.

6. Associating Objects and Observations Finding the best match between observations \(V_0\) and object tracks \(\mathcal{T}\) is done by generating all possible assignments and then maximising a matching likelihood \(m_{trk,obsj}\) between pairs of spatial

---

3) Finding the hypothesis and corresponding associations best explaining the visual observations using optimization, i.e., maximizing matching likelihood and minimizing event costs.

In the following we describe each step in detail:

**Step 1. Formulating the Problem Specification** The ASP\(^4\) problem specification for each time point \(t\) is given by the tuple \(<V_0, \mathcal{T}, \mathcal{L}_t, >\) and the sequence of events (\(H_{events}^t\)) before time point \(t\).

**Visual Observations** Scene elements derived directly from the visual input data are represented as spatial entities \(E\), i.e., \(V_0 = \{o_{obs1}, ..., o_{obsn}\}\) is the set of observations at time \(t\) (Table 3). For the examples and empirical evaluation in this paper (Sec. 4) we focus on Obstacle / Object Detections – detecting cars, pedestrians, cyclists, traffic lights etc using YOLOv3 [64]. Further we generate scene context using Semantic Segmentation – segmenting the road, sidewalk, buildings, cars, people, trees, etc. using DeepLabv3+ [26], and Lane Detection – estimating lane markings, to detect lanes on the road, using SCNN [61]. Type and confidence score for each observation is given by type \(o_{obsi}\) and conf \(o_{obsi}\).

**Movement Prediction** For each track \(trk\) changes in position and size are predicted using kalman filters; this results in an estimate of the spatial entity \(e\) for the next time-point \(t\) of each motion track \(\mathcal{T}_t = \{e_{trk1}, ..., e_{trkn}\}\).

**Matching Likelihood** For each pair of tracks and observations \(e_{trk}\) and \(e_{obsj}\), where \(e_{trk}\) \(\in\) \(\mathcal{T}_t\) and \(e_{obsj}\) \(\in\) \(V_0\), we compute the likelihood \(\mathcal{L}_t = \{m_{trk,obs1}, ..., m_{trk,obsn}\}\) that \(e_{obsj}\) belongs to \(e_{trk}\). The intersection over union (IoU) provides a measure for the amount of overlap between the spatial entities \(e_{obsj}\) and \(e_{trk}\).

**Abduction based Association** Following perception as logical abduction most directly in the sense of Shanahan [69], we define the task of abducing visual explanations as finding an association (\(H_{assign}^t\)) of observed scene elements (\(V_0\)) to the motion tracks of objects (\(\mathcal{T}\)) given by the predictions \(\mathcal{T}_t\), together with a high-level explanation (\(H_{events}^t\)), such that \(\mathcal{H}_{assign}^t \land H_{events}^t\) is consistent with the background knowledge and the previously abduced event sequence \(H_{events}^t\), and entails the perceived scene given by \(<V_0, \mathcal{T}, \mathcal{L}_t, >\):

\[
\Sigma \land H_{events}^t \land [H_{assign}^t \land H_{events}^t] \models V_0 \land \mathcal{T} \land \mathcal{L}_t
\]

where \(H_{assign}^t\) consists of the assignment of detections to object tracks, and \(H_{events}^t\) consists of the high-level events \(\Theta\) explaining the assignments.

**Associating Objects and Observations** Finding the best match between observations \(V_0\) and object tracks \(\mathcal{T}\) is done by generating all possible assignments and then maximising a matching likelihood \(m_{trk,obsj}\) between pairs of spatial

---

\(^4\) An introduction to Answer Set Programming from the viewpoint of key aspects relevant to this paper are summarised in Appendix A.
Table 3
Computational Steps for Online Visual Abduction.

For each $t \in T$

**Step 1.** Formulating the Problem Specification
(1) detect Visual Observations ($V_O_t$) e.g., People, Cars, Objects, Roads, Lanes,
(2) Predictions ($P_t$) of next position and size of object tracks using kalman filters, and
(3) calculate Matching Likelihood ($ML_t$) based on Intersection over Union (IoU) between predictions and detections.

```
obs(obs_0,car,99). obs(...). ... box2d(obs_16,1078,86,30,44). ...
trk(trk_0,car). trk(...). ... box2d(trk_0,798,146,113,203). ...
iou(trk_0,obs_0,83921). iou(...). ... iou(trk_23,obs_16,0). ...
```

**Step 2.** Abduction based Association
generate hypothesis for (1) matching of tracks and observations ($H^assign_t$), and (2) and high-level events ($H^{event}_t$) explaining (1).

```
holds-at(visibility(trk),fully_visible,tn). holds-at(visibility(trk),partially_occluded,tn).
occurs-at(gets_behind(trk,trk),tn). occurs-at(hides_behind(trk,trk),tn).
```

**Step 3.** Finding the Optimal Hypothesis
Jointly optimize $H^assign_t$ and $H^{event}_t$ by maximizing matching likelihood $ML_t$ and minimizing event costs.

**RESULT.** Visuo-Spatial Scene Semantics
Resulting motion tracks ($MT$) and the corresponding event sequence ($H^{event}_t$) explaining the low-level motion:

```
... occurs_at(missing_detections(trk_10),35) ... occurs_at(recover(trk_10),36)
... occurs_at(lost(trk_18),41) ... occurs_at(hides_behind(trk_9,trk_10),41)
... occurs_at(...) ... occurs_at(unhides_from_behind(trk_9,trk_10),42) ...
```

entities for matched observations $e_{obs_j}$ and predicted track region $e_{trk_i}$ (see Step 3). Towards this we use choice rules [40] (i.e., one of the heads of the rule has to be in the stable model) for $e_{obs_j}$ and $e_{trk_i}$, generating all possible assignments in terms of assignment actions: assign, start, end, halt, resume, ignore_det, ignore_trk. The corresponding ASP code is as follows:

```
1{ assign(Trk, Det): det(Det, _, _); end(Trk); ignore_trk(Trk); resume(Trk); resume(Trk,Det): det(Det, _, _) }1 :- trk(Trk, _).
```
For each assignment action we define integrity constraints\textsuperscript{5} that restrict the set of answers generated by the choice rules, e.g., the following constraints are applied to assigning an observation $\epsilon_{\text{obs}}$ to a track $\text{trk}_i$, applying thresholds on the $\text{IoU}_{\text{trk}, \text{obs}}$ and the confidence of the observation $\text{conf}_{\text{obs}}$, further we define that the type of the observation has to match the type of the track it is assigned to (e.g., also see Fig. 6):

\begin{verbatim}
:- assign(Trk, Det), not assignment_constraints(Trk, Det).

assignment_constraints(Trk, Det) :-
  trk(Trk, Trk_Type),
  trk_state(Trk, active),
  det(Det, Det_Type, Conf),
  Conf > conf_thresh_assign,
  match_type(Trk_Type, Det_Type),
  iou(Trk, Det, IOU),
  IOU > iou_thresh.
\end{verbatim}

\textsuperscript{5} Integrity constraints restrict the set of answers by eliminating stable models where the body is satisfied.
Table 4
ABDUCIBLES: Events Relevant to Explaining (Dis)Appearance.

<table>
<thead>
<tr>
<th>EVENTS</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>enters_fov(Trk)</td>
<td>Track Trk enters the field of view.</td>
</tr>
<tr>
<td>leaves_fov(Trk)</td>
<td>Track Trk leaves the field of view.</td>
</tr>
<tr>
<td>hides_behind(Trk1, Trk2)</td>
<td>Track Trk1 hides behind track Trk2.</td>
</tr>
<tr>
<td>unhides_from_behind(Trk1, Trk2)</td>
<td>Track Trk1 unhides from behind track Trk2.</td>
</tr>
<tr>
<td>missing_detections(Trk)</td>
<td>Missing detections for track Trk.</td>
</tr>
</tbody>
</table>

Table 5
ABDUCIBLES: Fluents Relevant to Explaining (Dis)Appearance.

<table>
<thead>
<tr>
<th>FLUENTS</th>
<th>Values</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>in_fov(Trk)</td>
<td>(true;false)</td>
<td>Track Trk is in the field of view.</td>
</tr>
<tr>
<td>hidden_by(Trk1, Trk2)</td>
<td>(true;false) partially_occcluded; fully_occcluded</td>
<td>Track Trk1 is hidden by Trk2. Visibility state of track Trk2.</td>
</tr>
<tr>
<td>visibility(Trk)</td>
<td>(true) fully_visible; (false) not_visible, T</td>
<td>Track Trk is interrupted, e.g., missing detection(s).</td>
</tr>
<tr>
<td>clipped(Trk)</td>
<td>(true;false)</td>
<td></td>
</tr>
</tbody>
</table>

- **Abducible High-Level Events** For the length of this paper, we restrict to high-level visuo-spatial abducibles pertaining to object persistence and visibility (Tables 4 and 5): (1) **Occulsion**: Objects can disappear or reappear as result of occlusion with other objects; (2) **Noise and Missing Observation**: (Missing-)observations can be the result of faulty detections.

Let’s take the case of **occlusion**: functional fluent visibility could be denoted **fully_visible, partially_occcluded or fully_occcluded**:

```prolog
fluent(visability(Trk)) :- trk(Trk, _).

possVal(visability(Trk), fully_visible) :- trk(Trk, _).
possVal(visability(Trk), partially_visible) :- trk(Trk, _).
possVal(visability(Trk), not_visible) :- trk(Trk, _).
```

We define the event **hides_behind**, stating that an object hides behind another object by defining the conditions that have to hold for the event to possibly occur, and the effects the occurrence of the event has on the properties of the objects, i.e., the value of the visibility fluent changes to **fully_occcluded**.

```prolog
event(hides_behind(Trk1, Trk2)) :- trk(Trk1, _), trk(Trk2, _).

causesValue(hides_behind(Trk1, Trk2), visability(Trk1), not_visible, T) :-
  trk(Trk1, _), trk(Trk2, _), time(T).
```

For abducing the occurrence of an event we use choice rules that connect the event with assignment actions, e.g., a track getting halted may be explained by the event that the track hides behind another track.

```prolog
1{ occurs_at(hides_behind(Trk, Trk2), curr_time): trk(Trk2, _); ...
}1 :- halt(Trk).
```

**Step 3. Finding the Optimal Hypothesis** To ensure an **optimal assignment**, we use ASP based optimization to maximize the matching likelihood between matched pairs of tracks and detections. Towards this, we first define the matching likelihood based on the Intersection over Union (IoU) between the observations and the predicted boxes for each track as described in [9]:

```
...
Fig. 7. Safety-Critical Situation (select prototypes): (a) momentarily occluded / hidden entities; (b) overtaking / lane-crossing situation; (c) blocked visibility; and (d) suddenly appearing objects.

Table 6: Select Safety-Critical Situations.

<table>
<thead>
<tr>
<th>SITUATION</th>
<th>Objects</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>OVERTAKING</td>
<td>vehicle, vehicle</td>
<td>vehicle is overtaking another vehicle in front of the car.</td>
</tr>
<tr>
<td>HIDDEN_ENTITY</td>
<td>entity, object</td>
<td>traffic participant may be hidden by an obstacle, e.g., another car or van.</td>
</tr>
<tr>
<td>REDUCED_VISIBILITY</td>
<td>object</td>
<td>visibility is reduced by some object in front of the car.</td>
</tr>
<tr>
<td>SUDDEN_STOP</td>
<td>vehicle</td>
<td>vehicle in front of the car is suddenly stopping.</td>
</tr>
<tr>
<td>BLOCKED_LANE</td>
<td>lane, object</td>
<td>lane of the road is blocked by some object.</td>
</tr>
<tr>
<td>EXITING_PARKED_VEHICLE</td>
<td>person, vehicle</td>
<td>person is exiting a parked vehicle.</td>
</tr>
</tbody>
</table>

We then maximize the matching likelihood for all assignments, using the build in maximize statement:

\[
\text{#maximize \{ML@10, Trk, Det : assign(Trk, Det), matching\_likelihood(Trk, Det, ML)\}.}
\]

To find the best set of hypotheses with respect to the observations, we minimize the occurrence of certain events and association actions, e.g., the following optimization statements minimize starting and ending tracks; the resulting assignment is then used to update the motion tracks accordingly.

\[
\text{#minimize \{S@2, Trk: end(Trk)\}.}
\]
\[
\text{#minimize \{S@2, Det: start(Det)\}.}
\]

It is important here to note that: (1), by jointly abducting the object dynamics and high-level events we can impose constraints on the assignment of detections to tracks, i.e., an assignment is only possible if we can find an explanation supporting the assignment; and (2), the likelihood that an event occurs guides the assignments of observations to tracks. Instead of independently tracking objects and interpreting the interactions, this yields to event sequences that are consistent with the abducted object tracks, and noise in the observations is reduced (See evaluation in Sec. 4).

4. Evaluation: application and empirical performance analysis

We demonstrate applicability towards identifying and interpreting safety-critical situations (e.g., Table 6; Figs. 7, 8; Fig. 4); these encompass those scenarios where interpretation of spacetime dynamics, driving behaviour, environmental characteristics is necessary to anticipate and avoid potential dangers. We also provide an empirical evaluation of the active sensemaking framework in the context of community benchmark datasets.
Fig. 8. Sample Safety-Critical Episodes: (a) overtaking event in front of the car; (b) occlusion while turning left; (c) abrupt lane change on the highway; (d) pedestrian suddenly appearing from between two parked cars; and (e) (relatively) crowded and chaotic inner city traffic.

Fig. 9. Abducting ‘Hiding Behind’ Event.

4.1. Application: visual perception by abduction

In the context of active vision as relevant to autonomous driving, we demonstrate select examples focussing on abducting appearance, disappearance, and occlusion events.

4.1.1. Abducting explanations: appearance and disappearance

Consider the scene in Fig. 9, where a car is passing behind a bus and is getting hidden during this. When the car hides behind the bus (time point 235), the track trk.13 gets halted and the event hides_beind is abduced to explain why the car is not detected anymore and the corresponding track is halted.

The problem specification for time point 235 (< V_O 235, P 235, M_L 235 >) is given as follows:

- V_O 235 the visual observation, consisting of the object detections, given by the bounding box, the type and the confidence:

  det(det_0, person, 99). det(det_1, bus, 99). det(det_2, traffic_light, 86).
  det(det_3, traffic_light, 81). det(det_4, traffic_light, 78).
  det(det_5, traffic_light, 59).
Solving the assignment of detections to tracks can now be done based on the choice rules for associating objects and observations, detailed in Section 3.1 Step 2. To restrict the assignment we can impose constraints on the matching, by stating integrity constraints, e.g., for ensuring that only tracks and detections with the same type are matched, we could state the following integrity constraint. Stating that any stable model where the body is satisfied can not be in the set of answers, i.e., any model assigning a track and a detection which are not of the same type can not be an answer. Further, the track has to be active, the confidence of the detection has to be above a threshold, and the IoU between the track and the detection has to be above a threshold:

\begin{verbatim}
:- assign(Trk, Det), not assignment_constraints(Trk, Det).
assignment_constraints(Trk, Det) :- 
  trk(Trk, Trk_Type), det(Det, Det_Type, Conf),
  trk_state(Trk, active),
  match_type(Trk_Type, Det_Type),
  Conf > conf_thresh_assign,
  iou(Trk, Det, IOU), IOU > iou_threshold.
\end{verbatim}

By maximizing the matching likelihood we get the optimal assignment of detections to tracks, in our example the bus is detected by detection det_1 which gets assigned to the corresponding track trk_12, but as the car is hiding behind the bus, there is no corresponding detection, thus the track of the car trk_13 gets halted:

\begin{verbatim}
halt(trk_13) assign(trk_15, det_0) assign(trk_12, det_1) assign(trk_8, det_2)
assign(trk_3, det_3) assign(trk_7, det_4)
\end{verbatim}

The assignment actions are linked with high-level events for explaining the assignments, i.e., the halted track trk_13 can be explained either by missing detections or by the track hiding behind another track. In this case track trk_13 is hiding behind track trk_12, this can be abduced based on possible events, which in this case is the hides_behind event.

For the event hides_behind/2 the predicted tracks have to be overlapping. This is ensured by (spatial) preconditions of the event, given by the predicate poss/1:

\[ \text{Poss} \text{ _12, } \text{Poss} \text{ _13, } \text{Poss} \text{ _7, } \text{Poss} \text{ _4} \]

Note, only those IoU values are stated which are bigger than 0.
In our example we can now abduce that the track \textit{trk\_13} representing the car is ended, because the car got hidden by the bus represented by track \textit{trk\_12}. In the formal representation of event calculus this is represented by the predicate \textit{occurs\_at} as follows:

\[
\text{poss(hides\_behind(Trk1, Trk2)) :-}
\text{trk(Trk1, _), trk(Trk2, _),}
\text{position(overlapping\_top, Trk1, Trk2),}
\text{not holds\_at(visibility(Trk1), not\_visible, curr\_time),}
\text{not holds\_at(visibility(Trk2), not\_visible, curr\_time).}
\]

At time point 268 the car reappears, after passing behind the bus. Due to the previously abduced event \textit{hides\_behind}, the visibility fluent for the track of the car \textit{trk\_13} has now the value \textit{not\_visible}.

For the detection \textit{det\_1} we can then abduce that track \textit{trk\_13} unhides from behind track \textit{trk\_12} based on the following event definition, stating that the event \textit{unhides\_from\_behind} is possible when Trk1 is \textit{not\_visible} and Trk2 is not \textit{not\_visible}:

\[
\text{poss(unhides\_from\_behind(Trk1, Trk2)) :-}
\text{trk(Trk1, _), trk(Trk2, _),}
\text{holds\_at(visibility(Trk1), not\_visible, curr\_time),}
\text{not holds\_at(visibility(Trk2), not\_visible, curr\_time).}
\]

\[
\text{resume(trk\_13,det\_1) assign(trk\_15,det\_0) assign(trk\_12,det\_2) assign(trk\_7,det\_3)}
\text{assign(trk\_8,det\_4) assign(trk\_3,det\_5)}
\]

\[
\text{occurs\_at(unhides\_from\_behind(trk\_13,trk\_12),268))}
\]

Similarly, when looking at a slightly more complex scene, like the one depicted in Fig. 10, we get an event sequence describing the interactions happening in the scene:

\[
\text{... occurs\_at(hides\_behind(trk\_34,trk\_16),283)}
\text{occurs\_at(unhides\_from\_behind(trk\_34,trk\_16),293)}
\text{occurs\_at(hides\_behind(trk\_37,trk\_34),296)}
\text{occurs\_at(unhides\_from\_behind(trk\_37,trk\_34),311)}
...\]

This event sequence explains the visuospatial dynamics of the scene and can be used for reasoning about the scene.
4.1.2. Reasoning about hidden entities

Consider the situation of Fig. 11: a car gets occluded by another car turning left and reappears in front of the autonomous vehicle. Using online abduction for abducing high-level interactions of scene objects we can hypothesize that the car got occluded and anticipate its reappearance based on the perceived scene dynamics.

The prediction for each track is given by the predicted bounding box, the state in which the track currently is, and the type of the tracked object:

```
trk(trk_3, car). trk_state(trk_3, active).
...
trk(trk_41, car). trk_state(trk_41, active).
...
box2d(trk_3, 660, 460, 134, 102).
...
box2d(trk_41, 631, 471, 40, 47).
...
```

Based on this problem specification for time point 179, the event \texttt{hides\_behind(trk}\texttt{\_41, trk}\texttt{\_3)} is abduced, as there is no detection that could be associated with \texttt{trk}\texttt{\_41} and \texttt{trk}\texttt{\_3} is partially overlapping with \texttt{trk}\texttt{\_41}:

```
... occurs\_at(hides\_behind(trk}\texttt{\_41, trk}\texttt{\_3), 179) ...
```

The abduced explanation together with the object dynamics may then be used for visual reasoning and anticipation of events, which can serve for decision support. Towards this we define a rule stating that a hidden object may unhide from behind the object it is hidden by and anticipate the time point \( t \) based on the object \texttt{movement} as follows:

```
anticipate(unhides\_from\_behind(Trk1, Trk2), T) :-
time(T), curr\_time < T, holds\_at(hidden\_by(Trk1, Trk2), curr\_time),
 topology(proper\_part, Trk1, Trk2),
 movement(moves\_out\_of, Trk1, Trk2, T).
```

We then interpolate the objects position at time point \( t \) to predict where the object may \texttt{reappear}:

```
point2d(interpolated\_position(Trk, T), PosX, PosY) :-
time(T), curr\_time < T, T1 = T-curr\_time,
 box2d(Trk1, X, Y, _, _), trk\_mov(Trk1, MovX, MovY),
 PosX = X+MovX*T1, PosY = Y+MovY*T1.
```

For the occluded car in our example we get the following prediction for time \( t \) and position \( x, y \):

