

# TALKING ABOUT THE MOVING IMAGE

## A Declarative Model for Image Schema Based Embodied Perception Grounding and Language Generation

Jakob Suchan<sup>1,2</sup>, Mehul Bhatt<sup>1,2</sup>, and Harshita Jhavar<sup>2,3</sup>

<sup>1</sup> University of Bremen, Germany

<sup>2</sup> The DesignSpace Group  
[www.design-space.org/Next](http://www.design-space.org/Next)

<sup>3</sup> MANIT (Bhopal, India)

**Abstract.** We present a general theory and corresponding declarative model for the embodied grounding and natural language based analytical summarisation of dynamic visuo-spatial imagery. The declarative model —encompassing spatio-linguistic abstractions, image schemas, and a spatio-temporal feature based language generator— is modularly implemented within Constraint Logic Programming (CLP). The implemented model is such that primitives of the theory, e.g., pertaining to space and motion, image schemata, are available as first-class objects with *deep semantics* suited for inference and query. We demonstrate the model with select examples broadly motivated by areas such as *film*, *design*, *geography*, *smart environments* where analytical natural language based externalisations of *the moving image* are central from the viewpoint of human interaction, evidence-based qualitative analysis, and sensemaking.

**Keywords:** *moving image, visual semantics and embodiment, visuo-spatial cognition and computation, cognitive vision, computational models of narrative, declarative spatial reasoning*

## 1 INTRODUCTION

Spatial thinking, conceptualisation, and the verbal and visual (e.g., gestural, iconic, diagrammatic) communication of commonsense as well as expert knowledge about the world —the *space* that we exist in— is one of the most important aspects of everyday human life [Tversky, 2005, 2004, Bhatt, 2013]. Philosophers, cognitive scientists, linguists, psycholinguists, ontologists, information theorists, computer scientists, mathematicians have each investigated *space* through the perspective of the lenses afforded

by their respective field of study [Freksa, 2004, Mix et al., 2009, Bateman, 2010, Bhatt, 2012, Bhatt et al., 2013a, Waller and Nadel, 2013]. Interdisciplinary studies on visuo-spatial cognition, e.g., concerning ‘visual perception’, ‘language and space’, ‘spatial memory’, ‘spatial conceptualisation’, ‘spatial representations’, ‘spatial reasoning’ are extensive. In recent years, the fields of *spatial cognition and computation*, and *spatial information theory* have established their foundational significance for the design and implementation of computational cognitive systems, and multimodal interaction & assistive technologies, e.g., especially in those areas where *processing and interpretation* of potentially large volumes of highly *dynamic spatio-temporal data* is involved [Bhatt, 2013]: cognitive vision & robotics, geospatial dynamics [Bhatt and Wallgrün, 2014], architecture design [Bhatt et al., 2014] to name a few prime examples.

Our research addresses ‘*space and spatio-temporal dynamics*’ from the viewpoints of visuo-spatial cognition and computation, computational cognitive linguistics, and formal representation and computational reasoning about space, action, and change. We especially focus on space and motion as interpreted within artificial intelligence and knowledge representation and reasoning (KR) in general, and *declarative spatial reasoning* [Bhatt et al., 2011, Schultz and Bhatt, 2012, Walega et al., 2015] in particular. Furthermore, the concept of *image schemas* as “*abstract recurring patterns of thought and perceptual experience*” [Johnson, 1990, Lakoff, 1990] serves a central role in our formal framework.

**Visuo-Spatial Dynamics of the Moving Image** *The Moving Image*, from the viewpoint of this paper, is interpreted 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. ■

As examples, consider the following:

- ▶ *cognitive studies of film* aimed at investigating attention and recipient effects in observers vis-a-vis the motion picture [Nannicelli and Taberham, 2014, Aldama, 2015]
- ▶ *evidence-based design* [Hamilton and Watkins, 2009, Cama, 2009] involving analysis of post-occupancy user behaviour in buildings, e.g., pertaining visual perception of signage
- ▶ *geospatial dynamics* aimed at human-centered interpretation of (potentially large-scale) geospatial satellite and remote sensing imagery [Bhatt and Wallgrün, 2014]
- ▶ *cognitive vision and control* in robotics, smart environments etc, e.g., involving human activity interpretation and real-time object / interaction tracking in professional and everyday living (e.g., meetings, surveillance and security at an airport) [Vernon, 2006, 2008, Dubba et al., 2011, Bhatt et al., 2013b, Spranger et al., 2014, Dubba et al., 2015].