```
anticipate(unhides\_from\_behind(trk}\texttt{\_41, trk}\texttt{\_2), 202}
point2d(interpolated\_position(trk}\texttt{\_41, 202), 738, 495)
```
### Table 7

<table>
<thead>
<tr>
<th>BENCHMARK</th>
<th>MOTA↑</th>
<th>MOTP↑</th>
<th>ML↓</th>
<th>MT↑</th>
<th>FP↓</th>
<th>FN↓</th>
<th>ID sw↓</th>
<th>Frag↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>KITTI tracking – Cars (8008 frames, 636 targets)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>– baseline</td>
<td>45.72%</td>
<td>76.89%</td>
<td>19.14%</td>
<td>23.04%</td>
<td>785</td>
<td>11182</td>
<td>1097</td>
<td>1440</td>
</tr>
<tr>
<td>– with Abd.</td>
<td><strong>50.5%</strong></td>
<td>74.76%</td>
<td><strong>20.21%</strong></td>
<td>23.23%</td>
<td>1311</td>
<td>10439</td>
<td>165</td>
<td>490</td>
</tr>
<tr>
<td>KITTI tracking – Pedestrians (8008 frames, 167 targets)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>– baseline</td>
<td>28.71%</td>
<td>71.43%</td>
<td>26.94%</td>
<td>9.58%</td>
<td>1261</td>
<td>6119</td>
<td>539</td>
<td>833</td>
</tr>
<tr>
<td>– with Abd.</td>
<td><strong>32.57%</strong></td>
<td>70.68%</td>
<td><strong>22.15%</strong></td>
<td>14.37%</td>
<td>1899</td>
<td>5477</td>
<td>115</td>
<td>444</td>
</tr>
<tr>
<td>MOT17 (5316 frames, 346 targets)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>– baseline</td>
<td>41.4%</td>
<td>88.0%</td>
<td>35.53%</td>
<td>16.48%</td>
<td>4877</td>
<td>60164</td>
<td>779</td>
<td>741</td>
</tr>
<tr>
<td>– with Abd.</td>
<td><strong>46.2%</strong></td>
<td>87.9%</td>
<td><strong>31.32%</strong></td>
<td>20.7%</td>
<td>5195</td>
<td>54421</td>
<td>800</td>
<td>904</td>
</tr>
<tr>
<td>MOT2020 (8931 frames, 2332 targets)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>– baseline</td>
<td>49.5%</td>
<td>87.1%</td>
<td>17.9%</td>
<td>18.19%</td>
<td>5271</td>
<td>531529</td>
<td>36560</td>
<td>39874</td>
</tr>
<tr>
<td>– with Abd.</td>
<td><strong>50.7%</strong></td>
<td>87.2%</td>
<td><strong>18.65%</strong></td>
<td>17.16%</td>
<td>4120</td>
<td>537427</td>
<td>17658</td>
<td>38346</td>
</tr>
</tbody>
</table>

Based on this prediction we can then define a rule that gives a warning if a hidden entity may reappear in front of the vehicle, which could be used by the control mechanism, e.g., to adapt driving and slow down in order to keep safe distance:

```prolog
warning(hidden_entity_in_front(Trk1, T)) :-
  time(T), T-cur_r_time < anticipation_threshold,
  anticipate(unhides_from_behind(Trk1, __, T), position(in_front, interpolated_pos(Trk1, T))).
```

#### 4.2. Empirical performance analysis

For online sensemaking, evaluation focusses on accuracy of abduced motion tracks, real-time performance, and the trade-off between performance and accuracy. Our evaluation uses the KITTI object tracking dataset [41], which is a community established benchmark dataset for autonomous cars: it consists of 21 training and 29 test scenes, and provides accurate track annotations for 8 object classes (e.g., car, pedestrian, van, cyclist). We also evaluate tracking results using the more general cross-domain Multi-Object Tracking (MOT) dataset [58] established as part of the MOT Challenge; We evaluate on MOT 2017 consisting of 7 training and 7 test scenes which are highly unconstrained videos filmed with both static and moving cameras, and MOT 2020 consisting of 4 training and 4 test scenes filmed in crowded environments. We evaluate on the available groundtruth for tracking scenes of both KITTI using YOLOv3 detections, and MOT17 / MOT20 using the provided faster RCNN (Region Based Convolutional Neural Network [65]) detections.

#### 4.2.1. Evaluating object tracking

For evaluating accuracy (MOTA) and precision (MOTP) of abduced object tracks we follow the Clear MOT [8] evaluation schema.

- **MOTA** describes the accuracy of the tracking, taking into account the number of missed objects / false negatives (FN), the number of false positives (FP), and the number of miss-matches (MM).
- **MOTP** describes the precision of the tracking based on the distance of the hypothesised track to the ground truth of the object it is associated to.

These metrics are used to assess how well the generated visual explanations describe the low-level motion in the scene.

Results (Table 7) show that jointly abducting high-level object interactions together with low-level scene dynamics increases the accuracy of the object tracks, i.e., we consistently observe an improvement of about 5% on KITTI and MOT 2017. On KITTI MOTA improves from 45.72% to 50.5% for cars and 28.71% to 32.57% for pedestrians, and on MOT 2017 it improves from 41.4% to 46.2%. On MOT 2020 we still observe an improvement of 1.2% from 49.5% to 50.7%. This relatively small improvement is mainly because of the different nature of the dataset, i.e., the focus on crowded scenes filmed from a slightly above perspective, which leads to only few targets that get fully occluded by others, and thus there are fewer corrected tracks when using abductive sensemaking compared to the scenes in KITTI and MOT 2017.

#### 4.2.2. Online performance and scalability

Performance of online abduction is evaluated with respect to its real-time capabilities.\(^7\) (1). We compare the time & accuracy of online abduction for state of the art (real-time) detection methods: YOLOv3, SSD [54], and Faster RCNN [65].

---

\(^7\) Evaluation using a dedicated Intel Core i7-6850K 3.6GHz 6-Core Processor, 64GB RAM, and a NVIDIA Titan V GPU 12GB.
(Fig. 12). (2) We evaluate scalability of the ASP based abduction on a synthetic dataset with controlled number of tracks and percentage of overlapping tracks per frame. Results (Fig. 13) show that online abduction can perform with above 30 frames per second for scenes with up to 10 highly overlapping object tracks, and more than 50 tracks with 1fps (for the sake of testing, it is worth noting that even for 100 objects per frame it only takes about an average of 4 secs per frame). Importantly, for realistic scenes such as in the KITTI dataset, abduction runs realtime at 33.9fps using YOLOv3, and 46.7 using SSD with lower accuracy but providing good precision.

4.2.3. Discussion of empirical results

Results show that integrating high-level abduction and object tracking improves the resulting object tracks and reduce the noise in the visual observations. For the case of online visual sense-making, ASP based abduction provides the required performance: even though the complexity of ASP based abduction increases quickly, with large numbers of tracked objects the framework can track up to 20 objects simultaneously with 30fps and achieve real-time performance on the KITTI benchmark dataset. It is also important to note that the tracking approach in this paper is based on tracking by detection using a naive measure, i.e., the IoU (Sec. 3.1; Step 1), to associate observations and tracks, and it is not using any visual information in the prediction or association step. Naturally, this results in a lower accuracy, in particular when used with noisy detections and when tracking fast moving objects in a benchmark dataset such as KITTI. That said, due to the modularity of the implemented framework, extensions with different methods for predicting motion (e.g., using particle filters or optical flow based prediction) are straightforward: i.e., improving tracking is not the aim of our research.

5. Discussion and related work

Answer Set Programming is now widely used as a foundational declarative language and robust methodology for a range of (non-monotonic) knowledge representation and reasoning tasks [25,66,39,38,40]. With ASP as a foundation, and driven by semantics, commonsense and explainability [32,31], this research aims to bridge the gap between high-level formalisms for logical visual reasoning (e.g., by abduction) and low-level visual processing by tightly integrating semantic abstractions of space and change with their underlying numerical representations. More broadly, this goal is pursued within the larger agenda of cognitive vision and perception [12], 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. 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 [78,70,77,75,76]. In this broader context, the principal motivation and developmental goal of this research follows a one-point agenda [12,79], namely:

- to develop a systematic, general, and modular integration of (methods in) Computer Vision and AI, particularly emphasizing the integration of high-level knowledge representation and reasoning techniques with low-level (i.e., quantitatively) based visual computing techniques (which in the present scientific status quo are primarily driven by end-to-end, black-box deep learning pipelines).

The integration of Vision and AI addressed in our research is motivated by the need to realise human-centred criteria pertinent to the design and implementation of high-level visual sensemaking technology, e.g., within autonomous driving systems where such criteria emanating from standardisation and regulation considerations are of utmost priority. Although this paper selectively focusses on the needs and challenges of active / online sensemaking in autonomous driving, the generality and modularity of the developed framework ensures foundational applicability in diverse applied contexts requiring perception, interaction and control; e.g., a case in point here being the fact that the demonstrated application and evaluation
also directly function with general datasets such as MOT concerned with moving objects (Sec 4). 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 reasoning about space, actions, and change [10,11] that could either be embedded directly within answer set programming, or possibly be utilised independently as part of other declarative frameworks for knowledge representation and reasoning.

Perception and Abduction: A KR Perspective. Within KR, the significance of general abduction and high-level abductive explanations in a range of contexts is long established: planning & process recognition [46,45], vision & abduction [69], probabilistic abduction [19], reasoning about spatio-temporal dynamics [11], reasoning about continuous spacetime change [60,43], general abduction in ASP and related logics [53,23,22,35,36] etc. Dubba et al. [34] formalises abductive reasoning in an inductive-abductive loop within inductive logic programming (ILP). Aditya et al. [1] formalise general rules for image interpretation with ASP. Closely related to this research is [77], which uses a two-step approach (with one huge problem specification), first tracking and then explaining (and fixing) tracking errors; such an approach is not runtime / realtime capable. Within computer vision research there has recently been an interest to synergise with cognitively motivated methods; in particular, e.g., for perceptual grounding and inference [89], and combining video analysis with textual information for understanding events and answering queries about video data [82].

Perception in Autonomous Driving. The present industrial relevance and market potential of autonomous driving technology can be primarily attributed to recent advances in deep learning driven computer vision. A typical engineering stack for autonomous driving consists of perception, prediction, planning and control modules [91]: perception gives the location, pose of the objects in the world while prediction forecasts the motion of the objects; planning involves creating a trajectory for the motion of the vehicle which is then executed by the controller. In object detection, Tan and Le [81] introduced EfficientDet which achieves order-of-magnitude better efficiency than previous works [64,65,54,81] without any drop in performance. Large datasets and self-supervised methods [27,92] enable end to end joint learning of flow [86], depth and camera pose estimation more accurately, exploiting the inherent relation between each other. More specialised research on object detection has investigated specific cases relevant in driving, such as detection of smaller objects [88], partially occluded pedestrians [62] and 3D object detection [87,51,93]. Recent advancements in object tracking involves neural methods like [7] which employ a tracking by detection paradigm and predicts the next object position using a simple neural network. Multi-object tracking is also extended to multi-object tracking and segmentation by [83]. Semantic and Instance segmentation of the object [28] [94] [90] [80] provides accurate boundaries. Advancements in robust lane detection [44] make it possible to extend automatic lane keeping and lane switching. Recent neural methods estimate visual odometry, ego motion, depth and flow through a set of multi-task learning methodologies. [92] [56] show that depth and ego-motion can be learnt in a joint manner. Flow and depth are also learnt using a multi-task approach [27] [95] [86].

Hybrid Methods to Meet Multi-Faceted Challenges. Critical challenges in driving, e.g., pertaining to perception, prediction, planning and control modules [91], are researched and developed individually which leads to a sub-optimal overall performance. End-to-end driving methodologies [21,91] are constructed in such a way that the sensor outputs such as images, LiDAR (Light Detection and Ranging) are directly used to predict control signals like steering and acceleration. Furthermore, these methods are generally black-box and are unable to model the complex multi-faceted nature of autonomous driving encompassing dimensions of human factors and usability, (natural) roadside multimodal interaction [48,47] etc, or support the range of human-centred AI considerations related to declarative explainability, queryability etc that have been the principal impulse underlying the aims of the methods developed in this paper.

Our research achieves a systematic integration of KR and Vision methods hitherto developed, evaluated, and applied in completion isolation of one another; we believe that our resulting framework can serve as a one possible interpretation and exemplar for the neurosymbolic integration of relational AI and neural (visual) feature detectors. Furthermore, it offers a novel potential for a multifaceted but integrated applied evaluation and benchmarking of visual sensemaking technologies; e.g., it is common practice within computer vision research to evaluate and benchmark visual computing capabilities, e.g., for object detection, tracking, using absolute performance benchmarks either solely or primarily centred on incremental improvements in accuracy. Naturally, this is necessary for fundamental progress in vision research, but such an evaluation metric misses out on other crucial requirements as they pertain to human-centred AI considerations in applications domains such as autonomous driving. For instance, in light of ethically driven standardisation and regulatory considerations (Section 1), this research has been motivated and directly addresses interpretability and explainability challenges (C1 – C4):

C1. Active visual sensemaking, e.g., involving (real-time) commonsense visuospatial abduction and (simulated) prediction of grounded percepts.

---

1 Industrial initiatives in autonomous driving. Autonomous driving research within industry is now well established: there exist cab-sharing companies like Uber and Lyft attempting to replace human drivers with “fully autonomous self-driving” vehicles. Companies such as Baidu, Comma AI and organisations like Udacity are creating an open source platform for various technologies of the self driving stack. Manufacturing giants such as GM, Toyota, Ford, Daimler, Bosch are also taking steps to offer varying levels of autonomy to consumer and industrial vehicles either directly or indirectly; GM acquired Cruise Automation while Toyota has invested in and collaborates with pony.ai. Ford and Volkswagen has partnered with Argo AI to bring self driving capabilities to their respective vehicles. Waymo, a subsidiary of Alphabet has already deployed autonomous ride-sharing operations in two cities. Last, but not the least, is Tesla with its competitive advantage of already having close to 500000 cars on the road collecting data with (Level 2 assistance) AutoPilot enabled.
C2. Posthoc analysis of quantitative archives, e.g., requiring semantic search / retrieval / visualisation for diagnosis, dispute settlement, inspection

C3. Natural human-machine interaction, e.g., involving natural language interfaces for (explanatory) communication between vehicle and passengers (or other stakeholders)

C4. Standardisation for vehicular licensing & validation, e.g., involving creation of diverse, naturalistic datasets usable in testing of autonomous vehicle performance; how to access the quality and distribution of training datasets utilised? (Sec 6)

Our research, by its integrative approach, makes it possible to explicitly address “human-centred interpretability and explainability challenges” such as in (C1–C4) for autonomous driving systems at the practical level of methods and tools. This is especially beneficial and timely since not everything in autonomous vehicles is about realtime control / decision-making; several human-machine interaction requirements (e.g., for interpretable diagnostic communication, universal design) also exist. The Federal Ministry of Transport and Digital Infrastructure in Germany (BMVI) has taken a lead in eliciting 20 key propositions (with possible legal implications) for the fulfilment of ethical commitments for automated and connected driving systems [20]. The BMVI report highlights a range of factors pertaining to safety, utilitarian considerations, human rights, statutory liability, technological transparency, data management and privacy etc. We claim that what appears as spectrum of complex challenges (in autonomous driving) that may possibly delay technology adoption is actually rooted to one fundamental methodological consideration that needs to be prioritised, namely: the design and implementation of human-centred technology based on a confluence of techniques and perspectives from AI+ML, Cognitive Science & Psychology, Human-Machine Interaction, and Design Science. Like in many applications of AI, such an integrative approach has so far not been explored also within autonomous driving research.

6. Summary and outlook

We have developed a novel neurosymbolic abduction-driven online (i.e., realtime, incremental) visual sensemaking framework: general, systematically formalised, modular, and fully implemented. Integrating robust state-of-the-art methods in knowledge representation and computer vision, the framework has been evaluated and demonstrated with established community benchmarks. We highlight application prospects of semantic vision for autonomous driving, a domain of emerging and long-term significance for research in Artificial Intelligence and Machine Learning. From the applied viewpoint of autonomous driving, our work is motivated by interpretability and explainability benchmarks (e.g., in active visual sensemaking, posthoc analysis, natural human-machine interaction, standardisation for licensing & validation; Sec 4.1) that go far beyond basic considerations in contemporary autonomous driving research, namely: how fast to drive and which way to steer, and testing performance by clocking mileage alone by the use of deep learning based methods in training and testing phases.

Technical Extensions. Our development of a systematic, modular, and general visual sensemaking methodology opens up several possibilities for further technical developments / extensions:

- **Commonsense.** Specialised commonsense theories about multi-sensory integration, multi-agent belief merging, incorporation of contextual knowledge and situational norms within the declarative framework of ASP merits individual strands of further development.

- **Tracking by detection.** Given the modularity of the developed framework, incorporating and experimenting with specialised / emerging low-level visual computing methods becomes feasible with relative ease. For instance, in this paper we have not attempted to develop a new tracking algorithm as such; instead, one of our aims has been to showcase the manner in which perceptual sensemaking by visual abduction can be integrated into a standard “tracking by detection” paradigm, which is most widely used approach in state of the art tracking (Sec 5). Nevertheless, extensions and variations of this approach deserve further investigation where tracking itself takes a centre-stage.

- **Uncertainty.** The present work handles the uncertainty involved in low-level object tracking using a naïve approach, which suffices for the present purposes, i.e., a full-scale systematic formalisation of a probabilistic model has not been attempted herein. However, handling uncertainty calls for its systematic treatment, e.g., requiring either integrating a declarative probabilistic model directly within the answer set programming framework, or possibly independently as a separate module. One seemingly natural approach towards this would be to explore possibilities with probabilistic ASP [50].

Towards a Dataset: Reasoning and Scenario Visuospatial Complexity Coverage. The application demonstrations of this paper have been conducted in the backdrop of select safety-critical situations (Table 6; e.g., Fig. 7), without aiming to achieve an exhaustive collection (if at all it is even possible to be comprehensive in this respect). The scenarios and corresponding safety-criticality are exemplary, with the selection emanating from a behavioural study of human-factors in everyday driving situations, and safety criticality determined based on analysis of empirical data about roadside accidents / hazardous situations from publicly available data published in accident research reports, e.g., by the German Insurance Association (“Unfallforschung der Versicherer”) [37]. Work is presently in progress to develop novel benchmark datasets (in synergy with behavioural human studies; refer below) that centralise range and distribution vis-à-vis commonsense explainability and vi-
suospatial complexity ([47,48]) criteria classes within a dataset, as opposed to merely collecting accumulating “mileage” / “big data”.

**Human-Factors in Autonomous Driving: A Cognitive Methodology Combining Behavioural and Computational Approaches**

In addition to continuing (aforediscussed) technical developments in computational cognitive vision pertaining to the integration of “vision & AI”, our ongoing focus is to develop a novel dataset emphasising (visuospatial) semantics and (commonsense) explainability. For instance, we develop a methodology —focussing on visuospatial complexity [48] of stimuli and multimodal interactions [47] in ecologically valid naturalistic [3,63] driving conditions— for establishing human-centred benchmarks and corresponding testing & validation datasets for visual sensemaking primarily from a human cognitive factors viewpoint. Our particular focus here is on embodied multimodal interactions (e.g., gestures, joint attention, visual search complexity) amongst drivers, pedestrians, cyclists etc under ecologically valid naturalistic conditions. This initiative is driven by bottom-up interdisciplinary research—encompassing AI, Psychology, HCI, and Design— for the study of driving behaviour particularly in diverse low-speed, complex urban environments possibly with unstructured traffic. Such interdisciplinary studies —at the confluence of Cognition, AI, Interaction, and Design— are needed to better appreciate the complexity and spectrum of varied human-centred challenges in autonomous driving, and demonstrate the significance of integrated vision & AI solutions [12] in those contexts.

**Declaration of competing interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Acknowledgements**

This article is an extended version of an IJCAI 2019 publication [78]. We thank the reviewers at IJCAI 2019 for their constructive feedback and support of [78]; all reviewer suggestions that could not be included in the conference length paper have been fully incorporated in the present article. We are also grateful to the anonymous reviewers and editors at the AIJ journal whose comments have helped us further improve the paper; we remain especially appreciative for their timely service during the special times of 2020.

We acknowledge partial funding by the German Research Foundation (DFG) via the Collaborative Research Center 1320 EASE – “Everyday Activity Science and Engineering” (http://www.ease-crc.org) project:

**Spatial Reasoning in Everyday Activity**, Number 329551904.

http://gepris.dfg.de/gepris/projekt/37412335

The overall scientific agenda (pertaining to Cognitive Vision and Deep Semantics [12]) driving this research is available at: CoDesign Lab (EU) > Cognitive Vision / https://codesign-lab.org/cognitive-vision/

Related select publications: https://codesign-lab.org/select-papers/#cognitive_vision

**Appendices A–D**

Appendix A. Answer Set Programming
Appendix B. Select Answer Set Programming Code
Appendix C. Additional Examples
Appendix D. Example Data

Appendix A provides a general overview of the Answer Set Programming paradigm in a manner that is independent of the rest of the paper; Appendix B provides annotations of select Answer Set Programming source code relevant to the declarative model presented in Section 3. Appendix C presents additional examples chosen from community benchmark datasets together with sample data; it also includes an elaborated version of a running example used in the paper. Appendix D provide a succinct view of (select) data corresponding to (select) scenes.

**Appendix A. Answer set programming**

Answer Set Programming (ASP) is an influential declarative programming methodology designed –at its core– to solve complex (NP-hard) search and optimisation problems [25]. An outgrowth and culmination of state-of-the-art KR research in logic programming, constraint satisfaction, and the stable model semantics of logic programs, ASP is now established as a general and powerful knowledge-centred reasoning method within AI [52]. Answer set programs follow the generate and test paradigm, where first a set of candidate solutions is generated and subsequently invalid candidates are eliminated (e.g., based on domain constraints, choice rules).
In the following, we include basic minimal examples of the key aspects of an ASP program as relevant to understanding the examples developed in this paper, including the code snippets provided in the following appendices.\footnote{For a detailed tutorial on ASP, please consult the material developed by Gebser et al. \cite{gebser2018}, also available via online courses at: \url{https://teaching.potassco.org}.}

**ASP Programs.**

An answer set program, similar to a logic program, is essentially a set of rules, where each rule is composed of the head of the rule and its body. Rules are defined as per the following notation:

head :- body\(_1\), body\(_2\), \ldots , body\(_n\)

Here, `:-` reads as *if* and separates the head of the rule from its body. The symbol `,' denotes a conjunction, and `;' denotes a disjunction. Rules where the body is empty constitute primitive facts and rules where the head is empty constitute constraints. A rule is considered to be true if all elements of its body are also true.

Negation may be denoted in two different ways, i.e., ASP provides constructs for negation by failure using `not`, stating that a rule or atom is not true, and for strong or classical negation `\neg`, explicitly stating that a rule or atom is *false*.

head :- not body.

head :- \neg body.

ASP also supports the use of constants and variables, where variables represent a collection of ground instances given by the constants. Constants are denoted by names where the first letter is lowercase, and variables are denoted by names where the first letter is uppercase. The symbol `_' denotes an anonymous variable that is not necessary to be named in a given context (e.g., if it remains unused).

fact(a).

fact(b).

\vdots

fact(n).

\textbf{conclude} :- fact(\_).

ASP supports optimisation statements, which can be used to find optimal solutions (i.e., answer sets) based on so called weak constraints, where a cost is given by a weight assigned to a constraint. An answer set is considered optimal if its cost is minimal compared to all other possible solutions. Within the examples in this paper we use optimisation statements of the following form:

\textbf{optimise} \{ w \_@p, Var\(_1\), Var\(_2\), \ldots , atom\(_1\)(Var\(_1\), Var\(_2\)), atom\(_n\)(Var\(_n\)) \}.

where, optimise is either: \textbf{maximise} or \textbf{minimise} as applicable to \textbf{atom\(_1\)} to \textbf{atom\(_n\)} with corresponding optimisation weight \textbf{w} and priority \textbf{p}.

**Visuospatial Scene Semantics: A Context-Specific Minimal ASP Example.**

In the context of interpreting scene semantics consider the following situation, where we have two detections and two tracks and the task is to decide how we assign the detections to the tracks.

\begin{itemize}
  \item \textbf{Facts:} Detected objects and tracks are represented as facts, i.e., for detections we define facts providing an \textit{id}, \textit{type} and a \textit{confidence value}, and for tracks we define facts providing an \textit{id} and a \textit{type}.
\end{itemize}

\begin{verbatim}
% Detections det_0 and det_1
det(det_0, car, 99).
det(det_1, person, 99).

% Tracks trk_0 and trk_1
trk(trk_0, car).
trk(trk_1, person).
\end{verbatim}
• **Rules:** Further we define a rule to state that a track and a detection are of the same type, e.g., *person*, or *car*. Here the head of the rule is true if all of the atoms in the body are true, i.e., there is a track and a detection for which the type value represented by the variable *Type* is the same.

```prolog
same_type(Trk, Det) :- trk(Trk, Type), det(Det, Type, _).
```

• **Choice Rules:** Choice rules state possible heads of the rule that have to be included in the answer set. I.e., the choice rule below states that for every track *Trk* there has to be either a detection *Det* that gets assigned to the track, or the track is ended.

Additionally, the numbers before and after the brackets specify how many of the possible heads have to be included in the included, i.e., the lower and the higher limit are stated. In the case below there has to be exactly one of the heads included in the answer set.

```prolog
1{ assign(Trk, Det) : det(Det, _, _); end(Trk) }
1 :- trk(Trk, _).
```

• **Integrity Constraints:** Integrity constraints restrict the set of answers by eliminating stable models where the body is satisfied. The integrity constraint below states that it is not possible that a detection gets assigned to a track when the track and the detection do not have the same type, i.e, both are cars, or both are persons.

```prolog
:- assign(Trk, Det), not same_type(Trk, Det).
```

• **Optimization:** To assign detections *Det* to tracks *Trk*, we may maximise a matching likelihood between matched tracks and detections. This is expressed by the following maximisation statement:

```prolog
#maximize { (ML)@1, Trk, Det : assign(Trk, Det), matching_likelihood(Trk, Det, ML) }.
```

To make sure that tracks and detections are matched when ever possible, the main priority is to minimise the ending of tracks *Trk* using the following minimisation statement:

```prolog
#minimize { 1@2, Trk : end(Trk) }.
```

**Appendix B. Select answer set programming code**

Select code snippets in support of the examples in the paper are included below:

**B.1. Abduction based association**

Following the generate and test paradigm of ASP, **choice rules** are used to generate all assignments between detections and tracks to resulting on all possible assignments; assignments are tested using **integrity constraints**.

• Choice rules for generating assignment actions generate the set of assignments actions for all tracks and all detections; for example:

```prolog
1{ assign(Trk, Det) : det(Det, _, _); end(Trk); ignore_trk(Trk); halt(Trk); resume(Trk, Det) : det(Det, _, _) }
1 :- trk(Trk, Trk_Type).
```
1{ assign(Trk, Det): trk(Trk, _); start(Det); ignore_det(Det); resume(Trk, Det): trk(Trk, _) }  
:- det(Det, Det_Type, Conf).

- Generated assignments are tested based on (spatio-temporal) constraints for each assignment action. Assignments not consistent with these constraints are eliminated from the set of answers using integrity constraints:

   :- assign(Trk, Det), not assignment_constraints(Trk, Det).
   :- start(Det), not start_constraints(Det).
   :- end(Trk), not trk_state(Trk, halted).
   :- ignore_trk(Trk), not trk_state(Trk, halted).
   :- halt(Trk), not trk_state(Trk, active).
   :- resume(Trk, Det), not resume_constraints(Trk, Det).

   assignment_constraints(Trk, Det) :-
   trk(Trk, Trk_Type), det(Det, Det_Type, Conf),
   trk_state(Trk, active), match_type(Trk_Type, Det_Type),
   Conf > conf_thresh_assign,
   iou(Trk, Det, IOU), IOU > iou_thresh.

   resume_constraints(Trk, Det) :-
   trk(Trk, Trk_Type),
   det(Det, Det_Type, Conf), Conf > conf_thresh_resume,
   match_type(Trk_Type, Det_Type),
   trk_state(Trk, halted).

   start_constraints(Det) :-
   det(Det, _, Conf), Conf > conf_thresh_new_track,
   size(bigger, Det, size_threshold).

This results in the set of all possible assignments, which further gets optimized based on optimization statements in B.4.

B.2. Abducible high-level events

Event hypotheses with respect to background fluents and events are generated to explain assignment actions.

- Functional fluent visibility of a track can be fully_visible, partially_visible, or not_visible.

   fluent(visibility(Trk)) :- trk(Trk, _).

   possVal(visibility(Trk), fully_visible) :- trk(Trk, _).
   possVal(visibility(Trk), partially_visible) :- trk(Trk, _).
   possVal(visibility(Trk), not_visible) :- trk(Trk, _).

- Boolean fluent hidden_by for two tracks can be true or false.

   fluent(hidden_by(Trk1, Trk2)) :- trk(Trk1, _), trk(Trk2, _).

   possVal(hidden_by(Trk1, Trk2), true) :- trk(Trk1, _), trk(Trk2, _).
   possVal(hidden_by(Trk1, Trk2), false) :- trk(Trk1, _), trk(Trk2, _).

- Boolean fluent clipped for a track can be true or false.

   fluent(clipped(Trk)) :- trk(Trk, _).

   possVal(clipped(Trk), true) :- trk(Trk, _).
   possVal(clipped(Trk), false) :- trk(Trk, _).

- Fluents corresponding to all tracks and pairs of tracks are initialised as follows: all tracks are initialised as fully visible, not hidden by another track, and not clipped (however, note that it can be the case that events occurring with the start of a track have an effect on initialised fluent values, e.g., an event for a track starting partially occluded).
Events and causal effects are defined to describe changes in the fluents as effects of events occurring in the world. Here we show examples for the events hides_behind and missing_detections.

The event hides_behind is defined on two tracks as follows:

\[
\text{event(hides_behind(Trk1, Trk2)) :- trk(Trk1, _), trk(Trk2, _).}
\]

One object hiding behind another object causes the visibility fluent for the hidden object to change its value to not_visible. Further the fluent hidden_by for the two tracks changes its value to true.

\[
\text{causesValue(hides_behind(Trk1, Trk2), visibility(Trk1), not_visible, T) :- trk(Trk1, _), trk(Trk2, _), time(T).}
\]

\[
\text{causesValue(hides_behind(Trk1, Trk2), hidden_by(Trk1, Trk2), true, T) :- trk(Trk1, _), trk(Trk2, _), time(T).}
\]

The event missing_detections is defined on a single track as follows.

\[
\text{event(missing_detections(Trk)) :- trk(Trk, _).}
\]

\[
\text{causesValue(missing_detections(Trk), clipped(Trk), true, T) :- trk(Trk, _), time(T).}
\]

B.3. Abducing high-level events explaining assignments

Possible explanations are generated using choice rules for explaining association actions, i.e., for each association a possible explanation in terms of high-level events is generated based on preconditions and causal effects. Here we show examples for the events hides_behind and missing_detections.

• Choice rule (snippet) for explaining halted tracks:
A track can be halted because it is hiding behind another track, or there are missing detections within the track.

\[
\begin{cases}
\text{occurs_at(hides_behind(Trk, Trk2), curr_time): trk(Trk2, _);}
\text{occurs_at(missing_detections(Trk), curr_time)}
\end{cases}
\]

:- halt(Trk).

• Constraints for events are defined using integrity constraints for each event:
Integrity constraint for event hides_behind can not occur if poss(hides_behind(_, _)) is not true.

\[
:- \text{occurs_at(hides_behind(Trk1, Trk2), curr_time), not poss(hides_behind(Trk1, Trk2)).}
\]

• The event hides_behind is possible if the tracks are overlapping and both tracks visible.

\[
\text{poss(hides_behind(Trk1, Trk2)) :-}
\text{trk(Trk1, _), trk(Trk2, _),}
\text{position(overlapping_top, Trk1, Trk2),}
\text{not holds_at(visibility(Trk1), not_visible, curr_time),}
\text{not holds_at(visibility(Trk2), not_visible, curr_time).}
\]
• Integrity constraint for event missing detections.

```prolog
:- occurs_at(missing_detections(Trk), curr_time), not poss(missing_detections(Trk)).
```

• The event missing detections is possible if the track is not clipped and it is visible.

```prolog
poss(missing_detections(Trk)) :-
  holds_at(clipped(Trk), false, curr_time),
  not holds_at(visibility(Trk), not_visible, curr_time).
```

B.4. Optimisation

• Finding best fitting hypothesis on assignments and high-level events is achieved using ASP optimisation statements as follows:
  – Matching likelihood is maximised to ensure matching of best fitting detections to tracks, i.e., here maximising IoU between bounding rectangles of predicted tracks and the detections:

```prolog
matching_likelihood(Trk, Det, IOU) :- det(Det, _, _), trk(Trk, _, _), iou(Trk, Det, IOU).
```

```prolog
#maximize {(ML)@10, Trk, Det : assign(Trk, Det), matching_likelihood(Trk, Det, ML)}.
```

  – Maximising assignment of detections to tracks to avoid segmented tracks, i.e., assign detections to tracks whenever possible:

```prolog
#maximize {1@10, Trk, Det: assign(Trk, Det)}.
```

  – Resume tracks if possible; start / end tracks if resuming is not possible:

```prolog
#minimize {1@2, Trk, Det: resume(Trk, Det)}.
#minimize {5@2, Trk: end(Trk)}.
#minimize {5@2, Det: start(Det)}.
```

  – Only if no other explanation can be found, tracks and detections are ignored:

```prolog
#minimize {10@3, Det: ignore_det(Det)}.
#minimize {10@3, Trk: ignore_trk(Trk)}.
```

Appendix C. Additional examples

C.1. Occlusion example from MOT 2017 benchmark dataset

Abduced event sequence for scene 04 from the MOT 2017 benchmark, involving people moving in a crowded environment, with various occlusions (Fig. C.14).

C.2. Results for select (complete) scenes

The following are results for select scenes from the datasets being used in the evaluation (Sec 4): KITTIMOD, MOT, and safety-criticality set of scenarios developed as part of this work. For lack of space, we only choose to illustrate one select frames per sec of input stimuli:
occurs_at(enters_fov(trk_30),172)
occurs_at(hides_behind(trk_14,trk_22),188)
occurs_at(hides_behind(trk_16,trk_30),191)
occurs_at(enters_fov(trk_32),203)
occurs_at(unhides_from_behind(trk_14,trk_22),205)
occurs_at(hides_behind(trk_14,trk_20),222)
occurs_at(hides_behind(trk_8,trk_22),230)
occurs_at(unhides_from_behind(trk_16,trk_32),238)
occurs_at(unhides_from_behind(trk_14, trk_20),245)
occurs_at(unhides_from_behind(trk_8,trk_22),250)

Fig. C.14. Abduced events for scene MOT17-04 between time point 270 and time point 310.

Fig. C.15. Complete example scene from KITTIMOD tracking benchmark. High traffic highway situation including high and low speed driving.

• Fig. C.15: Scene 20 from KITTIMOD [41] tracking dataset
• Fig. C.16: Scene 02 from the MOT Challenge [58]
• Fig. C.17: Scene from safety-critical scenario dataset (Sec 4.1.2)

Appendix D. Example data

The problem specification for each time point $t$, which is the input data for the answer set programming based abduction, is generated online based on the visual stimuli; because of the size of the data (visual observations, predictions, and matching likelihood for each frame of the video) we only include a snippet for one frame to illustrate the nature of the data.
Fig. C.16. Complete example scene from MOT16 benchmark dataset for people tracking.

Fig. C.17. Tracking results for the complete scene of the occlusion example (Fig. 11; Section 4.1.2) involving tracking of cars, pedestrians, and traffic lights.
Example problem specification \(< VO_t, P_t, ML_t >=\) generated for KITTI 0020, time point 79.

```prolog
#const curr_time=79.

\(\mathcal{V}_0^{79}\) – Spatial entities of detected objects as bounding boxes:

\$\$ detections
\det(det_0,\text{car}, 99).
\det(det_1,\text{car}, 99).
\det(det_2,\text{car}, 99).
\det(det_3,\text{car}, 99).
\det(det_4,\text{car}, 95).
\det(det_5,\text{car}, 91).
\det(det_6,\text{car}, 84).
\det(det_7,\text{car}, 75).
\det(det_8,\text{car}, 72).
\det(det_9,\text{car}, 46).
\det(det_{10},\text{car}, 40).
\det(det_{11},\text{car}, 25).
\det(det_{12},\text{car}, 22).
\det(det_{13},\text{truck}, 52).
\det(det_{14},\text{truck}, 52).

\$\$ boxes for detections
\bbox2d(det_0, 0, 189, 208, 119).
\bbox2d(det_1, 697, 187, 105, 68).
\bbox2d(det_2, 220, 178, 215, 138).
\bbox2d(det_3, 401, 183, 89, 72).
\bbox2d(det_4, 640, 179, 38, 28).
\bbox2d(det_5, 520, 179, 27, 23).
\bbox2d(det_6, 473, 182, 39, 33).
\bbox2d(det_7, 588, 179, 30, 22).
\bbox2d(det_8, 494, 184, 29, 29).
\bbox2d(det_9, 557, 176, 11, 14).
\bbox2d(det_{10}, 475, 173, 28, 18).
\bbox2d(det_{11}, 422, 174, 39, 13).
\bbox2d(det_{12}, 453, 176, 24, 12).
\bbox2d(det_{13}, 586, 174, 32, 22).
\bbox2d(det_{14}, 579, 172, 22, 20).

\(\mathcal{P}_0^{79}\) – Spatial entities of predicted tracks for time-point 79 as bounding boxes:

\$\$ tracks and track states
\trk(trk_0,\text{car}).
\trk_state(trk_0, halted).
\trk(trk_1,\text{car}).
\trk_state(trk_1, active).
\trk(trk_4,\text{car}).
\trk_state(trk_4, active).
\trk(trk_5,\text{car}).
\trk_state(trk_5, active).
\trk(trk_6,\text{car}).
\trk_state(trk_6, active).
\trk(trk_7,\text{car}).
\trk_state(trk_7, active).
\trk(trk_8,\text{car}).
\trk_state(trk_9, halted).
\trk(trk_{11},\text{car}).
\trk_state(trk_{11}, active).
\trk(trk_{12},\text{car}).
\trk_state(trk_{12}, active).
\trk(trk_{13},\text{car}).
\trk_state(trk_{13}, active).

\$\$ boxes for tracks
\bbox2d(trk_0, -42, 227, 249, 159).
\bbox2d(trk_1, 698, 186, 102, 68).
\bbox2d(trk_4, 590, 179, 26, 21).
```
$\mathcal{ML}_{79}$ — Matching likelihood for pairs of tracks and detections at time point 79 given by the IoU between them.