Within all these areas, high-level semantic interpretation and qualitative analysis of the moving image requires the representational and inferential mediation of (declarative)

embodied, qualitative abstractions of the visuo-spatial dynamics, encompassing *space, time, motion, and interaction*.

**Declarative Model of Perceptual Narratives** With respect to a broad-based understanding of the moving image (as aforesaid), we define 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. ■

Declarativeness denotes the existence of grounded (e.g., symbolic, sub-symbolic) models coupled with **deep semantics** (e.g., for spatial and temporal knowledge) and systematic formalisation that can be used to perform reasoning and query answering, embodied simulation, and relational learning.<sup>4</sup> With respect to methods, this paper particularly alludes to declarative KR frameworks such as logic programming, constraint logic programming, description logic based spatio-terminological reasoning, answer-set programming based non-monotonic (spatial) reasoning, or even other specialised commonsense reasoners based on expressive action description languages for handling *space, action, and change*. Declarative representations serve as basis to externalise explicit and *inferred* knowledge, e.g., by way of modalities such as visual and diagrammatic representations, natural language, etc.

**Core Contributions.** We present a declarative model for the embodied grounding of the visuo-spatial dynamics of the moving image, and the ability to generate corresponding textual summaries that serve an analytical function from a computer-human interaction viewpoint in a range of cognitive assistive technologies and interaction system where reasoning about space, actions, change, and interaction is crucial. The overall framework encompasses:

**(F1).** a formal theory of qualitative characterisations of *space and motion* with deep semantics for spatial, temporal, and motion predicates

**(F2).** formalisation of the embodied *image schematic* structure of visuo-spatial dynamics wrt. the formal theory of space and motion

**(F3).** a declarative *spatio-temporal feature-based natural language generation engine* that can be used in a domain-independent manner

The overall framework (**F1–F3**) for the embodied grounding of the visuo-spatial dynamics of the moving image, and the externalisation of the declarative perceptual narrative model by way of natural language has been fully modelled and implemented in an elaboration tolerant manner 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 with spatio-temporal relations, image

<sup>4</sup> Broadly, we refer to methods for abstraction, analogy-hypothesis-theory formation, belief revision, argumentation.

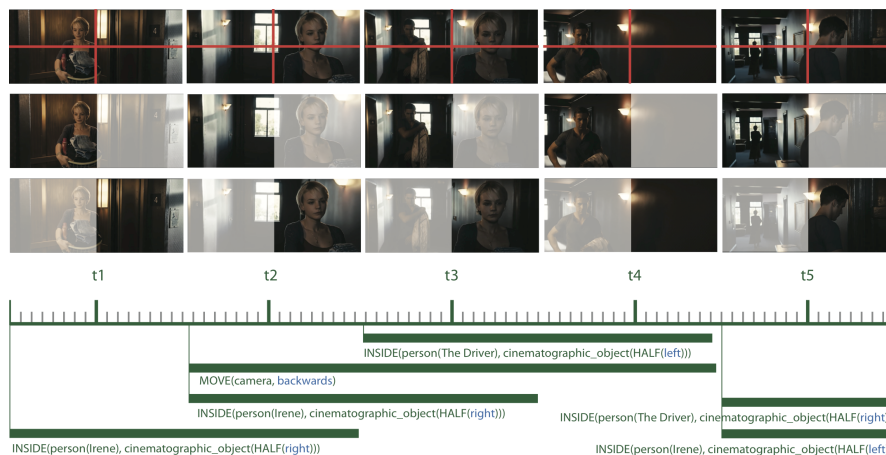


Fig. 1: Analysis based on the Quadrant system (Drive 2011)

schemas, path & motion predicates, syntax trees etc as first class objects within the CLP environment.

**Organization of the Paper.** Section 2 presents the application scenarios that we will directly demonstrate as case-studies in this paper; we focus on a class of cognitive interaction systems where the study of visuo-spatial dynamics in the context of the moving image is central. Sections 3–4 present the theory of space, motion, and image schemas elaborating on its formalisation and declarative implementation within constraint logic programming. Section 5 presents a summary of the declarative natural language generation component. Section 6 concludes with a discussion of related work.