```plaintext
% IoU for overlapping tracks and detections
iou(trk_5, det_0, 34625).
iou(trk_11, det_0, 83400).
iou(trk_1, det_1, 96879).
iou(trk_6, det_2, 71071).
iou(trk_9, det_2, 7556).
iou(trk_12, det_2, 6388).
iou(trk_6, det_3, 7333).
iou(trk_12, det_3, 30871).
iou(trk_5, det_4, 94024).
iou(trk_7, det_5, 4551).
iou(trk_13, det_5, 84391).
iou(trk_7, det_6, 30907).
iou(trk_12, det_6, 8403).
iou(trk_4, det_7, 84365).
iou(trk_7, det_8, 86273).
iou(trk_13, det_8, 1145).
iou(trk_7, det_10, 5146).
iou(trk_12, det_10, 2138).
iou(trk_12, det_11, 3087).
iou(trk_12, det_12, 2341).
iou(trk_4, det_13, 51257).
iou(trk_4, det_14, 13495).
```

Abduced Event Sequence for time point 79 (snippet for 10 time points)

```plaintext
occurs_at(missing_detections(trk_9), 69)
occurs_at(missing_detections(trk_6), 70)
occurs_at(missing_detections(trk_7), 72)
occurs_at(noise(trk_10), 73)
occurs_at(recover(trk_7), 73)
```

References


8 - Out of Sight But Not Out of Mind

Published in:
International Joint Conference on Artificial Intelligence (IJCAI 2019)

Citation:

This conference paper was published at the International Joint Conference on Artificial Intelligence (IJCAI 2019) and was nominated for the Distinguished Paper Award as one of three papers from overall 850 accepted (and 4752 submitted), received Honorable Mention. It integrates the general method for visual abduction developed in (Suchan, Bhatt, Wałęga, et al., 2018) into an online incremental visual sensemaking method, and applies it in the context of safety criticality in autonomous driving. The core technical contribution of the paper is a method for integrated high-level event abduction and low-level tracking of object motion, systematically implemented with ASP, and utilising incremental solving of the visual abduction problem to support online (real-time) execution. In essence, the method jointly tracks object motion and generates high-level declarative models explaining the resulting motion tracks in terms of spatio-temporal motion patterns and events.

The method is demonstrated and evaluated on community established real-world datasets and benchmarks in autonomous driving, KITTI and MOT (Section 2.3.3, D3 and D4), showing that integrated motion tracking and abductive reasoning with semantic characterisations of object occlusion helps to reduce common tracking errors, such as missing detections or identity switches.
Out of Sight But Not Out of Mind: An Answer Set Programming Based Online Abduction Framework for Visual Sensemaking in Autonomous Driving

Jakob Suchan\textsuperscript{1,3}, Mehul Bhatt\textsuperscript{2,3} and Srikrishna Varadarajan\textsuperscript{3}
\textsuperscript{1}University of Bremen, Germany \textsuperscript{2}Örebro University, Sweden \textsuperscript{3}CoDesign Lab / Cognitive Vision
jsuchan@uni-bremen.de, mehul.bhatt@oru.se, krish@codesign-lab.org

Abstract

We demonstrate the need and potential of systematically integrated vision and semantics solutions for visual sensemaking (in the backdrop of autonomous driving). A general method for online visual sensemaking using answer set programming is systematically formalised and fully implemented. The method integrates state of the art in visual computing, and is developed as a modular framework usable within hybrid architectures for perception & control. We evaluate and demo with community established benchmarks KITTIMOD and MOT. As 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.

1 Motivation

Autonomous driving research has received enormous academic & industrial interest in recent years. This surge has coincided with (and been driven by) advances in deep learning based computer vision research. Although deep learning based vision & control has (arguably) been successful for self-driving vehicles, we posit that there is a clear need and tremendous potential for hybrid visual sensemaking solutions (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).

Visual Sensemaking Needs Both “Vision & Semantics”

We demonstrate the significance of semantically-driven methods rooted in knowledge representation and reasoning (KR) in addressing research questions pertaining to explainability and human-centred AI particularly from the viewpoint of sensemaking of dynamic visual imagery. Consider the occlusion scenario in Fig. 1:

Car (c) is in-front; 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.

The 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 hypoth-

c beauty in Germany taking a lead in eliciting 20 key propositions (with legal implications) for the fulfillment of ethical commitments for automated and connected driving systems [BMVI, 2018].
Key Contributions. We develop a general and systematic declarative visual sensemaking method capable of online abduction: 

1. human-centric representations semantically rooted in spatio-linguistic primitives as they occur in natural language [Bhatt et al., 2013; Mani and Pustejovsky, 2012];
2. driven by Answer Set Programming (ASP) [Brewka et al., 2011], the ability to abductively compute commonsense interpretations and explanations in a range of (a)typical everyday driving situations, e.g., concerning safety-critical decision-making; 
3. online performance of the overall framework modularly integrating high-level commonsense reasoning and state of the art low-level visual computing for practical application in real-world settings. We present the formal framework & its implementation, and demo & empirically evaluate with community established real-world datasets and benchmarks, namely: KIT-TIMOD [Geiger et al., 2012] and MOT [Milan et al., 2016].

2 Visual Sensemaking: A General Method Driven by ASP

Our proposed framework, in essence, jointly solves the problem of assignment of detections to tracks and explaining overall scene dynamics (e.g. appearance, disappearance) in terms of high-level events within an online integrated low-level visual computing and high level abductive reasoning framework (Fig. 2). Rooted in answer set programming, the framework is general, modular, and designed for integration as a reasoning engine within (hybrid) architectures designed for real-time decision-making and control where visual perception is needed as one of the several components. In such large scale AI systems the declarative model of the scene dynamics resulting from the presented framework can be used for semantic QA, inference etc. to support decision-making.

2.1 Space, Motion, Objects, Events

Reasoning about dynamics is based on high-level representations of objects and their respective motion & mutual interactions in spacetime. Ontological primitives for commonsense reasoning about spacetime ($\Sigma_{dyn}$) and dynamics ($\Sigma_{dyn}^*$) are:

- $\Sigma_{dyn}$: domain-objects $O = \{o_1, ..., o_n\}$ represent the visual elements in the scene, e.g., people, cars, cyclists; elements in $O$ are geometrically interpreted as spatial entities $E = \{e_1, ..., e_n\}$; spatial entities $E$ may be regarded as points, line-segments or (axis-aligned) rectangles based on their spatial properties (and a particular reasoning task at hand). The temporal dimension is represented by time points $T = \{t_1, ..., t_n\}$. $MT_{ts} = (e_{ts_1}, ..., e_{ts_n})$ represents the motion track of a single object $o_s$ where $t_s$ denotes the start and end time of the track and $e_{ts_i}$ denotes the spatial entity (E) —e.g., the axis-aligned bounding box—corresponding to the object $o_i$ at time points $t_s$ to $t_e$. The spatial configuration of the scene and changes within it are characterised based on the qualitative spatio-temporal relationships (R) between the domain objects. For the running and demo examples of this paper, positional relations on axis-aligned rectangles based on the rectangle algebra (RA) [Balbiani et al., 1999] suffice; RA uses the relations of Interval Algebra (IA) [Allen, 1983] $\mathcal{RA} = \{\text{before, after, during, contains, starts, started by, finishes, finished by, overlaps, overlapped by, meets, met by, equal}\}$ to relate two objects by the interval relations projected along each dimension separately (e.g., horizontal and vertical dimensions).
- $\Sigma_{fl}^*$: the set of fluents $\Phi = \{\phi_1, ..., \phi_n\}$ and events $\Theta = \{\theta_1, ..., \theta_n\}$ respectively characterise the dynamic properties of the objects in the scene and high-level abducibles (Table
1. For reasoning about dynamics (with $\langle \Phi, \Theta \rangle$), we use a variant of event calculus as per [Ma et al., 2014; Miller et al., 2013]; in particular, for examples of this paper, the functional event calculus fragment ($\Sigma_{\text{dyn}}$) of Ma et al. [2014] suffices:

$\Sigma \equiv_{\text{def}} \Sigma_{\text{dyn}} < \Phi, \Theta > \cup \Sigma_{\text{st}} < \mathcal{O}, \mathcal{T}, \mathcal{M}, R >$

2.2 Tracking as Abduction

Scene dynamics are tracked using a detect and track approach: we tightly integrate low-level visual computing (for detecting scene elements) with high-level ASP-based abduction to solve the assignment of observations to object tracks in an incremental manner. For each time point $t$ we generate a problem specification consisting of the object tracks and visual observations and use ASP to abductively solve the corresponding assignment problem incorporating the ontological structure of the domain / data (abstracted with $\Sigma$). Steps 1–3 (Alg. 1 & Fig. 3) are as follows:

Step 1. Formulating the Problem Specification

The ASP problem specification for each time point $t$ is given by the tuple $\langle \mathcal{V} \mathcal{O}_t, \mathcal{P}_t, \mathcal{M}_t, \mathcal{L}_t \rangle$ and the sequence of events ($\mathcal{H}_{\text{events}}$) before time point $t$.

- **Visual Observations** Scene elements derived directly from the visual input data are represented as spatial entities $\mathcal{E}$, i.e., $\mathcal{V} \mathcal{O}_t = \{ e_{\text{obs}1}, ..., e_{\text{obs}n} \}$ is the set of observations at time $t$ (Fig. 3). For the examples and empirical evaluation in this paper (Sec. 3) we focus on Obstacle / Object Detections – detecting cars, pedestrians, cyclists, traffic lights etc using YOLOv3 [Redmon and Farhadi, 2018]. Further we generate scene context using Semantic Segmentation – segmenting the road, sidewalk, buildings, cars, people, trees, etc. using DeepLabv3+ [Chen et al., 2018] and Lane Detection – estimating lane markings, to detect lanes on the road, using SCNN [Pan et al., 2018]. Type and confidence score for each observation is given by $\mathcal{V} \mathcal{P}_t$, and $\mathcal{C} \mathcal{O}_t$.

- **Movement Prediction** For each track $trk_i$, changes in position and size are predicted using kalman filters; this results in an estimate of the spatial entity $e$ for the next time point $t$ of each motion track $\mathcal{P}_t = \{ e_{\text{trk}1}, ..., e_{\text{trk}n} \}$.

- **Matching Likelihood** For each pair of tracks and observations $e_{\text{trk}}$, and $e_{\text{obs}}$, where $e_{\text{trk}} \in \mathcal{P}_t$ and $e_{\text{obs}} \in \mathcal{V} \mathcal{O}_t$, we compute the likelihood $\mathcal{M}_t = \{ ml_{\text{trk}, e_{\text{obs}}}, ..., ml_{\text{trk}, e_{\text{obs}}} \}$ that $e_{\text{obs}}$ belongs to $e_{\text{trk}}$. The intersection over union (IoU) provides a measure for the amount of overlap between the spatial entities $e_{\text{obs}}$ and $e_{\text{trk}}$.

Step 2. Abduction based Association

Following perception as logical abduction most directly in the sense of Shanahan [2005], we define the task of abducting visual explanations as finding an association ($\mathcal{H}_{\text{assign}}$) of observed scene
elements ($\mathcal{VO}_t$) to the motion tracks of objects ($\mathcal{MT}$) given by the predictions $P_t$, together with a high-level explanation ($H_{events}^t$), such that $H_{events}^t \cap H_{assign}^t$ is consistent with the background knowledge and the previously abduced event sequence $H_{events}^{t-1}$, and entails the perceived scene given by $< \mathcal{VO}_t, P_t, \mathcal{ML}_t >$.

$$\Sigma \land H_{events}^t \land [H_{assign}^t \land H_{events}^t] = \mathcal{VO}_t \land P_t \land \mathcal{ML}_t$$

where $H_{assign}^t$ consists of the answers of detections to object tracks, and $H_{events}^t$ consists of the high-level events $\Theta$ explaining the assignments.

- **Associating Objects and Observations** Finding the best match between observations ($\mathcal{VO}_t$) and object tracks ($P_t$) is done by generating all possible assignments and then maximising a matching likelihood $\mathcal{ML}_{trk, obs}$ between pairs of spatial entities for matched observations $\mathcal{OB}_t$ and predicted track region $\mathcal{TRK}_t$ (See Step 3). Towards this we use choice rules [Gebser et al., 2014] (i.e., one of the heads of the rule has to be in the stable model) for $\mathcal{BLU}_t$ and $\mathcal{IRK}_t$, generating all possible assignments in terms of assignment actions: assign, start, end, halt, resume, ignore_det, ignore_trk.

For each assignment action we define integrity constraints that restrict the set of answers generated by the choice rules, e.g., the following constraints are applied to assigning an observation $\mathcal{OB}_t$ to a track $\mathcal{TRK}_t$, applying thresholds on the IOU $\mathcal{TRK}_t$, $\mathcal{OB}_t$, and the confidence of the observation $\mathcal{BLO}_t$. Further we define the assignment of an object to match the type of the track it is assigned to:

$$\Sigma \land H_{events}^t \land [H_{assign}^t \land H_{events}^t] = \mathcal{VO}_t \land P_t \land \mathcal{ML}_t$$

that have to hold for the event to possibly occur, and the effects the occurrence of the event has on the properties of the objects, i.e., the value of the fluency fluent changes to fully_occluded.

Table 1: Events and Fluents Explaining (Dis)Appearance

<table>
<thead>
<tr>
<th>EVENTS</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>hides_behind(Trk1, Trk2)</td>
<td>Trk1 hides behind Trk2.</td>
</tr>
<tr>
<td>unhides_from_behind(Trk1, Trk2)</td>
<td>Trk1 unhides from behind Trk2.</td>
</tr>
<tr>
<td>missing_detections(Trk)</td>
<td>Trk's missing detections in Trk.</td>
</tr>
</tbody>
</table>

Step 3: Finding the Optimal Hypothesis To ensure an optimal assignment, we use ASP based optimization to maximize the matching likelihood between matched pairs of tracks and detections. Towards this, we first define the matching likelihood based on the Intersection over Union (IoU) between the observations and the predicted boxes for each track as described in [Bewley et al., 2016]:

$$\text{maximize} \{ \text{Prob} \} _t \text{Trk}, \text{Det}, \text{Det}: \text{assign} \text{Det}, \text{assign} \text{Prob} \text{Det}, \text{Prob} \}.$$

We then maximize the matching likelihood for all assignments, using the build in maximise statement:

$$\text{maximize} \{ \text{Prob} \} _t \text{Trk}, \text{Det}, \text{Det}: \text{assign} \text{Det}, \text{assign} \text{Prob} \text{Det}, \text{Prob} \}.$$

To find the best set of hypotheses with respect to the observations, we minimize the occurrence of certain events and association actions, e.g., the following optimization statements minimize starting and ending tracks; the resulting assignment is then used to update the motion tracks accordingly.

$$\text{minimize} \{ \text{Start} \} _t \text{Trk}, \text{Det}: \text{end} \text{Det} \}.$$

It is important here to note that: (1) by jointly abducting the object dynamics and high-level events we can impose constraints on the assignment of detections to tracks, i.e., an assignment is only possible if we can find an explanation supporting the assignment; and (2) the likelihood that an event
occurs guides the assignments of observations to tracks. Instead of independently tracking objects and interpreting the interactions, this yields to event sequences that are consistent with the abduced object tracks, and noise in the observations is reduced (See evaluation in Sec. 3).

3 Application & Evaluation

We demonstrate applicability towards identifying and interpreting safety-critical situations (e.g., Table 2); these encompass those scenarios where interpretation of spacetime dynamics, driving behaviour, environmental characteristics is necessary to anticipate and avoid potential dangers.

Reasoning about Hidden Entities

Consider the situation of Fig. 4: a car gets occluded by another car turning left and reappears in front of the autonomous vehicle. Using online abduction for abducting high-level interactions of scene objects we can hypothesize that the car got occluded and anticipate its reappearance based on the perceived scene dynamics. The following shows data and abduction events.

\[
\begin{align*}
\text{trk(trk}_3, \text{ car). trk_state(trk}_3, \text{ active). ...} \\
\text{trk(trk}_41, \text{ car). trk_state(trk}_41, \text{ active). ...} \\
\text{box2d(trk}_3, 600, 460, 136, 302). ... \\
\text{box2d(trk}_41, 631, 471, 40, 47). ... \\
\text{occurs_at(hides_behind(trk}_1, trk}_2), 179)) ...
\end{align*}
\]

We define a rule stating that a hidden object may unhide from behind the object it is hidden by and anticipate the time point \( t \) based on the object movement as follows:

\[
\text{anticipate(unhides_from_behind(trk}_1, trk}_2, t) :- \text{time(t), curr_time < t, ...} \\
\text{holds_at(hidden_by(trk}_1, trk}_2, \text{ curr_time), ...} \\
\text{topology(proper_part, trk}_1, \text{ trk}_2), ... \\
\text{movement(moves_out_of, trk}_1, \text{ trk}_2, t).
\]

We then interpolate the objects position at time point \( t \) to predict where the object may reappear.

\[
\text{point2d(interpolated_position(Trk}, T), \text{ PosX, PosY) :-} \\
\text{time(T), curr_time < T, ...} \\
\text{box2d(Trk}, X, Y, W, H), \text{ trk_mov(Trk}, \text{ MovX, MovY), ...} \\
\text{PosX = X+MovX*T, PosY = Y+MovY*T).
\]

For the occluded car in our example we get the following prediction for time \( t \) and position \( x, y \):

\[
\text{anticipate(unhides_from_behind(trk}_41, trk}_3, 202) \\
\text{point2d(interpolated_position(trk}_41, 202), 738, 495)
\]

Based on this prediction we can then define a rule that gives a warning if a hidden entity may reappear in front of the vehicle, which could be used by the control mechanism, e.g., to adapt driving and slow down in order to keep safe distance:

\[
\text{warning(hidden_entity_in_front(Trk}_1, T) :-} \\
\text{time(T), curr_time < anticipation_threshold, ...} \\
\text{position(in_front, interpolated_pos(Trk}_1, T).}
\]

Table 2: Safety-Critical Situations

<table>
<thead>
<tr>
<th>Situation</th>
<th>Objects</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIDDEN_ENTITY</td>
<td>entity, object</td>
<td>traffic participant hidden by obstacle</td>
</tr>
<tr>
<td>REDUCED_VISIBILITY</td>
<td>object</td>
<td>visibility reduced by object in front.</td>
</tr>
<tr>
<td>SUDEN_STOP</td>
<td>vehical</td>
<td>vehicle in front stopping suddenly</td>
</tr>
<tr>
<td>BLOCKED_LANE</td>
<td>lane, object</td>
<td>lane of the road is blocked by an object.</td>
</tr>
</tbody>
</table>

Empirical Evaluation

For online sensemaking, evaluation focusses on accuracy of abduced motion tracks, real-time performance, and the tradeoff between performance and accuracy. Our evaluation uses the KITTI object tracking dataset [Geiger et al., 2012], which is a community established benchmark dataset for autonomous cars: it consists of 21 training and 29 test scenes, and provides accurate track annotations for 8 object classes (e.g., car, pedestrian, van, cyclist). We also evaluate tracking results using the more general cross-domain Multi-Object Tracking (MOT) dataset [Milan et al., 2016] established as part of the MOT Challenge; it consists of 7 training and 7 test scenes which are highly unconstrained videos filmed with both static and moving cameras. We evaluate on the available groundtruth for training scenes of both KITTI using YOLOv3 detections, and MOT17 using the provided faster RCNN detections.

- **Evaluating Object Tracking** For evaluating accuracy (MOTA) and precision (MOTP) of abduced object tracks we follow the Clear MOT [Bernardin and Stiefelhagen, 2008] evaluation schema. Results (Table 3) show that jointly abducting high-level object interactions together with low-level scene dynamics increases the accuracy of the object tracks, i.e., we consistently observe an improvement of about 5%, from 45.72% to 50.5% for cars and 28.71% to 32.57% for pedestrians on KITTI, and from 41.4% to 46.2% on MOT.

- **Online Performance and Scalability** Performance of online abduction is evaluated with respect to its real-time capabilities.\(^4\) (1). We compare the time & accuracy of online abduction for state of the art (real-time) detection methods: YOLOv3, SSD [Liu et al., 2016], and Faster RCNN [Ren et al., 2015] (Fig. 5). (2). We evaluate scalability of the ASP based abduction on a synthetic dataset with controlled number of tracks and % of overlapping tracks per frame. Results (Fig. 5) show that online abduction can perform with above 30 frames per second for scenes with up to 10 highly overlapping object tracks, and more than 50 tracks with 1fps (for the sake of testing, it is worth noting that even for 100 objects per frame it only takes about an average of 4 secs per frame). Importantly, for realistic scenes such as in the KITTI dataset, abduction runs realtime at 33.9fps using YOLOv3, and 46.7 using SSD with lower accuracy but providing good precision.

Discussion of Empirical Results

Results show that integrating high-level abduction and object tracking improves the

---

\(^4\)Evaluation using a dedicated Intel Core i7-6850K 3.6GHz 6-Core Processor, 64GB RAM, and a NVIDIA Titan V GPU 12GB.
resulting object tracks and reduce the noise in the visual observations. For the case of online visual sense-making, ASP based abduction provides the required performance: even though the complexity of ASP based abduction increases quickly, with large numbers of tracked objects the framework can track up to 20 objects simultaneously with 30 fps and achieve real-time performance on the KITTI benchmark dataset. It is also important to note that the tracking approach in this paper is based on tracking by detection using a naive measure, i.e., the IoU (Sec. 2.2; Step 1), to associate observations and tracks, and it is not using any visual information in the prediction or association step. Naturally, this results in a lower accuracy, in particular when used with noisy detections and when tracking fast moving objects in a benchmark dataset such as KITTI. That said, due to the modularity of the implemented framework, extensions with different methods for predicting motion (e.g., using particle filters or optical flow based prediction) are straightforward: i.e., improving tracking is not the aim of our research.

4 Related Work

Answer Set Programming is now widely used as an underlying knowledge representation language and robust methodology for non-monotonic reasoning [Brewka et al., 2011; Gebser et al., 2012]. With ASP as a foundation, and driven by semantics, commonsense and explainability [Davis and Marcus, 2015], this research aims to bridge the gap between high-level formalisms for logical abduction and low-level visual processing by tightly integrating semantic abstractions of space-change with their underlying numerical representations. Within KR, the significance of high-level (abductive) explanations in a range of contexts is well-established: planning & process recognition [Kautz, 1991], vision & abduction [Shanahan, 2005], probabilistic abduction [Blythe et al., 2011], reasoning about spatio-temporal dynamics [Bhatt and Loke, 2008], reasoning about continuous spacetime change [Muller, 1998; Hazarika and Cohn, 2002] etc. Dubba et al. [2015] uses abductive reasoning in an inductive-abductive loop within inductive logic programming (ILP). Aditya et al. [2015] formalise general rules for image interpretation with ASP. Similarly motivated to this research is [Suchan et al., 2018], which uses a two-step approach (with one huge problem specification), first tracking and then explaining (and fixing) tracking errors; such an approach is not runtime / real-time capable. In computer vision research there has recently been an interest to synergise with cognitively motivated methods; in particular, e.g., for perceptual grounding and inference [Yu et al., 2015], and combining video analysis with textual information for understanding events and answering queries about video data [Yu et al., 2014].

5 Conclusion & Outlook

We develop a novel abduction-driven online (i.e., realtime, incremental) visual sensemaking framework: general, systematically formalised, modular and fully implemented. Integrating robust state-of-the-art methods in knowledge representation and computer vision, the framework has been evaluated and demonstrated with established benchmarks. We highlight application prospects of semantic vision for autonomous driving, a domain of emerging and long-term significance. Specialised commonsense theories (e.g., about multi-sensory integration and multi-agent belief merging, contextual knowledge) may be incorporated based on requirements. Our ongoing focus is to develop a novel dataset emphasising semantics and (commonsense) explainability; this is driven by mixed-methods research – AI, Psychology, HCI – for the study of driving behaviour in low-speed, complex urban environments with unstructured traffic. Such interdisciplinary studies are needed to better appreciate the complexity and spectrum of varied human-centred challenges in autonomous driving, and demonstrate the significance of integrated vision and semantics solutions in those contexts.

Acknowledgements

Partial funding by the German Research Foundation (DFG) via the CRC 1320 EASE – Everyday Activity Science and Engineering” (www.ease-crc.org), Project P3: Spatial Reasoning in Everyday Activity is acknowledged.
References


Chapter 5
Learning Explainable Visuospatial Models

We develop a framework for computing declarative interpretation models of image characteristics; such declarative models provide the basis for inductive learning of relational visuospatial structure. We summarise how these declarative models of image characteristics are modelled based on symbolic and subsymbolic image elements and features. Further, we describe how this forms the basis for inductive learning, and discuss neurosymbolic aspects of learning visuospatial characteristics.

Included Publications:

5.1. Integrating Declarative Explainability and Inductive Generalization

Learning high-level concepts from visuospatial imagery is a central ability of artificially intelligent perception systems. Towards this visual perception systems have to be able to represent and reason about conceptual symbolic commonsense knowledge (i.e., pertaining to space, time, motion, and events) and ground these concepts with respect to sub-symbolic (neural) visual features of involved image elements. Such neurosymbolic reasoning and sensemaking may involve systematic integration of KR based semantics with powerful (deep learning based) methods for visual feature learning. Declarative explainability here refers to the ability to declaratively interpret visual imagery with respect to high-level visuospatial concepts and underlying visual features (as detailed in Section 2.1.2). 

Inductive Logic Programming (ILP) (Muggleton and Raedt, 1994) has evolved as a powerful mechanism to learn high-level relational concepts based on inductive generalisation and has produced various extensions to account for real-world conditions (i.e. noisy and incomplete), including probabilistic inductive learning (Kimmig, De Raedt, and Toivonen, 2007; Sato and Kameya, 2008) and inductive-abductive learning of spatial structures (K. Dubba et al., 2012; K. S. R. Dubba et al., 2015). However, learning complex visuospatial relational structures from naturalistic visual stimuli remains challenging. In this context, we argue that explainability driven declarative interpretation models of deep visuospatial semantics grounded in geometric representations and visual features of spatial entities provide a suitable abstraction layer for inductive learning of visuospatial structure directly from low-level (dynamic) visual imagery. As an example use-case we apply these models to analyse symmetrical structure in naturalistic images and learn subjective measures of symmetry perception. Furthermore, we show how inductive generalisation may be used to learn relational spatio-temporal rules of (visuospatial) symmetry.

Declarative Models of Visuospatial Characteristics. Interpreting visuospatial characteristics of an image is based on symbolic and sub-symbolic image elements, e.g., objects and people in the image, but also abstract regions and areas of interest with respect to the particular domain at hand, and (neural) visual features of these image elements, emanating directly from colour, texture, light, and shadow. Within the theory of space and motion (Σ, Section 2.2.2) the image elements are represented as primitive spatial entities $E$, and grounded in neural image features represented by feature vectors obtained from Deep Neural Networks (DNN). Visuospatial characteristics of an image may then be modelled based on the spatio-temporal configuration of image elements combined with visual features of these elements, e.g., similarity of image elements in symmetrical positions in the image. For interpretation and explanation with these models we incorporate distance measures for spatial configuration and neural features, providing a scale for the specific visuospatial characteristic. These distances involve spatial configuration, i.e. how close is the perceived configuration to the model, e.g., in size, position, orientation, of elements, and the neural visual features of these el-
5.1. Integrating Declarative Explainability and Inductive Generalization

Figure 5.1: Analysing Symmetry: (a) symmetrical, and (b) non-symmetrical elements of an image.

Credits: Figure adapted from (Suchan, Bhatt, Vardarajan, et al., 2018); Stills from the movie “Skyfall”, directed by Sam Mendes, produced by Michael G. Wilson and Barbara Broccoli, Eon Productions, Metro-Goldwyn-Mayer Pictures, and Columbia Pictures, UK and USA, 2012

Example 8. A MULTI-LAYER MODEL OF SYMMETRY

We develop a multi-level characterisation for analysing (reflectional) symmetry, encompassing three layers (L1–L3, as defined in (Suchan, Bhatt, Vardarajan, et al., 2018)):

L1. Symmetrical (spatial) composition: Spatial arrangement of objects in the scene with respect to a structural representation of position, size, orientation, etc.;

L2. Perceptual similarity: Perceptual similarity of features in symmetrical image patches, based on the low-level feature based appearance of objects, e.g., color, shape, patterns, etc.;

L3. Semantic similarity: Similarity of semantic categories of the objects in symmetrical image patches, e.g., people, object types, and properties of these objects, such as peoples gazing direction, foreground / background etc.

This characterisation may serve as the foundation for analysing and interpreting symmetrical structures in the images; in particular it can be used to identify the elements of the image supporting the symmetrical structure, and those not in line with the symmetry, e.g., elements breaking the symmetry (Figure 5.1).

ements. The model can then be used to declaratively inspect image characteristics and to learn weights on the importance of specific parts of the model, e.g., for learning subjective measures of symmetry.

Consider the characterisation of symmetrical image structure alluded to in Example 8. In (Suchan, Bhatt, Vardarajan, et al., 2018) this characterisation is used to train machine learning models, using subjective symmetry data obtained from a large scale empirical symmetry study conducted in the context of the Symmetry Dataset (Section 2.3.3, D2). In particular, we train a classifier and a regressor to predict the symmetry class (not_symmetric, somewhat_symmetric, symmetric, and highly_symmetric) and a value for the average symmetry of an image.
Example 9. SYMMETRICAL IMAGE STRUCTURE

Symmetrical spatial structure (as depicted in Figure 5.2) may be inductively learned by finding spatio-temporal relations ($R$) holding between basic spatial entities ($E$) representing the image elements.

Based on the ILP system Aleph (Srinivasan, 2001), the learning domain is defined using the predicates $modeh$, for specifying the head of the rule, and $modeb$ to specify the predicates used in the body of the rule, i.e., including relations on topology, orientation, distance, and size as follows:

\[
\begin{align*}
: & \quad modeh(1, \text{symmetric}(+img)). \\
: & \quad modeb(*, \text{entity}(\#ent,-obj,+img)). \\
: & \quad modeb(*, \text{topology}(\text{rcc8}(\#rel),+obj,+obj)). \\
: & \quad modeb(*, \text{distance}(\#rel,+obj,+obj)). \\
: & \quad modeb(*, \text{orientation}(\#rel,+obj,+obj)). \\
: & \quad modeb(*, \text{size}(\#rel,+obj,+obj)).
\end{align*}
\]

Rules about symmetrical spatial configuration are learned based on positive and negative examples. E.g., when trained on examples where the object level symmetry is determined by two people being equally far away from the symmetry axis, the system learns the following spatial structure:

\[
\text{sym}(A) :\text{entity(center(person(0)),B,A)}, \text{entity(center(person(1)),C,A)}, \text{entity(symmetry_object(center_axis),D,A)}, \text{distance(equidistant,D,C,B)}.
\]

**Inductive Visuospatial Generalisation.** In the context of (inductive) generalisation and learning, such interpretation models may constitute the basis for learning relational visuospatial structures and rules, by integrating the declarative characterisations of space and motion, as available from $\Sigma$, into the general inductive learning framework (Muggleton and Raedt, 1994), which is defined as follows:

**General ILP Setup:** Given a set of positive and negative examples, i.e., $E = E^+ \cup E^-$ and some background knowledge $B$, the learning task is to find hypotheses $H$ consisting of logic rules, such that the hypotheses $H$ together with the background knowledge $B$
entail the positive examples but do not entail the negative examples, i.e., $H \cup B \models E^+$, and $H \cup B \not\models E^-$.

For learning rules on visuospatial characteristics, e.g., symmetrical configuration of objects (as detailed in Example 9), the learning domain, as described above, is defined by spatio-temporal relations ($R$) holding between basic spatial entities ($E$) representing the image elements, and visual features of these elements. In particular, learning examples are abstracted, based on their underlying spatio-temporal structure, and inducible hypotheses ($H$) consist of rules regarding this visuospatial structure.

**Towards Integrated Neurosymbolic Explainability and Learning.** Explainability driven declarative interpretation models as defined above can be directly integrated into the inductive learning setup and can thus serve as a basis for inductive generalisation in the backdrop of the theory of space and motion, and to learn image characteristics directly from images in an integrated neurosymbolic manner. In particular, the incremental learning process of (neural) visual features 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), and facilitate (neurosymbolic) visuospatial structure learning (Bhatt, Suchan, and Varadarajan, 2019).

**INCLUDED PUBLICATIONS:** Declarative explainability of image characteristics and inductive spatio-temporal learning for the case of reflectional symmetry are the focus of the following two publications:

- A journal publication (Suchan, Bhatt, Vardarajan, et al., 2018, 9) published in the *Advances in Cognitive Systems Journal* (ACS), developing a general method to declaratively model visuospatial image characteristics, based on qualitative relational spatio-temporal structure between image elements and visual (neural) features of these elements; and


*Copies of the above stated publications in the given order follow in the next sections.*
This journal article was published in the *Advances in Cognitive Systems Journal* (ACS). It presents a general (neurosymbolic) framework to declaratively model image characteristics, based on qualitative visuospacial relational structure between image elements and visual (neural) features of these elements. The main technical focus is on explainable interpretation models for the analysis of visuospatial image characteristics, based on relational characterisations of spatio-temporal structure holding between image elements (using spatio-temporal relations as available from the theory about space and motion $\Sigma$, Section 2.2.2) and visual appearance of these image elements represented by feature vectors extracted from DNN. As such these interpretation models support human-centred deep semantic question-answering and explanation with visuospatial image characteristics.

Application of the framework focuses on the case of (reflectional) symmetry in naturalistic scenes. Here, human-centred declarative characterisations of visual symmetry are demonstrated to learn subjective symmetry measures based on a qualitative study with human subjects, whereby human subjects rank their subjective perception of visual symmetry for a set of images using (qualitative) distinctions. In this context, the proposed characterisation serves as the foundation for analysing and interpreting symmetrical structures in the images; in particular it can be used to identify the elements of the image supporting the symmetrical structure, but also those parts of the image that are not in line with the formal characterisation of symmetry, e.g., elements breaking the visual symmetry.
Semantic Analysis of (Reflectional) Visual Symmetry
A Human-Centred Computational Model for Declarative Explainability

Jakob Suchan JSUCHAN@UNI-BREMEN.DE
Mehul Bhatt MEHUL.BHATT@ORU.SE
Srikrishna Varadarajan KRISH@CODESIGN-LAB.ORG
CoDesign Lab EU / Cognitive Vision
www.codesign-lab.org / www.cognitive-vision.org
Örebro University, Sweden / University of Bremen, Germany

Seyed Ali Amirshahi S.ALI.AMIRSHAHI@NTNU.NO
Norwegian Colour and Visual Computing Laboratory
Norwegian University of Science and Technology, Norway

Stella Yu STELLAYU@BERKELEY.EDU
ICSI Vision Group, UC Berkeley, United States

Abstract
We present a computational model for the semantic interpretation of symmetry in naturalistic scenes. Key features include a human-centred representation, and a declarative, explainable interpretation model supporting deep semantic question-answering founded on an integration of methods in knowledge representation and deep learning based computer vision. In the backdrop of the visual arts, we showcase the framework’s capability to generate human-centred, queryable, relational structures, also evaluating the framework with an empirical study on the human perception of visual symmetry. Our framework represents and is driven by the application of foundational, integrated Vision and Knowledge Representation and Reasoning methods for applications in the arts, and the psychological and social sciences.

1. Introduction
Visual symmetry as an aesthetic and stylistic device has been employed by artists across a spectrum of creative endeavours concerned with visual imagery in some form, e.g., painting, photography, architecture, film and media design. Symmetry in (visual) art and beyond is often linked with elegance, beauty, and is associated with attributes such as being well-proportioned and well-balanced (Weyl, 1952). Closer to the "visual imagery” and “aesthetics” centred scope of this paper, symmetry...
Figure 1: The perception of symmetry. (a) Symmetry perception influenced by visual features, conceptual categories, semantic layering, and nuances of individual differences in perception, and (b) examples for symmetry in visual arts: “Delivery of the Keys” (ca.1481) by Perugino, “The Last Supper” (1495-98) by Leonardo Da Vinci, “View of the grand staircase at La Rinascente in Rome, designed by Franco Albini and Franca Helg” (1962) by Giorgio Casali, and “The Matrix” (1999) by The Wachowski Brothers.

has been employed by visual artists going back to the masters Giorgione, Titian, Raphael, da Vinci, and continuing till the modern times with Dali and other contemporary artists (Figure 1).

**Visual Symmetry: Perception and Semantic Interpretation**

There exist at least four closely related points of view pertaining to symmetry, namely, the physical, mathematical, psychological, and aesthetical points of view (Molnar & Molnar, 1986). As Molnar & Molnar (1986) articulate:

“But perceptual symmetry is not always identical to the symmetry defined by the mathematicians. A symmetrical picture is not necessarily symmetrical in the mathematical sense...Since the aesthetical point of view is strictly linked to the perceptive system, in examining the problems of aesthetics we find ourselves dealing with two distinct groups of problems: (1) the problem of the perception of symmetry; (2) the aesthetical effect of the perception of a symmetrical pattern.”

Indeed, the high-level semantic interpretation of symmetry in naturalistic visual stimuli by humans is a multi-layered perceptual phenomena operating at several interconnected cognitive levels involving, e.g., spatial organisation, visual features, semantic layers, individual differences (Section 2.1; and Figure 1a). Consider the select examples from movie scenes in Figure 2:

- in the shot from “2001: A Space Odyssey” (Figure 2a) a centre-perspective is being applied for staging the scene. The symmetry here is obtained by this, as well as by the layout of the room, the placement of the furniture, and the decoration of the room. In particular, the
black obelisk in the centre of the frame is emphasising the centre-perspective regularly used by Kubrick, with the bed (and person) being positioned directly on the central axis.

- Wes Anderson is staging his shot from "The Royal Tenenbaums" (Figure 2b) around a central point, but unlike Kubrick's shot, Anderson focuses on the people involved in it. Even though the visual appearance of the characters differs a lot, the spatial arrangement and the semantic similarity of the objects in the shot creates symmetry. Furthermore, the gazing direction of the characters, i.e., people on the right facing left and people on the left facing right, adds to the symmetrical appearance of the shot.

- In "The Big Lebowski" (Figure 2c), Joel and Ethan Coen use symmetry to highlight the surreal character of a dream sequence; the shot in Figure 2c uses radial symmetry composed of a group of dancers, shot from above, moving around the centre of the frame in a circular motion. This is characterised by moving entities along a circular path and centre-point, and the perceptual similarity in the appearance of the dancers.

The development of computational cognitive models focussing on a human-centred –semantic, explainable– interpretation of visuo-spatial symmetry presents a formidable research challenge demanding an interdisciplinary—mixed-methods—approach at the interface of cognitive science, vision & AI, and visual perception focussed human-behavioural research. Broadly, our research is driven by addressing this interdisciplinarity, with an emphasis on developing integrated KR-and-vision foundations for applications in the psychological and social sciences, e.g., archival, automatic annotation and pre-processing for qualitative analysis, studies in visual perception.

Key Contributions The core focus of the paper is to present a computational model with the capability to generate semantic, explainable interpretation models for the analysis of visuo-spatial symmetry. The explainability is founded on a domain-independent, mixed qualitative-quantitive representation of visuo-spatial relations based on which the symmetry is declaratively characterised. We also report on a qualitative evaluation with human-subjects, whereby human subjects rank their subjective perception of visual symmetry for a set stimuli using (qualitative) distinctions. The broader implications are two-fold: (1) the paper demonstrates the integration of vision and semantics, i.e., knowledge representation and reasoning methods with low-level (deep learning based) visual processing methods; and (2). from an applied viewpoint, the developed methodology can serve as the...
technical backbone for assistive and analytical technologies for visual media studies, e.g., from the viewpoint of psychology, aesthetics, cultural heritage.

2. The Semantics of Symmetry

Symmetry in visual imagery denotes that an image is invariant to certain types of transformation of the image, e.g., reflectional symmetry is the case where the image does not change, when it is mirrored along a specific symmetry-axis. Besides reflectional symmetry, there are various types of symmetry, including rotational symmetry, and translational symmetry. Perfect symmetry can be easily detected based on image level features, by comparing pixel in the image; however, in natural images, e.g., coming from the visual arts, perfect symmetry is a very rare case and mostly variations of symmetry are used as a stylistic device, with it being present only in some parts of the image. To address this, we focus on developing a semantic model capable of interpreting symmetrical structures in images.

2.1 A Multi-Level Semantic Characterisation

From the viewpoint of perceptual and aesthetic considerations, key aspects for interpreting visual-spatial symmetry (in scope of the approach of this paper) include (S1–S4; Figure 1):

(S1) **Spatial organisation**: High-level conceptual categories identifiable from geometric constructions by way of arbitrary shapes, relative orientation and placement, size of geometric entities, relative distance, and depth

(S2) **Visual features**: Low-level visual features and artefacts emanating directly from color, texture, light, and shadow

(S3) **Semantic layers**: Semantic-spatial layering and grouping based on natural scene characteristics involving, for instance, establishing foreground-background, clustering based on conceptual similarity, relative distance, and perceived depth, and application of commonsense knowledge possibly not directly available in the stimulus
**Individual differences:** Grounding of the visual features in the socio-cultural semiotic landscape of the perceiver (i.e., contextual and individualised nuances in perception and sensemaking).

We develop a multi-level characterisation of symmetry aimed at analysing (reflectional) symmetry. Visual symmetry—in this paper—encompasses three layers (L1–L3; Figure 3):

1. **Symmetrical (spatial) composition:** Spatial arrangement of objects in the scene with respect to a structural representation of a wrt. position, size, orientation, etc.;
2. **Perceptual similarity:** Perceptual similarity of features in symmetrical image patches, based on the low-level feature based appearance of objects, e.g., colour, shape, patterns, etc.;
3. **Semantic similarity:** Similarity of semantic categories of the objects in symmetrical image patches, e.g., people, object types, and properties of these objects, such as peoples gazing direction, foreground / background etc.

The proposed characterisation serves as the foundation for analysing and interpreting symmetrical structures in the images; in particular it can be used to identify the elements of the image supporting the symmetrical structure, but also those parts of the image that are not in line with the symmetry, e.g., elements breaking the symmetry. This may be used for investigating the use of balance and in-balance in visual arts, and for analysing how this can be used to guide peoples attention in the context of visual saliency.

### 2.2 A Model of Reflectional Symmetry

For the computational model presented in this paper (Figure 3), we focus on reflectional symmetry in the composition of the image based on layers L1–L3 (Section 2.1), i.e., we investigate image properties based on spatial configuration, low-level feature similarity, and semantic similarity. Towards this we extract image elements \( E \) of the image:

1. **Image patches** are extracted using selective search as described in Uijlings et al. (2013); resulting in structural parts of the image, potential objects and object parts;
2. **People and objects** are detected in the image using YOLO object detection (Redmon et al., 2016);
3. **Human body pose** consisting of body joints and facing direction is extracted using human pose estimation (Cao et al., 2017).

Potential symmetrical structures in the image are defined on the image elements \( E \) using a model of symmetry to identifying pairs of image elements (symmetry pairs) as well as single elements that are constituting a symmetrical configuration.

We consider compositional structure (C1) of images, and similarity (C2) of constituent elements, in particular perceptual similarity in the low-level features, and semantic similarity of objects and regions. The resulting model of symmetrical structure in the image consists of a set of image elements, and the pair-wise similarity relations between the elements.
Symmetrical composition in the case of reflectional symmetry consists of symmetrically arranged pairs of image elements, where one element is on the left and one is on the right of the symmetry axis, and single centred image elements, which are placed on the symmetry axis. To model this, we represent the extracted image elements as spatial entities, i.e. points, axis-aligned rectangles, and line-segments and define constraints on the spatial configuration of the image elements, using the following spatial properties of the spatial entities:

- **position**: the centre-point of a rectangle or position of a point in \( x, y \) coordinates;
- **size**: the width and height of a rectangle \( w, h \);
- **aspect ratio**: the ratio \( r \) between width and height of a rectangle;
- **distance**: euclidean distance \( d \) between two points \( p \) and \( q \);
- **rotation**: the \( \text{yaw}, \text{pitch}, \text{and roll} \) angles between two line-segments in 3D space.

**Symmetrical Spatial Configuration** We use a set of spatial relations holding between the image elements to express their spatial configuration; spatial relations (e.g., *left*, *right*, and *on*)\(^1\) holding between points and lines describe the relative orientation of image elements with respect to the symmetry axis. Towards this, we use the relative position (rel-pos) of an image element with respect to the symmetry axis, which is defined as the distance to the symmetry axis and the \( y \) coordinate of the element.

---

**Image Patches and Objects** Symmetrical configuration of image elements is defined based on their spatial properties using the following two rules.

In the case of a single element \( e \) the centre of the rectangle has to be on the symmetry axis.

\[
\text{symmetrical}(e) \supset \text{orientation}(on, \text{position}(e), \text{symmetry-axis}). \quad (1)
\]

In the case of pairs of elements \( e_i \) and \( e_j \) these have to be on opposite sites of the symmetry axis, and have same size and aspect ratio, further the position of \( e_i \) and \( e_j \) has to be reflected.

\[
\text{symmetrical}(p_i, p_j) \supset \\
\text{orientation}(left, \text{position}(p_i), \text{symmetry-axis}) \land \\
\text{orientation}(right, \text{position}(p_j), \text{symmetry-axis}) \land \\
\text{equal}(\text{aspect-ratio}(p_i), \text{aspect-ratio}(p_j)) \land \\
\text{equal}(\text{size}(p_i), \text{size}(p_j)) \land \text{equal}(\text{rel-pos}(p_i), \text{rel-pos}(p_j)). \quad (2)
\]

---

1. The semantics of spatial relations is based on specialised polynomial encoding as suggested in Bhatt et al. (2011) within constraint logic programming (CLP) (Jaffar & Maher, 1994); CLP is also the framework being used to demonstrate Q/A later in this section.
The model of symmetry serves as a basis for analysing symmetrical structures and defines the attributes that constitute a symmetrical configuration. Additionally to this basic definition of symmetrical configuration of arbitrary image elements, we define rules for symmetry in the placement and layout of humans in the images.

— Human Body Pose. Is given by a set of joints $j$, represented as points, i.e. $\text{pose} = \{j_0, \ldots, j_n\}$. The pose can be either symmetrical within itself, or two people can be arranged in a symmetrical way. Symmetrical body pose is analysed by defining joint pairs $\text{JP} = \{(j_k, j_l), \ldots, (j_m, j_n)\}$, such as (left shoulder, right shoulder), (left elbow, right elbow), etc. and compare the relative position of these pairs with respect to the centre of the person $c_p$.

$\text{symmetrical}(\text{pose}(p)) \supset \forall (j_k, j_l) \text{ equal}(\text{rel-pos}(j_k, c_p), \text{rel-pos}(j_l, c_p))$  

Accordingly, pose of two persons is analysed by defining joint pairs associating each joint of one person to the corresponding joint of the other person, e.g., the left hand of person 1 gets associated to the right hand of person 2.

Further we define symmetrical facing directions of two people based on the rotation of their heads. Towards this we use the $\text{yaw}$, $\text{pitch}$, and $\text{roll}$ angles of a persons head $h_p$, relatively to a front facing head, and define that the facing direction is symmetrical if the $\text{pitch}$ rotation is the same, and the $\text{yaw}$, and $\text{roll}$ rotation are opposite.

$\text{symmetrical}(\text{facing_dir}(p_1), \text{facing_dir}(p_2)) \supset$  
$\text{equal}(\text{pitch}(h_{p_1}), \text{pitch}(h_{p_2})) \land \text{equal}(\text{yaw}(h_{p_1}), -\text{yaw}(h_{p_2})) \land \text{equal}(\text{roll}(h_{p_1}), -\text{roll}(h_{p_2})))$  

Divergence from Symmetrical Configuration. To account for configurations that are only symmetrical in some aspects, as it typically occurs in naturalistic scenes, we calculate the divergences of the configuration from the symmetry model. For each element of the symmetry structure we calculate the divergence from the defined symmetry model, i.e., we focus on divergence with respect
to position, size, aspect ratio, and pose (involving configuration of body parts and joints). We use thresholds on the average of these values to identify hypotheses on (a)symmetrical structures.

(C2) Similarity Measures

Visual Symmetry is also based on similarity of image features; we assess similarity of image patches using CNN features, e.g., obtained from AlexNets (Krizhevsky et al., 2012), or ResNets (He et al., 2016), pre-trained on the ImageNet Dataset (Deng et al., 2009), i.e., we use the extracted features to evaluate perceptual similarity and use ImageNet classifications to evaluate semantic similarity of image patches.

Perceptual Similarity Visual Symmetry is based in perceptual similarity of image features, this denotes the similarity in visual appearance of the image patches. To analyse the perceptual similarity of image patches we use the feature vectors obtained from the network and use cosine similarity to evaluate the similarity of the feature vectors of two image patches. For the case of reflectional symmetry we compare the image patches of all potential symmetry pairs by comparing the features of one patch to the features of the mirrored second patch.

Semantic Similarity On the semantical level, we classify the image patches and compare their content for semantic similarities, i.e., we compare conceptual similarity of the predicted categories. Towards this we use the weighted ImageNet classifications for each image patch with WordNet (Miller, 1995), which is used as an underlying structure in ImageNet, to estimate conceptual similarity of the object classes predicted for the image patches in each symmetry pair. In particular, we use the top five predictions from the AlexNet classifiers and estimate similarity of each pair by calculating the weighted sum of the similarity values for each pair of predicted object categories.

<table>
<thead>
<tr>
<th>Predicate</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>symmetrical_element(E)</td>
<td>Symmetrical elements E.</td>
</tr>
<tr>
<td>non_symmetrical_element(E)</td>
<td>Non-symmetrical elements E.</td>
</tr>
<tr>
<td>symmetrical_objects(SO)</td>
<td>Symmetrical objects SO.</td>
</tr>
<tr>
<td>non_symmetrical_objects(NSO)</td>
<td>Non-symmetrical objects NSO.</td>
</tr>
<tr>
<td>symmetrical_body_pose(SP,SBP)</td>
<td>Symmetrical person SP (pair or single object), and symmetrical parts of body-pose SBP.</td>
</tr>
<tr>
<td>non_symmetrical_body_pose(SE,NSP)</td>
<td>Symmetrical person SP (pair or single object), and non-symmetrical parts of body-pose SBP.</td>
</tr>
<tr>
<td>symmetry_stats(NP,NSP,MD,MS)</td>
<td>Basic stats on symmetrical structure: number of patches NP, number of symmetrical patches NSP, mean divergence MD, and mean similarity MS.</td>
</tr>
<tr>
<td>symmetrical_objects_stats(NO,NSO,MD,MS)</td>
<td>Stats on symmetrical structure of objects: number of objects NO, number of symmetrical objects NSO, mean divergence MD, and mean similarity MS.</td>
</tr>
</tbody>
</table>

Table 1: Sample predicates for querying interpretation model.
2.3 Declarative Symmetry Semantics

The semantic structure of symmetry is described by the model in terms of a set of symmetry pairs and their respective similarity values with respect to the three layers of our model, i.e. for each symmetry pair it provides the similarity measures based on semantic similarity, spatial-arrangement, and low-level perceptual similarity (Table 2). This results in a declarative model of symmetrical structure, which is used for fine-grained analysis of symmetry features and question-answering about symmetrical configuration in images, i.e., using our framework, it is possible to define high-level rules and execute queries in (constraint) logic programming (Jaffar & Maher, 1994) (e.g., using SWI-Prolog (Wielemaker et al., 2012)) to reason about symmetry and directly query symmetrical features of the image.

Symmetrical Structure of the Image As an example consider the image in Table 2. Based on the symmetrical structure extracted from the image, the underlying interpretation model is queryable using utility predicates (see sample predicates in Table 1). The symmetry model as defined in Section 2.2 can be used to query symmetrical (and non-symmetrical) elements of the image using the following rules:

\[
\text{symmetrical\_element}(E) \leftarrow \text{symmetrical}(E).
\]

\[
\text{symmetrical\_element}(E) \leftarrow \text{symmetrical}(E, _); \text{symmetrical}(_, E).
\]

Aggregating results for the \(\text{symmetrical\_element}(E)\) predicate for the example image results in a list of all symmetrical image elements (depicted in the results in Table 2):

\[
\text{SYMETRICAL} = \{0, 2, 8, 10, 11, 12, 14, 15, 17\}...
\]

Similarly we can query for the non symmetrical elements of the image using the following rule:

\[
\text{non\_symmetrical\_element}(P) \leftarrow \text{image\_element}(P), \text{not}\(\text{symmetrical\_element}(P))\).
\]

\[
\text{NON\_SYMETRICAL} = \{1, 3, 4, 5, 6, 7, 9, 13, 16\}...
\]

Divergence The divergence of a image elements from the optimal symmetrical configuration can be directly queried using the \(\text{divergence}\) predicate:

\[
?- \text{divergence}(\text{id}(P1), \text{id}(P2)), \text{Div\_Size}, \text{Div\_AR}, \text{Div\_Pos}.
\]

\[
\text{P1} = 170, \text{P2} = 200, \text{Div\_Size} = \text{div\_size}(9.0, 18.0), \text{Div\_AR} = \text{div\_ar}(0.025384207737148917), \text{Div\_Pos} = \text{div\_pos}(7.5).
\]

Similarity Perceptual and semantic similarity of image elements are queried as follows:

\[
?- \text{similarity}(\text{pair}(\text{id}(P1), \text{id}(P2)), \text{Percept\_Sim}, \text{Semantic\_Sim})
\]

\[
\text{P1} = 170, \text{P2} = 200, \text{Percept\_Sim} = 0.70728711298, \text{Semantic\_Sim} = 0.6666666666666666.
\]

2. Within the (constraint) logic programming language PROLOG, ‘,’ corresponds to conjunction, ‘;’ to a disjunction, and ‘a + b, c’ denotes a rule where ‘a’ is true if both ‘b’ and ‘c’ are true; capitals are used to denote variables, whereas lower-case refers to constants; ‘_’ (i.e., the underscore) is a “don’t care” variable, i.e., denoting placeholders for variable in cases where one doesn’t care for a resulting value.
Table 2: Computational steps to generate the semantic symmetry model.

Step 1) Extracting Image Elements
extract image elements \( E \) consisting of \( E_1, E_2, \) and \( E_3 \):

- \( E_1 \) Image Patches are extracted using selective search as described in Uijlings et al. (2013);
- \( E_2 \) People and Objects are detected in the image using YOLO object detection (Redmon et al., 2016);
- \( E_3 \) Human Body Pose consisting of body joints and facing direction is extracted using human pose estimation (Cao et al., 2017).

Step 2) Semantic and Perceptual Similarity
Compute semantic and perceptual similarity for each pair of image elements \( e_i, e_j \in E \) based on features from CNN layers.
- Compute semantic similarity based on ImageNet classification of image patches
- Compute perceptual similarity based on cosign similarity between CNN features of image elements

Step 3) Symmetry Configuration and Divergence
Identify symmetrical structures in the image elements \( E \) based on the formal definition of symmetry in Section 2.2 and calculate the divergence of elements from this model.

Result A Declarative Model for Semantic Analysis of Symmetry
The process results in the declarative structure consisting of the symmetrical properties of the image given by the elements of the image \( E \) and the divergence from the formal definition of symmetry. This model serves as a basis for declarative query-answering about symmetrical characteristics of the image, e.g. the images showcase results from queries in Section 2.3, analysing symmetrical or non-symmetrical image elements.
The above predicates provide the basis for the semantic analysis of symmetry structures in the image as described in the following.

**Symmetrical Structure of Objects and People** Symmetrical structures in the configuration of objects and people in the image can be queried using the predicate `symmetrical_objects` to get symmetrically configured objects, i.e., pairs of symmetrically positioned objects and single objects in the centre of the image.