## 2 TALKING ABOUT THE MOVING IMAGE

Talking about the moving image denotes:

the ability to computationally generate semantically well-founded, embodied, multi-modal (e.g., **natural language**, iconic, diagrammatic) **externalisations** of dynamic visuo-spatial phenomena as perceived via visuo-spatial, auditory, or sensorimotor haptic interactions. ■

In the backdrop of the twin notions of *the moving image & perceptual narratives* (Section 1), we focus on a range of computer-human interaction systems & assistive technologies at the interface of language, logic, and cognition; in particular, visuo-spatial cognition and computation are most central. Consider the case-studies in (S1–S4):<sup>5</sup>

<sup>5</sup> The paper is confined to visual processing and analysis, and ‘talking about it’ by way of natural language externalisations. We emphasise that our underlying model is general, and elaboration tolerant to other kinds of input features.

**(S1). COGNITIVE STUDIES OF FILM** Cognitive studies of the moving image—specifically, *cognitive film theory*—has accorded a special emphasis on the role of *mental activity of observers* (e.g., subjects, analysts, general viewers / spectators) as one of the most central objects of inquiry [Nannicelli and Taberham, 2014, Aldama, 2015] (e.g., expert analysis in Listing L1; Fig 1). Amongst other things, cognitive film studies concern making sense of subject’s visual fixation or saccadic eye-movement patterns whilst watching a film and correlating this with deep 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), studies in embodiment [Sobchack, 2004, Coegnarts and Kravanja, 2012].

#### DRIVE (2011) | QUADRANT SYSTEM. VISUAL ATTENTION.

Director. Nicolas Winding Refn

This short scene, involving [The Driver](#) ([Ryan Gosling](#)) and [Irene](#) ([Carey Mulligan](#)), adopts a TOP-BOTTOM and LEFT-RIGHT quadrant system that is executed in a [SINGLE TAKE](#) / without any [CUTS](#)

The [CAMERA](#) MOVES BACKWARD tracking the movement of [The Driver](#) and [Irene](#); DURING MOVEMENT-1, [Irene](#) OCCUPIES the right quadrant, WHILE [The Driver](#) OCCUPIES the LEFT quadrant

Spectator eye-tracking data suggests that the audience is repeatedly switching their attention between the LEFT and RIGHT quadrants, with a majority of the audience fixating visual attention on [Irene](#) as she MOVES into an extreme [CLOSE-UP SHOT](#)

Credit. Quadrant system method based on study by Tony Zhou.

L1

#### (S2). EVIDENCE BASED DESIGN (EBD) OF THE

**BUILT ENVIRONMENT** Evidence-based building design involves the study of the post-occupancy behaviour of building users with the aim to provide a scientific basis for generating best practice guidelines aimed at improving building performance and user experience. Amongst other things, this involves an analysis of the visuo-locomotive navigational experience of subjects based on eye-tracking and egocentric video capture based analysis of visual perception and attention, indoor people-movement analysis, e.g., during a wayfinding task, within a large-scale built-up environment such as a hospital or an airport (e.g., see Listing L2). EBD is typically pursued as an interdisciplinary endeavour—involving environmental psychologists, architects, technologists—toward the development of new tools and processes for data collection, qualitative analysis etc.

#### THE NEW PARKLAND HOSPITAL | WAYFINDING STUDY.

Location. Dallas, Texas

This experiment was conducted with 50 subjects at the New Parkland Hospital in Dallas

[Subject 21](#) ([Barbara](#)) performed a wayfinding task (#T5), STARTING FROM the reception desk of the emergency department and FINISHING AT the Anderson Pharmacy. Wayfinding task #5 GOES THROUGH the long corridor in the emergency department, the main reception and the blue elevators, going up to Level 2 INTO the Atrium Lobby, PASSING THROUGH the Anderson-Bridge, finally ARRIVING AT the X-pharmacy

Eye-tracking data and video data analysis suggests that [Barbara](#) fixated on passerby [Person.5](#) for two seconds as [Person.5](#) PASSES FROM her RIGHT IN the long corridor. [Barbara](#) fixated most ON the big blue elevator signage AT the main reception desk. DURING the 12th minute, video data from external GoPro cameras and egocentric video capture and eye-tracking suggest that [Barbara](#) looked indecisive (*stopped walking, looked around, performed rapid eye-movements*)