```
?- symmetrical_objects(SymObj).
```

For the example image this results in the two people sitting on the bench in the centre of the image.
```
SymObj = pair(id(1), id(2)).
```

Similarly to symmetrical object configurations, objects placed in a non-symmetrical way can be queried as follows:

```
?- non_symmetrical_objects(NonSymObj).
```
```
NonSymObj = id(0).
```

Resulting in objects that are not part of a symmetrical structure, i.e., the person in the left of the image has no symmetrical correspondent in the right of the image.

**Body Pose.** Based on this, the extracted symmetrical objects can be analysed further, e.g., symmetrical configuration of *people* and their *body pose*, can be queried using the following rule:

```
symmetrical_body_pose(pair(P,Q), SymPose) :-
symmetrical_objects(pair(P, Q)),
type(P, 'person'), type(Q, 'person'),
symmetrical_pose(pair(P, Q)), SymPose.
```

This results in the symmetrically placed people, and the elements of the poses that are symmetrical, i.e., the upper-body of person 1 and person 2 are symmetrical.
```
P = id(1), Q = id(2),
SymPose = ['upperbody'].
```

Respectively, non-symmetrical parts of the *body pose* can be queried as follows:
```
?- non_symmetrical_body_pose(pair(P, Q), NonSymPose).
```
```
P = id(1), Q = id(2),
NonSymPose = ['facing direction', 'legs'].
```

As such, the above analysis states that the two people sitting on the bench are placed in a symmetrical way. Their pose is symmetrical in the upper-body, while the facing direction and the legs are not symmetrical.

**Statistics on Image Symmetry** Additionally the model can be used to query statistics on the symmetrical features of an image, e.g., to train a classifier based on the semantic characterisations of symmetry as shown in Section 3. (see Table 3 for additional examples of symmetry statistics)

```
P = id(1), Q = id(2),
NonSymPose = ['facing direction', 'legs'].
```

Similarly statistics on symmetry of objects and people can be queried.
Based on these rules, our model provides a declaratively interpretable characterisation of reflectional symmetry in visual stimuli.

3. Human Evaluation: A Qualitative Study

Experimental Dataset  Human-generated data from subjective, qualitative assessments of symmetry serves many useful purposes: we built a dataset of 150 images consisting of landscape and architectural photography, and movie scenes. The images range from highly symmetric images showing very controlled symmetric patterns to completely non symmetric images. Each participant was shown 50 images selected randomly from the dataset; subjects had to rank the images by selecting one of four categories: not symmetric, somewhat symmetric, symmetric, and highly symmetric. Each image was presented to approximately 100 participants; we calculated the symmetry value as the average of all responses.

Empirical Results  The results from the human experiment suggest, that perception of symmetry varies a lot between subjects. While in the case of no symmetry people tend to agree, i.e. variance
in the answers is very low (see Figure 5), in the case of high symmetry, there is a wider variance in the human perception of symmetry. In particular in the case of images with an average level of symmetry the variance in the answers tends to be high. Qualitatively, there are various aspects on the subjective judgement of symmetry that we can observe in the human evaluation (1 – 3): (1) absence of features decreases the subjective rating of symmetry, e.g., the image in Figure 6a has a nearly perfect symmetry in the image features, but as there are only very few features that can be symmetrical people only perceived it as medium symmetrical, with a high variance in the answers;
Figure 7: Results of empirical evaluation with three different feature set combinations, showing (a) mean accuracy, (b) mean error, and (c) class probability error.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>fs1</td>
<td>41.33</td>
<td>0.01806876383</td>
<td>0.0572886659</td>
</tr>
<tr>
<td>fs1+2</td>
<td>52.00</td>
<td>0.0126452444</td>
<td>0.0400713172</td>
</tr>
<tr>
<td>fs1+2+3</td>
<td>54.00</td>
<td>0.009900461023</td>
<td>0.0375853705</td>
</tr>
</tbody>
</table>

Table 4: Results from classification and prediction pipeline.

(2) symmetrical placement of people in the image has a higher impact on the subjective judgement of symmetry then other objects, e.g., the image in Figure 6b is judged as symmetrical based on the placement of the characters and the door in the middle, but the objects on the left and right side are not very symmetrical; (3) images that are naturally structured in a symmetrical way are judged less symmetrical then those arranged in a symmetrical way, e.g., images of centred faces as depicted in Figure 6c, are rated less symmetrical then other images with similar symmetry on the feature level.

Subjective symmetry interpretation. To evaluate how good our symmetry model reflects subjective human criteria for judging symmetry in naturalistic images, we use the results from the human study to train a classifier and a regressor to predict the symmetry class of an image and predict the average symmetry of the images. For our experiment, we extracted three sets of features \( fs1 - fs3 \) from the symmetry model: \( fs1 \) consists of the cosine similarity between the two halves of each image on each of the 5 convolution layers in an AlexNet; \( fs2 \) consists of the symmetrical properties between image patches, i.e., divergence from symmetrical spatial configuration, and perceptual similarity; and \( fs3 \) consists of the symmetrical properties of object configuration and people in the images. We have 2 models, a classifier and a regressor. A given image is classified into one of the 4 symmetry classes using the classifier. This model is evaluated using the mean accuracy as shown in Figure 7(a). The classifier model also predicts the per class probabilities, this is denoted by `multiclass_proba_model`. This model is evaluated by calculating the Mean Squared Error (MSE) between the predicted probabilities and the percentages from the human data for each class. The per class errors are shown in Figure 7(c) while the mean error is shown in Figure 7(b).

The regressor model predicts the average symmetry value of a given image. The model is evaluated by calculating the MSE between the predicted average symmetry value and average symmetry value from the human data. We use the pipeline optimization method of TPOT (Olson et al., 2016) to automatically build the classification and regression pipelines for the feature sets. This results in a
classification pipeline consisting of an ensemble of DecisionTrees, SVM, RandomForest classifiers while the regression pipeline consists of an ensemble of ExtraTrees and XGBoost regressors. The models are trained and tested on the 3-feature set using 5-fold cross validation, splitting the 150 images into 5 folds. Reported are mean error and classification accuracy (CA).

Results and Discussion The results (Figure 7; Table 4) show that using the features from our symmetry model improves performance in both tasks, i.e., the accuracy for the classification task improves by over 10% (see Table 4) from 41.33 % to 54%, and for the per class probabilities the errors decreases from 0.057 to 0.038. The biggest improvement in the classification and in the prediction of the average symmetry value happens when adding the image patch features $f_s^2$ (Figure 7(a), and (b)). Adding people centred features only results in a small improvement, which may be because only a subset of the images in the dataset involves people. The results on the predicted per class probabilities (Figure 7(c)) show that by adding features from our symmetry model we are able to make better predictions on the variances in the human answers.

4. Discussion and Related Work

Symmetry in images has been studied from different perspectives, including visual perception research, neuroscience, cognitive science, arts and aesthetics (Treder, 2010). The semantic interpretation of symmetry from the viewpoint of perception and aesthetics requires a mixed empirical-analytical methodology consisting of both empirical and analytical methods:

• **Empirical / Human Behaviour Studies.** This involves qualitative studies involving subjective assessments, as well as an evidence-based approach measuring human performance from the viewpoint of visual perception using eye-tracking, qualitative evaluations, and think-aloud analysis with human subjects; and

• **Analytical / Interpretation and Saliency.** This involves the development of computational models that serve an interpretation and a predictive function involving, for instance: (i) multi-level computational modelling of interpreting visuo-spatial symmetry; (ii) a saliency model of visual attention serving a predictive purpose vis-a-vis the visuo-spatial structure of visual media.

Symmetry and (computer) vision Symmetry is an important feature in visual perception and there are numerous studies in vision research investigating how symmetry affects visual perception (Cohen & Zaidi, 2013; Norcia et al., 2002; Machilsen et al., 2009; Bertamini & Makin, 2014), and how it is detected by humans (Wagemans, 1997; Freyd & Tversky, 1984; Árpád Csathó et al., 2004). Most relevant to our work is the research on computational symmetry in the area of computer vision (Liu et al., 2013, 2010). Typically, computational studies on symmetry in images characterise symmetry as reflection, translation, and rotation symmetry; here, reflection symmetry (also referred to as *bilateral* or *mirror symmetry*) has been investigated most extensively. Another direction of research in this area focuses on detecting symmetric structures in objects. In this context Teo et al. (2015) presents a classifier that detects curved symmetries in 2D images. Similarly, Lee & Liu (2012) presented an approach to detect curved glide-reflection symmetry in 2D and 3D images, and
Atadjanov & Lee (2016) uses appearance of structure features to detect symmetric structures of objects.

**Computational analysis of image structure**  Analysing image structure is a central topic in computer vision research and there are various approaches for different aspects involved in this task. Deep learning with convolutional neural networks (CNNs) provide the basis for analysing images using learned features, e.g., AlexNets (Krizhevsky et al., 2012), or ResNets(He et al., 2016), trained on the ImageNet Dataset (Deng et al., 2009). Most recent developments in object detection involve RCNN based detectors such as Ren et al. (2017) and Girshick et al. (2016), where objects are detected based on region proposals extracted from the image, e.g., using selective search (Uijlings et al., 2013) or region proposal networks for predicting object regions. For comparing images, Zagoruyko & Komodakis (2015) and Dosovitskiy & Brox (2016) measure perceptual similarity based on features learned by a neural network.

5. **Summary and Outlook**

Our research addresses visuo-spatial symmetry in the context of naturalistic stimuli in the domain of visual arts, e.g., film, paintings, and landscape and architectural photography. With a principal focus on developing a human-centred computational model of (interpreting) visuo-spatial symmetry, our approach is motivated and driven by three crucial and mutually synergistic aspects, namely: reception, interpretation, and synthesis:

- **Reception**: A behavioural study of the human perception (and explanation) of symmetry from the viewpoint of visual attention, and spatio-linguistic and qualitative characterisation(s);

- **Interpretation**: A computational model of deep semantic interpretation of visual symmetry with an emphasis on human-centred explainability and visual sensemaking;

- **Synthesis**: The ability to apply human-centred explainable models as a basis to directly or indirectly engineer visual media vis-a-via their (predictive) receptive effects, i.e., guiding attention by influencing visual fixation patterns, minimising / maximising saccadic movements (e.g., in animation, gaming, built environment planning, and design).
In this paper, we have focussed on the reception and interpretation aspects; we presented a declarative, computational model of reflectional symmetry integrating (visuospatial) composition, feature-level similarity, and semantic similarity in visual stimuli. Some possible next steps could be:

- **spatio-temporal symmetry and visual perception**: Going beyond static images to analyse symmetry in *space-time* (e.g., as in the films of Wes Anderson (Bhatt & Suchan, 2015)): here, a particular focus is on the influence of space-time symmetry on visual fixations and saccadic eye-movements (Suchan et al., 2016b)

- **visual processing aspect**: More advanced region proposals are possible, and can be naturally driven by newer forms of visual computing primitives and similarity measures. The framework is modular and may be extended with improved or new visual-computing features

- **Resynthesising images**: produce qualitatively distinct classes of (a)symmetry (e.g., Figure 8), and conducting further empirical studies involving qualitative surveys, eye-tracking, think-aloud studies etc

The most immediate outlook of our research on the computational front is geared towards extending the current symmetry model for the analysis of **spatio-temporal symmetry** particularly from the viewpoint of moving images as applicable in film, animation, and other kinds of narrative media. Towards this, we extend the symmetry model to include a richer spatio-temporal ontology, e.g., consisting of ‘space-time’ entities (Suchan & Bhatt, 2016b; Suchan et al., 2018; Schultz et al., 2018) for the analysis of spatio-temporal symmetry. Space-time symmetry analysis will also be complemented with specialised methods that provide a holistic view of the cinematographic “geometry of a scene” (Suchan & Bhatt, 2016a,b). Another promising line of work we are exploring involves relational learning of visuo-spatial symmetry patterns (e.g., based on inductive generalisation (Suchan et al., 2016a)). Explainable learning from (big) visual datasets promises to offer a completely new approach towards the study of media and art history, cultural studies, and aesthetics.

**Acknowledgements**

We thank the anonymous reviewers and ACS 2018 program chairs Pat Langley and Dongkyu Choi for their constructive comments and for helping improve the presentation of the paper. The empirical aspect of this research was made possible by the generous participation of of volunteers in the online survey; we are grateful to all participants. We acknowledge the support of Steven Kowalzik towards digital media work pertaining to manual re-synthesis of categories of stimuli.

**References**


10 - Deeply Semantic Inductive Spatio-Temporal Learning

Published in:
International Conference on Inductive Logic Programming (ILP 2016)

Citation:

This conference paper was published in the International Conference on Inductive Logic Programming (ILP 2016). It develops a general spatio-temporal learning framework implemented within the general Inductive Logic Programming (ILP) setup and based on declarative abstractions of space and motion. The framework supports learning with relational spatio-temporal structures, as available within the theory of space and motion $\Sigma$ (Section 2.2.2), and thus, facilitates learning of visuospatial characteristics directly from positive and negative examples of dynamic visuospatial imagery.

The framework is demonstrated in the context of an example application in learning relational spatio-temporal structure of symmetrical configuration in images. Here, declaratively abstractions of relational structure between image elements constitute the basis to learn spatio-temporal rules about symmetrical configuration.
Deeply Semantic Inductive Spatio-Temporal Learning

Jakob Suchan¹, Mehul Bhatt¹, and Carl Schultz²

Human-Centred Cognitive Assistance¹ hcc.uni-bremen.de and
The DesignSpace Group.,¹,² www.design-space.org
University of Bremen¹, and University of Münster², GERMANY

Abstract. We present an inductive spatio-temporal learning framework rooted in inductive logic programming. With an emphasis on visuo-spatial language, logic, and cognition, the framework supports learning with relational spatio-temporal features identifiable in a range of domains involving the processing and interpretation of dynamic visuo-spatial imagery. We present a prototypical system, and an example application in the domain of computing for visual arts and computational cognitive science.

Keywords: Spatio-Temporal Learning; Dynamic Visuo-Spatial Imagery; Declarative Spatial Reasoning; Inductive Logic Programming; AI and Art

1 INTRODUCTION

Cognitive assistive technologies and computer-human interaction systems involving an interplay of space, dynamics, and cognition necessitate capabilities for explainable reasoning, learning, and control about space, actions, change, and interaction [1]. Prime application scenarios, for instance, include (A1–A5): (A1). activity grounding from video and point-clouds; (A2). modelling and analysis of environmental processes at the geospatial scale; (A3). medical computing scenarios replete with visuo-spatial imagery; (A4). visuo-locomotive human behavioural data concerning aspects such as mobility or navigation, eye-tracking based visual perception research; (A5). embodied human-machine interaction and control for commonsense cognitive robotics. A crucial requirement in relevant application contexts (such as A1–A5) pertains to the semantic interpretation of multi-modal human behavioural or socio-environmental data, with objectives ranging from knowledge acquisition (e.g., medical computing, computer-aided learning) and data analyses (e.g., activity interpretation) to hypothesis formation in experimental settings (e.g., empirical visual perception studies). The focus of our research is the processing and interpretation of dynamic visuo-spatial imagery with a particular emphasis on the ability to learn commonsense knowledge that is semantically founded in spatial, temporal, and spatio-temporal relations and patterns.

DEEP VISUO-SPATIAL SEMANTICS The high-level semantic interpretation and qualitative analysis of dynamic visuo-spatial imagery requires the representational and inferential mediation of commonsense abstractions of space, time, action, change, interaction and their mutual interplay thereof. In this backdrop, deep visuo-spatial semantics denotes the existence of declaratively grounded models —e.g., pertaining to
Recent perspectives on deep visuo-spatial semantics encompass methods for declarative (spatial) representation and reasoning —e.g., about space and motion— within frameworks such as constraint logic programming (rule-based spatio-temporal inference [4, 24]), answer-set programming (for non-monotonic spatial reasoning [27]), description logics (for spatio-terminological reasoning [3]), inductive logic programming (for inductive-abductive spatio-temporal learning [5, 6]) and other specialised forms of commonsense reasoning based on expressive action description languages for modelling space, events, action, and change [1, 2]. In general, deep visuo-spatial semantics driven by declarative spatial representation and reasoning pertaining to dynamic visuo-spatial imagery is relevant and applicable in a variety of cognitive interaction systems and assistive technologies at the interface of (spatial) language, (spatial) logic, and (visuo-spatial) cognition.

**INDUCTIVE SPATIO-TEMPORAL LEARNING (WITH DEEP SEMANTICS)**

This research is motivated by the need to have a systematic inductive logic programming [15] founded spatio-temporal learning framework and corresponding system that:

- provides an expressive spatio-linguistically motivated ontology to predicate primitive and complex (domain-independent) relational spatio-temporal features identifiable in a broad range of application domains (e.g., A1–A5) involving the processing and interpretation of dynamic visuo-spatial imagery.
- supports spatio-temporal relations natively such that the semantics of these relations is directly built into the underlying ILP-based learning framework.
- supports seamless mixing of, and transition between, quantitative and qualitative spatial data.

We particularly emphasise and ensure compatibility with the general setup of (constraint) logic programming framework such that diverse knowledge sources and reasoning mechanisms outside of inductive learning may be directly interfaced, and reasoning / learning capabilities be combined within large-scale integrated systems for cognitive computing.

**2 LEARNING FROM RELATIONAL SPATIO-TEMPORAL STRUCTURE: A GENERAL FRAMEWORK AND SYSTEM**

We present a general framework and working prototype for an inductive spatio-temporal learning system with an elaborate ontology supporting a range of space-time features;
we demonstrate the functional capabilities from the viewpoint of AI-based computing for the arts & social sciences, and computational cognitive science.

2.1 THE SPATIO-TEMPORAL DOMAIN $O_{SP}$, AND QS

The spatio-temporal ontology $O_{SP} \equiv_{df} \langle E, R \rangle$ is characterised by the basic spatial entities ($E$) that can be used as abstract representations of domain-objects and the relational spatio-temporal structure ($R$) that characterises the qualitative spatio-temporal relationships amongst the supported entities in ($E$). The following primitive spatial entities are sufficient to characterise the learning mechanism and its sample application for this paper:

- A point is a pair of reals $x$, $y$; a vector is a pair of reals $v_x$, $v_y$; an oriented point consists of a point $p$ and a vector $v$; a line segment is a pair of end points $p_1$, $p_2$ ($p_1 \neq p_2$); a rectangle is a point $p$ representing the bottom left corner, a direction vector $v$ defining the orientation of the base of the rectangle, and a real width and height $w$, $h$ ($0 < w$, $0 < h$); an axis-aligned rectangle is a rectangle with fixed direction vector $v = (1, 0)$; a circle is a centre point $p$ and a real radius $r$ ($0 < r$); a simple polygon is defined by a list of $n$ vertices (points) $p_1, \ldots, p_n$ (spatially ordered counter-clockwise) such that the boundary is non-self-intersecting, i.e., there does not exist a polygon boundary edge between vertices $p_i, p_{i+1}$ that intersects some other edge $p_j, p_{j+1}$ for all $1 \leq i < j < n$ and $i + 1 < j$.

Spatio-temporal relationships ($R$) between the basic entities in $E$ may be characterised with respect to arbitrary spatial and spatio-temporal domains such as mereotopology, orientation, distance, size, motion; Table 1 lists the relevant supported relations from the viewpoint of established spatial abstraction calculi such as the Region Connection Calculus [16], Rectangle Algebra and Block Algebra [7], LR Calculus [20], Oriented-Point Relation Algebra (OPRA) [14], and Space-Time Histories [8, 9].

QS – ANALYTIC SEMANTICS FOR $O_{SP}$ We adopt an analytic approach to spatial reasoning, where the semantics of spatial relations are encoded as polynomial constraints within a (constraint) logic programming setup. The analytic method supports the integration of qualitative and quantitative spatial information, and provides a means for sound, complete and approximate spatial reasoning [4]. For example, let axis-aligned rectangles $a, b$ each be defined by a bottom-left vertex $(x_1, y_1)$ and a width and height $w_i, h_i$, for $i \in \{a, b\}$ such that $x_1, y_1, w_i, h_i$ are reals. The relation that $a$ is a non-tangential proper part of $b$ corresponds to the polynomial constraint:

$$(x_b < x_a) \land (x_a + w_a < x_b + w_b) \land (y_b < y_a) \land (y_a + h_a < y_b + h_b)$$

Continuing with the example, this is generalised to arbitrarily oriented rectangles. Determining whether a point is inside an arbitrary rectangle is based on vector projection. Point $p$ is projected onto vector $v$ by taking the dot product:

$$(x_p, y_p) \cdot (x_v, y_v) = x_p x_v + y_p y_v.$$

With this approach, the task of determining whether a set of spatial relations is consistent then becomes the task of determining whether a system of polynomial constraints is satisfiable. We emphasise that our approach and framework are not limited to the above
<table>
<thead>
<tr>
<th><strong>SPATIAL DOMAIN (QS)</strong></th>
<th><strong>Formalisms</strong></th>
<th><strong>Spatial Relations (R)</strong></th>
<th><strong>Entities (E)</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mereotopology</td>
<td>RCC-5, RCC-8, [16]</td>
<td>disconnected (dc), external contact (sc), partial overlap (po), tangential proper part (tpp), non-tangential proper part (ntpp), proper part (pp), part of (p), discrete (d), overlap (o), contact (c)</td>
<td>arbitrary rectangles, circles, polygons, cuboids, spheres</td>
</tr>
<tr>
<td></td>
<td>Rectangle &amp; Block algebra [1]</td>
<td>proceeds, meets, overlaps, starts, during, finishes, equals</td>
<td>axis-aligned rectangles and cuboids</td>
</tr>
<tr>
<td>Orientation</td>
<td>LR [20]</td>
<td>left, right, collinear, front, back, on</td>
<td>2D point, circle, polygon with 2D line</td>
</tr>
<tr>
<td></td>
<td>ORPA [14]</td>
<td>facing towards, facing away, same direction, opposite direction</td>
<td>oriented points, 2D/3D vectors</td>
</tr>
<tr>
<td>Distance, Size</td>
<td>QDC [10]</td>
<td>adjacent, near, far, smaller, equi-sized, larger</td>
<td>rectangles, circles, polygons, cuboids, spheres</td>
</tr>
<tr>
<td>Dynamics, Motion</td>
<td>Space-Time Histories [8, 9]</td>
<td>moving: towards, away, parallel, growing / shrinking: vertically, horizontally: passing: in front, behind, splitting / merging</td>
<td>rectangles, circles, polygons, cuboids, spheres</td>
</tr>
</tbody>
</table>

Table 1. The Spatio-Temporal Domain $O_{sp}$ supported within the Learning Framework

entities; a wider class of 2D and 3D spatial entities are supported and may be defined as per domain-specific and computational needs [4, 18, 27, 19].

**INDUCTIVE LEARNING WITH THE SPATIAL SYSTEM $< O_{SP}, QS >$** Learning is founded on the Aleph ILP system [21]. Learning spatio-temporal structures, is based on integrating the spatial ontology $O_{sp}$ described above, into the basic learning setup of ILP.

**Given:** (1) A set of examples $E$, consisting of positive and negative examples for the desired spatio-temporal structure, i.e., $E = E^+ \cup E^-$, where each example is given by a set of spatio-temporal observations in the domain; (2) the (spatio-temporal) background knowledge $B$.

The spatio-temporal learning domain is defined by basic spatial entities ($E$) constituting the domain objects, the relational spatial structure ($R$) describing the spatio-temporal configuration of spatial entities in the domain, and rules defining spatio-temporal phenomena and characteristics of the domain. In this context, spatio-temporal facts characterising the learning examples $E$ can be given as, (a) numerical representation of domain objects, (b) qualitative relations between spatial entities, or (c) a mixed qualitative-quantitative representations, where the facts are partially grounded in numerical observations.

**Learning:** The learning task is defined as finding hypothesis $H$ consisting of spatio-temporal relations ($R$) holding between basic spatial entities ($E$), such that $H \cup B \models E^+$, and $H \cup B \not\models E^-$. As such, the spatial ontology $O_{sp}$ constitutes an integrated part of the learning setup and spatio-temporal semantics are available throughout the learning process.

3 **LEARNING CINEMATOGRAPHIC PATTERNS AND THEIR VISUAL RECEPTION: THE CASE OF SYMMETRY**

Aimed at cognitive film studies and visual perception research, we present a use-case pertaining to the (visual) learning of cinematographic patterns of symmetry and its vi-
Fig. 1. Positive examples for symmetric scene structures at the object level

ual reception (by means of eye-tracking) by subjects.¹ To demonstrate the temporal aspect of the learning framework, we demonstrate the capability to learn “axioms of visual perception” from dynamic eye-tracking data; both the chosen films and their corresponding eye-tracking data are obtained from a large-scale experiment in visual perception of films [23, 22]. The presented example translates to a variety of cases involving visual perception and human behaviour studies.

Learning Spatial Structures: Object-Level Symmetry As an example for learning spatial structures, we consider symmetry in the relative object placement in a movie scene (see Fig. 1). In particular, learning is based on the spatial configuration of people, faces, and their facing direction, directly obtained from computer vision algorithms as described in [23]. In this context, positive and negative examples, are given as numerical spatial facts about domain objects in the image.

```
... detection(id(0), image(3), class(person), rectangle(point(319, 194), 319, 456)).
  detection(id(1), image(3), class(person), rectangle(point(678, 215), 367, 452)).
  detection(id(0), image(3), class(face), rectangle(point(438, 246), 86, 86)).
  detection(id(1), image(3), class(face), rectangle(point(745, 284), 87, 87)).
  2d_facing_dir(id(0), image(3), vector(0.550864, 0.834595), magnitude(6.26042)).
  2d_facing_dir(id(1), image(3), vector(-0.500519, 0.865726), magnitude(4.82556)).
... 
```

We define representations of domain objects linking the numerical description of objects in the image to basic spatial entities describing different aspects of these objects, e.g. the bounding box (rectangles), or the center-point (points).

```
entity(center(person(P)), point(X, Y), image(Img)) :-
  detection(_, image(Img), class(person), rectangle(point(Xr, Yr), W, H)),
  X is Xr + W/2, Y is Yr + H/2.
```

In addition to the detected domain objects, we define abstract geometric objects needed to describe symmetry, e.g. the symmetry axis in the center of the image.

```
entity(symmetry_obj(center_axis), line(X, 0, X, Y), image(Img)) :-
  img(image(Img)), media_size(size(MediaWidth, MediaHeight), image(Img)),
  X is MediaWidth/2, Y is MediaHeight.
```

**Learning:** We learn the relational spatial structure consisting of qualitative spatial relationships characterising symmetry in the configuration of the spatial entities in the image, i.e. we consider relations of topology, orientation, distance, and size.