Credit. Based on joint work with Corgan Associates (Dallas)

L2

**(S3). GEOSPATIAL DYNAMICS** The ability of semantic and qualitative analytical capability to complement and synergize with statistical and quantitatively-driven methods has been recognized as important within geographic information systems. Research in geospatial dynamics [Bhatt and Wallgrün, 2014] investigates the theoretical foundations necessary to develop the computational capability for high-level commonsense, qualitative analysis of dynamic geospatial phenomena within next generation event and object-based GIS systems.

**(S4). HUMAN ACTIVITY INTERPRETATION** Research on embodied perception of vision —termed *cognitive vision* [Vernon, 2006, 2008, Bhatt et al., 2013b]— aims to enhance classical computer vision systems with cognitive abilities to obtain more robust vision systems that are able to adapt to unforeseen changes, make “narrative” sense of perceived data, and exhibit interpretation-guided goal directed behaviour. The long-term goal in cognitive vision is to provide general tools (integrating different aspects of space, action, and change) necessary for tasks such as real-time human activity interpretation and dynamic sensor (e.g., camera) control within the purview of vision, interaction, and robotics.

### 3 Space, Time, and Motion

Qualitative Spatial & Temporal Representation and Reasoning (QSTR) [Cohn and Hazarika, 2001] abstracts from an exact numerical representation by describing the relations between objects using a finite number of symbols. Qualitative representations use a set of relations that hold between objects to describe a scene. Galton [Galton, 1993, 1995, 2000] investigated movement on the basis of an integrated theory of space, time, objects, and position. Muller [Muller, 1998] defined continuous change using 4-dimensional regions in space-time. Hazarika and Cohn [Hazarika and Cohn, 2002] build on this work but used an interval based approach to represent spatio-temporal primitives.

We use spatio-temporal relations to represent and reason about different aspects of space, time, and motion in the context of visuo-spatial perception as described by [Suchan et al., 2014]. To describe the spatial configuration of a perceived scene and the dynamic changes within it we combine spatial calculi to a general theory for declaratively reason about spatio-temporal change. The domain independent theory of *Space*, *Time*, and *Motion* ( $\Sigma_{STM}$ ) consists of:

- ▶  $\Sigma_{Space}$  – Spatial Relations on topology, relative position, relative distance of spatial objects
- ▶  $\Sigma_{Time}$  – Temporal Relations for representing relations between time points and intervals
- ▶  $\Sigma_{Motion}$  – Motion Relations on changes of distance and size of spatial objects

The resulting theory is given as:  $\Sigma_{STM} \equiv_{def} [\Sigma_{Space} \cup \Sigma_{Time} \cup \Sigma_{Motion}]$ .

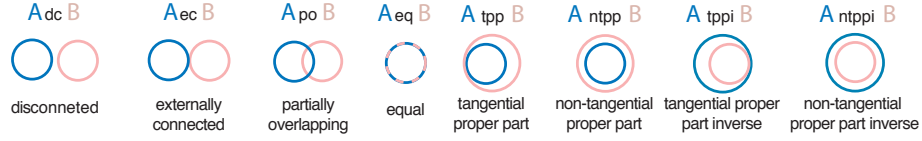


Fig. 2: Region Connection Calculus (RCC-8)