```
:- modeh(1,symmetric(+img)).
:- modeb(*,entity(#ent,-obj,+img)).
:- modeb(*,topology(rcc8(#rel),+obj,+obj)).
:- modeb(*,distance(#rel,+obj,+obj)).
:- modeb(*,orientation(#rel,+obj,+obj)).
:- modeb(*,size(#rel,+obj,+obj)).
```

¹ Our case-study is motivated by a broader multi-level interpretation of symmetry from the viewpoint of film cinematography [25]; however, the specific example of this paper focusses on one aspect of this multi-level symmetry characterisation involving relative object placement in a movie scene.
Exemplary symmetrical spatial structures, learned by the system include the following.

\[
\text{symmetric}(A) \leftarrow \text{entity}(\text{center(person(0))},B,A), \text{entity}(\text{center(person(1))},C,A), \text{entity}([\text{symmetry_object}](\text{center_axis}),D,A), \text{distance}(\text{equidistant},D,C,B).
\]

\[
\text{symmetric}(A) \leftarrow \text{entity}(\text{person(0)},B,A), \text{entity}(\text{person(1)},C,A), \text{size}(\text{same},C,B).
\]

**Learning Spatio-Temporal Dynamics: Axioms of Perception**

We illustrate learning of spatio-temporal dynamics in the context of visual perception, by learning perceptual patterns from eye-tracking data and people tracks in a movie scene. As an example we focus on attention of a person switching from one individual to another.

\[
\text{detection}(\text{id}(0), \text{frame}(426), \text{class}(\text{person}), \text{rectangle}(\text{point}(385,66),244,271)).
\]

\[
\text{detection}(\text{id}(1), \text{frame}(426), \text{class}(\text{person}), \text{rectangle}(\text{point}(111,68),332,276)).
\]

\[
\text{gazepoint}(\text{frame}(426), \text{point}(859,212)).
\]

**Learning:**

We adapt the general learning setup of the example above, for learning spatio-temporal dynamics by introducing the predicate \(\text{holds-in/2}\) to denote that a spatial relation holds between two entities at a time point.

\[
\text{:- mode}(-, \text{holds-in}(\text{topology}(\text{rel}, +\text{ent}, +\text{ent}), +\text{time})).
\]

\[
\text{:- mode}(+, \text{time}(\text{rel}, +\text{time}, +\text{time})).
\]

\[
\ldots
\]

Spatio-temporal dynamics constituting attention switches include the following.

\[
\text{att_switch}(B) \leftarrow \text{holds-in}(\text{topology}(\text{inside}, \text{gaze}, \text{person}(1)), A), \text{holds-in}(\text{topology}(\text{inside}, \text{gaze}, \text{person}(2)), B), \text{time}(\text{consecutive}, A, B).
\]

### 4 DISCUSSION AND OUTLOOK

Directly comparable to this research is the line of work on integrated inductive-abductive reasoning for learning spatio-temporal relational models from video in [5, 6]; here, spatio-temporal learning in the context of ILP has only been addressed for the case of topological relations. Furthermore, the ILP learning framework does not have built-in semantics for the topological relations. Aside from this, learning relational spatial structures was investigated in the context of learning spatial relations from language [12], and within the geospatial domain [13, 26]. Probabilistic Logic Programming frameworks such as PRISM [17] and ProbLog [11] have been used for learning parameters, and the structure, of probabilistic logic programs, although (qualitative) spatial reasoning has not been directly addressed. The main point-of-departure of this paper with respect to the state of the art in (qualitative) spatial learning is that the semantics of spatial, temporal, and spatio-temporal relations are directly built within the inductive learning framework of ILP. Pragmatically, what this implies is that it is possible to seamlessly describe a learning problem using a generic relational spatio-temporal ontology directly as part of a logic programming based learning environment. To the best of our knowledge, such a general spatio-temporal learning framework with built in semantics for mixed qualitative-quantitative spatio-temporal reasoning capabilities has not been available before. Furthermore, the ontology of space-time features supported in our framework goes much beyond topological relations addressing orientation, distance, and size. Future research will focus on enhancing the expressivity of the spatio-temporal relations to cover a wider range of domain-independent features characterising spatio-temporal dynamics.
Bibliography


Chapter 6
Discussion and Outlook

We summarise the main aspects of this thesis, discuss the results, and point out future research directions. The thesis is discussed, with respect to (1), the core technical and developmental results, (2), application focus and interdisciplinary aspects, and (3), the broader aim of this thesis on the integration of vision and semantics. Future research directions are discussed including potential technical developments emanating from the technical results of this thesis and general considerations towards developing cognitive vision systems.
This thesis develops a domain neutral theory for declarative reasoning about space and motion, and general methods for semantic sensemaking with visuospatial imagery were presented, with a focus on three core reasoning tasks (R1-R3, details in Section 2.3.1):

R1. **Semantic Question-Answering**, where declarative characterisations of spatio-temporal interactions, including qualitative abstractions of space, motion, and events, serve as a backbone for grounding high-level symbolic representations of (human) interactions in perceived object motion, i.e., directly coming from low-level sensor data, facilitating query answering within logic programming.

R2. **Visuospatial Abduction**, where hypotheses on event sequences, explaining observed visuospatial dynamics, are generated based on deep semantic representations of object interactions grounded in visuospatial scene dynamics, consisting of object detections, motion tracks, etc.

R3. **Learning Explainable Visuospatial Models**, where explainability driven declarative interpretation models of visuospatial image characteristics form the basis for analysing visuospatial phenomena and for inductive generalisation based learning of relational spatio-temporal structures.

We demonstrated the applicability of declarative reasoning about visuospatial dynamics in diverse application scenarios. Examples developed and demonstrated in this thesis include (A1-A3, details in Section 2.3.2):

A1. **Autonomous Driving**, focusing on perceptual sensemaking in the context of human-centred considerations, where the interaction between the car and its environment, e.g. other traffic participants, pedestrians and cyclists, people on the road, are of interest.

A2. **Commonsense Cognitive Robotics**, focusing on embodied interactions, where declaratively characterised image schematic representations are used to provide an abstraction for the semantic interpretation of sensor data.

A3. **Behavioural Studies**, focusing on interpretation and question-answering with human behaviour data. Examples in this thesis include, (a) cognitive media studies, using declarative semantics for analysing human engagement with visuo-auditory media, i.e., focusing on cognitive film studies analysing spectators reception of the moving image; and (b) visual perception studies, where declarative interpretation models of image characteristics are used to analyse and learn subjective factors in the reception and interpretation of arts and aesthetics, i.e., focusing on the perception of (reflectional) symmetry.

Further, we have shown the potential of tightly integrated AI and Computer Vision methods for semantic sensemaking, and have evaluated these with real-world community established benchmarks.
In the following we discuss the key results of this thesis and highlight main technical outcomes with respect to the aforementioned reasoning tasks. Finally, we discuss future research directions emanating from the presented work.

### 6.1. Results

Outcomes and results of our research on cognitive vision and declarative reasoning about space and motion are multifaceted, involving core technical developments resulting in general methods and tools, evaluated with real-world datasets; interdisciplinary and application focused results demonstrated in diverse example domains; and general findings on the integration of vision and semantics.

**Core Technical Outcomes.** From a technical point of view, key results of this thesis are general methods and tools focusing on the core reasoning capabilities (R1-R3) semantic question-answering, visuospatial abduction, and learning of explainable visuospatial models. These methods and tools include:

- A domain neutral theory about space and motion $\Sigma$ for reasoning about relational spatio-temporal structures based on foundations in declarative spatial reasoning;
- Deep semantic representations for (embodied) visuospatial sensemaking based on the theory about space and motion and implemented in CLP, suitable for semantic interpretation and question-answering with visuospatial imagery;
- A general and real-time (online) capable method founded in ASP for jointly abducing semantic explanations and corresponding motion tracks, based on visuospatial observations and deep semantic abstractions of space and motion;
- A neurosymbolic pipeline for characterising explainability driven interpretation models for representing and analysing visuospatial characteristics and phenomena in visual imagery;
- A general method for inductive generalisation with spatio-temporal primitives, based on ILP and suitable for learning relational spatio-temporal structures from visual imagery.

Importantly, building on powerful logic programming paradigms (CLP, ASP, and ILP), the developed tools and methods can be directly integrated within larger AI systems where perceptual sensemaking is a concern.

**Quantitative Evaluation.** Empirical results have shown that deep semantic reasoning is capable of improving performance in classical computer vision tasks, i.e., multi-object tracking, and provides a robust representation for human-centred visual sense-making, e.g., for autonomous driving, embodied cognitive robotics, natural language processing, and visual perception research.
Core results from empirical evaluation are:

- Visual abduction based hypothesising of object interactions for tracking scene dynamics leads to improved tracking of low-level object motion (Suchan, Bhatt, and Varadarajan, 2019, 2021; Suchan, Bhatt, Wałęga, et al., 2018);
- Further, ASP-based visual abduction can achieve real-time performance in real-world scenarios, as shown for abducting scene dynamics in autonomous driving (Suchan, Bhatt, and Varadarajan, 2019, 2021);
- Declarative models of space and motion provide meaningful abstractions for modelling human-centred visuospatial characteristics in images, and facilitate learning of subjective perception of visuospatial characteristics, as demonstrated for the case of subjective perception of reflectional symmetry (Suchan, Bhatt, Vardarajan, et al., 2018); and
- Declarative abstractions of spatial dynamics improve robustness of natural language generation and parsing in embodied robot interaction tasks, and provide an efficient way to deal with noisy visual perception data (Spranger, Suchan, and Bhatt, 2016).

**Application and Interdisciplinarity.** We demonstrated deep semantic visuospatial reasoning within diverse application areas (A1-A3) and showed how declaratively grounded characterisations of visuospatial dynamics facilitate semantic sensemaking with dynamic visual imagery. In this sense, dynamic spatio-temporal relations and motion patterns are powerful abstraction mechanisms for declarative grounding of dynamic interactions, capable of providing rich semantics for commonsense interpretation, explanation and learning of scene dynamics.

This was demonstrated for:

- Reasoning about safety-criticality in driving situations, where deep semantics provide an abstraction layer for hypothesising and anticipating object dynamics, e.g., for reasoning about (re)appearance of occluded entities;
- Embodied grounding in cognitive robotics, where deep semantics were utilised for representing human-object interactions based on a declarative model of the human body grounded in skeleton data from RGB-D sensing;
- Analysing spectators gaze vis-a-vis the moving image, where deep semantics are used for representing the dynamics of scene objects extracted from the movie scene together with perceptual artefacts obtained from spectators eye-tracking data; and
- Representing visuospatial characteristics grounded in perceptual features for generating declarative interpretation models of visual perception, demonstrated for the case of reflectional visual symmetry.
In particular, application of the developed methods and tools has demonstrated generality, domain independence, and transferability of the underlying visuospatial abstraction and reasoning capabilities.

From an interdisciplinary perspective, cognitive vision systems have demonstrated great value as an analysis toolbox in the social sciences, in particular, within this thesis, we presented application in visual perception research in cognitive film studies (Suchan, Bhatt, and Yu, 2016). Another application in this direction, not included in this thesis, are empirical studies of human visuo-locomotive way-finding behaviour (Bhatt, Suchan, Kondyli, et al., 2016; Bhatt, Suchan, Schultz, et al., 2016). Further, we have shown that the developed theory of space and motion can serve as a meaningful abstraction layer for embodied natural language processing.

Broad Intellectual Merit. On a broader level, this thesis highlights declarative models of space and motion as a means for integrating semantics and vision within human-centred AI systems. Qualitative and quantitative results demonstrate that systematic and general tools integrating semantic, knowledge driven, representation and reasoning, with data-driven visual computing can help to improve the performance of visual perception methods. In particular, integrating semantically grounded representations of high-level scene dynamics and low-level object motion yield tremendous potential for improving computational perceptual sensemaking, where deep semantic abstractions ensure consistency of semantic (commonsense) interpretations and perceptual (subsymbolic) scene structure, obtained from (deep learning based) computer vision. Further, deep visuospatial semantics provide the formal foundations for (visual) commonsense reasoning on a symbolic (human understandable) level.

6.2. Future Directions

The presented theoretical foundations and the results of this thesis offer several opportunities for future research within computational cognitive vision and perceptual sensemaking. Firstly, our development of a systematic, modular, and general visual sensemaking methodology opens up several possibilities for further technical developments and extensions, including the integration with specialised methods and theories for commonsense reasoning, probabilistic models, and neurosymbolic reasoning & learning. Secondly, the aim of a systematic integration of Semantics and Vision poses a range of important questions for future research. These include practical questions, such as how to benchmark and evaluate semantic sensemaking in visual perception using specialised datasets, as well as long-term considerations concerning the development of human-like perceptual sensemaking, including research topics such as multi-level abstractions, natural (embodied) learning, and semantics in computer vision.

In the following we present select directions for future work, directly founded on declarative space and motion and the broader arc of integrating vision and semantics, as presented in this thesis.
Commonsense Reasoning
Research in the fields of AI and KR has developed a huge collection of theories and formalisms for reasoning about commonsense knowledge. These include for instance action languages for reasoning about actions, events, and changes, epistemic reasoning for modelling knowledge and beliefs, reasoning about (object) affordances, naive physics, etc. Such commonsense reasoning capabilities are important for high-level understanding of event dynamics, e.g., for reasoning about (spatio-temporal) effects and preconditions of perceived events, or for reasoning about intention and beliefs of people in the scene.

The general, modular, and elaboration tolerant nature of the presented framework and its systematic development within mainstream KR methods directly supports integration with specialised KR methods. Thus, a natural continuation of the work presented in this thesis lies in extending the reasoning capabilities and increasing the expressivity of the underlying formalisations using such established theories for commonsense reasoning. In this sense, each of the above mentioned reasoning strands constitute an interesting direction for further development, and could for instance, be directly implemented within the declarative framework of ASP.

Spatio-Temporal Uncertainty
The ability to process and reason about uncertainty and probabilities within spatio-temporal representations is important in a range of different tasks within visuospatial sensemaking, including hypothesis optimisation, interpolation and simulation of (spatio-temporal) scene dynamics, modelling preferences and event likelihoods, reasoning about noise and missing / faulty information. For instance, this is important for finding the most likely hypothesis form a set of hypotheses, e.g., based on probability distributions, or for predicting object motion, based on statistical evidence and / or human preferences. Further, a systematic treatment of uncertainty is also needed to deal with noise inherent to visual perception systems at all levels of the sensemaking process. In this thesis uncertainty within the spatio-temporal dynamics of a scene and noise in visual observations is handled directly within the low-level visual processing methods, using naive weights / costs. While this suffices for the methods and examples in this thesis, a systematic integration of uncertainty is needed to increase robustness and to model more complex probabilistic relationships. Concerning future work, a natural next step is to handle uncertainty in a systematic way, by either integrating declarative probabilistic models directly within the logic programming framework, or possibly independently as a separate module. One approach towards this would be to explore probabilistic reasoning in ASP, or to integrate with probabilistic logic programming as outlined in (Schultz, Bhatt, and Suchan, 2016).

Neurosymbolic Integration and Learning
Visuospatial explainability and learning is based on characterisations of visual imagery on different levels of abstraction. In particular, interpreting and making sense of natural visuospatial dynamics and phenomena involve symbolic relational characteristics
and sub-symbolic features. Reflectional symmetry, as alluded to in this thesis, constitutes one example for such multi-level characterisations, where the characterisation of symmetry integrates sub-symbolic visual similarity with relational spatio-temporal structure and semantics. Such neurosymbolic representations are not only important for characterising symmetry but they are also fundamental in most other visuospatial sensemaking, be it identity maintenance (e.g., of reappearing entities in the context of motion tracking), or the interpretation and learning of (fine-grained) interactions and object manipulation.

Learning of visuospatial structures, as presented in Chapter 5, constitutes a first step towards the goal of neurosymbolic integration and provides prototypical implementations for inductive learning of relational visuospatial structures. However, a systematic integration of low-level neural / weight learning with inductive structure learning was not pursued. Such an integration would be crucial for learning natural visuospatial dynamics and phenomena directly from (multimodal) sensory data. Preliminary considerations on how declarative explainability and learning can lead to integrated neurosymbolic learning are outlined in (Bhatt, Suchan, and Varadarajan, 2019).

**Evaluation and Benchmarking**

Developing a real-world dataset of dynamic (human) interactions categorically from the viewpoint of human-centred AI, together with appropriate evaluation schemes for cognitively motivated sensemaking abilities, such as *anticipation / projection, causal reasoning*, and *explanation* would constitute an important step, moving towards a cognitive benchmark for deep visuospatial semantics. For instance, in the context of autonomous driving, such a dataset could provide an evaluation scheme for benchmarking multimodal interactions between traffic participants and human-factors in driving (e.g., visual complexity) (Kondyli and Bhatt, 2020; Kondyli, Bhatt, and Suchan, 2020). As most popular benchmark datasets lack the necessary annotations and evaluation schemes for evaluating semantically grounded structures, such a dataset would fill a gap in the current landscape of computer vision driven datasets and could provide a semantic perspective on evaluating visual processing and visuospatial sensemaking.

**KR and Vision**

The integration of semantics and commonsense knowledge with state of the art computer vision methods remains an open and challenging research field offering a range of interesting opportunities for further development. Whereas research in computer vision primarily focusses on bottom-up data and task driven development of powerful visual processing methods, research on semantic reasoning about visual data (in **KR**) focusses on top-down, general and high-level approaches for structured knowledge representing and reasoning, e.g., driven by formal (logic-based) methods. A true integration of **Vision and Semantics** would also require that these two strands of research come together with a shared agenda (encompassing research questions, benchmarks) and that methods and tools developed in the respective fields of research function together, thereby closing the gap between abstract conceptual (symbolic) reasoning and real-world (sub-
symbolic) visual processing.
The work presented in this thesis constitutes one approach towards this goal and provides an initial step towards this integration of vision and semantics. However, there is still a long way to go and future research on cognitive vision needs to address a range of key challenges on its way to human-like (multimodal) perceptual sensemaking. Select fundamental questions worth pointing out include:

- Developing and maintaining holistic, multi-level characterisations of semantic knowledge, beliefs, perceived information, etc., capable of adapting to new information available at different levels of abstraction and ensuring consistency within its representation of the world;

- Learning conceptual knowledge from embodied (sensory-motor) experience, including natural interaction, demonstration, and description, involving topics such as semi-, and unsupervised learning, and lifelong learning;

- Exploring relational visuospatial structures as an intermediate layer for semantic grounding in vision and language, e.g., by utilising image schematic concepts;

- Using deep semantics as articulated in this thesis for characterising domain knowledge and context for visual relation detection and for modelling conceptual knowledge for object detection and scene understanding.

At a general level, we posit that future research on cognitive vision systems and next generation AI systems has to be categorically pursued from a human-centred perspective in order to develop systems that function in an environment populated by humans, and that are able to assist and interact with humans in everyday situations.


McCarthy, J. (1963). “Situations, Actions, and Causal Laws”. In: *Stanford Artificial Intelligence Laboratory and Memo (Stanford Artificial Intelligence Laboratory)*.


Appendices
Chapter A
List of All Publications

Publications included in this thesis are marked with a star (⋆). Furthermore, the list of Select Publications Included in this Thesis can be found in the preface of this thesis (Pg. xiii).


Chapter B
Dissemination and Collaboration

B.1. Tutorials

In addition to publications, select aspects of the research in this thesis are also disseminated in form of tutorials to the AI community in the following events:

- **Tutorial on:**
  - Spatial Cognition in the Wild: Methods for Large-Scale Behavioural Research in Visuo-Locomotive Perception.
  - at: ACM Symposium on Eye Tracking Research & Applications (**ETRA 2022**), Seattle, USA, June 8 - 11, 2022
  - Presenters: Mehul Bhatt, Vasiliki Kondyli, Jakob Suchan, Vipul Nair

- **Tutorial on:**
  - at: the Advanced Course on AI on Human Centered AI (**ACAI 2021**), Berlin, Germany, October 11-14, 2021
  - Presenters: Mehul Bhatt, Jakob Suchan

- **Tutorial on:**
  - at: the Thirty-First International Joint Conference on Artificial Intelligence (**IJCAI 2021**)
  - Montreal, Canada, August 21-26, 2021
  - Presenters: Mehul Bhatt, Jakob Suchan

- **Tutorial on:**
  - at: the Twenty-Fourth European Conference on Artificial Intelligence (**ECAI 2020**)
  - Santiago de Compostela, Spain, August 29-September 2, 2020
  - Presenters: Mehul Bhatt, Jakob Suchan
Tutorial on:
at: the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018)
New Orleans, Louisiana, USA, February 2-7, 2018
Presenters: Mehul Bhatt, Jakob Suchan

Tutorial on:
Spatial Cognition in the Wild: Methods for Large-Scale Behavioural Research in Visuo-Locomotive Perception.
at: the ACM Symposium on Eye Tracking Research & Applications (ETRA 2018)
Warsaw, Poland, June 14-17, 2018
Presenters: Mehul Bhatt, Jakob Suchan

Invited Plenary Workshop on:
Computational Cognitive Vision for Human-Behaviour Interpretation.
at: the International Conference on Multimodal Communication (ICMC 2017)
Osnabrück, Germany, June 9-11, 2017
Presenter: Jakob Suchan

Tutorial on:
at: the Twenty-Second European Conference on Artificial Intelligence (ECAI 2016)
The Hague, Netherlands, August 29 - September 2, 2016
Presenters: Mehul Bhatt, Jakob Suchan
B.2. External Collaboration

1. Vasiliki Kondyli, Örebro University, Sweden
   on human centred visual perception studies focusing on architectural design (Bhatt, Suchan, Kondyli, et al., 2016) and human factors in driving (Kondyli, Bhatt, and Suchan, 2020).

2. Michael Spranger, Sony CSL, Tokyo, Japan
   on declarative models of spatio-temporal dynamics for language processing in robotic interactions, resulting in (Spranger, Suchan, Bhatt, and Eppe, 2014) and (Spranger, Suchan, and Bhatt, 2016)

3. Stella Yu, UC Berkeley and ICSI Vision Group, United States, and Ali Amirshahi, NTNU, Norway
   on explainability driven declarative characterisation of visual structure, focusing on the case of reflectional symmetry. Resulting in (Suchan, Bhatt, Vardarajan, et al., 2018) and (Suchan, Bhatt, and Yu, 2016).