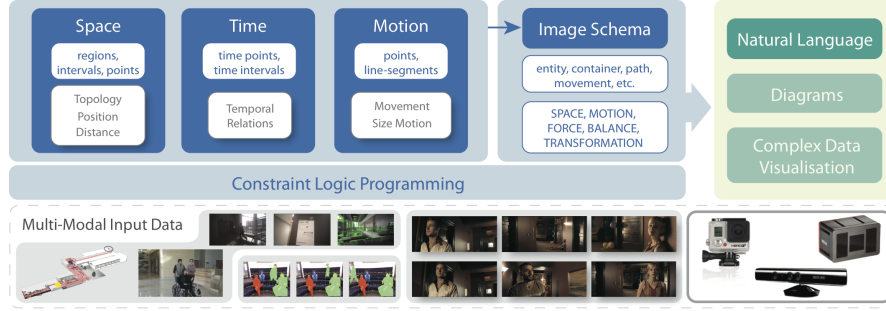


Fig. 3: General Theory of Space, Time, Motion, and Image Schema

Objects and individuals are represented as spatial primitives according to the nature of the spatial domain we are looking at, i.e., *regions of space*  $\mathcal{S} = \{s_1, s_2, \dots, s_n\}$ , *points*  $\mathcal{P} = \{p_1, p_2, \dots, p_n\}$ , and *line segments*  $\mathcal{L} = \{l_1, l_2, \dots, l_n\}$ . Towards this we use functions that map from the object or individual to the corresponding spatial primitive. The spatial configuration is represented using  $n$ -ary *spatial relations*  $\mathcal{R} = \{r_1, r_2, \dots, r_n\}$  of an arbitrary spatial calculus.  $\Phi = \{\phi_1, \phi_2, \dots, \phi_n\}$  is a set of propositional and functional fluents, e.g.  $\phi(e_1, e_2)$  denotes the spatial relationship between  $e_1$  and  $e_2$ . Temporal aspects are represented using *time points*  $\mathcal{T} = \{t_1, t_2, \dots, t_n\}$  and *time intervals*  $\mathcal{I} = \{i_1, i_2, \dots, i_n\}$ .  $Holds(\phi, r, at(t))$  is used to denote that the fluent  $\phi$  has the value  $r$  at time  $t$ . To denote that a relation holds for more than one contiguous time points, we define time intervals by its start and an end point, using  $between(t_1, t_2)$ .  $Occurs(\theta, at(t))$ , and  $Occurs(\theta, between(t_1, t_2))$  is used to denote that an event or action occurred.

### 3.1 $\Sigma_{\text{Space}}$ – Spatial Relations

The theory consists of spatial relations on objects, which includes relations on *topology* and *extrinsic orientation* in terms of left, right, above, below relations and depth relations (distance of spatial entity from the spectator).

► **Topology.** The Region Connection Calculus (RCC) [Cohn et al., 1997] is an approach to represent topological relations between regions in space. We use the RCC8 subset of the RCC, which consists of the eight base relations in  $\mathcal{R}_{\text{top}}$  (Figure 2), for representing regions of perceived objects, e.g. the projection on an object on the image plan.

$$\mathcal{R}_{\text{top}} \equiv \{\text{dc, ec, po, eq, tpp, ntpp, tpp}^{-1}, \text{ntpp}^{-1}\}$$

► **Relative Position.** We represent the position of two spatial entities, with respect to the observer’s viewpoint, using a 3-Dimensional representation that resemble Allen’s interval algebra [Allen, 1983] for each dimension, i.e. *vertical*, *horizontal*, and *depth* (distance from the observer).  $\mathcal{R}_{\text{pos}} \equiv [\mathcal{R}_{\text{pos-v}} \cup \mathcal{R}_{\text{pos-h}} \cup \mathcal{R}_{\text{pos-d}}]$

$$\mathcal{R}_{\text{pos-v}} \equiv \{\text{above, overlaps\_above, along\_above, vertically\_equal, overlaps\_below, along\_below, below}\}$$

$$\mathcal{R}_{\text{pos-h}} \equiv \{\text{left, overlaps\_left, along\_left, horizontally\_equal, overlaps\_right, along\_right, right}\}$$

$$\mathcal{R}_{\text{pos-d}} \equiv \{\text{closer, overlaps\_closer, along\_closer, distance\_equal, overlaps\_further, along\_further, further}\}$$

► **Relative Distance.** We represent the relative distance between two points  $p_1$  and  $p_2$  with respect to a third point  $p_3$ , using ternary relations  $\mathcal{R}_{\text{dist}}$ .

$$\mathcal{R}_{\text{dist}} \equiv \{\text{closer, further, same}\}$$

► **Relative Size.** For comparison of the size of two regions we use the relations in  $\mathcal{R}_{\text{size}}$ .

$$\mathcal{R}_{\text{size}} \equiv \{\text{smaller, bigger, same}\}$$

### 3.2 $\Sigma_{\text{Time}}$ – Temporal Relations

Temporal relations are used to represent the relationship between actions and events, e.g. one action happened before another action. We use the extensions of Allen’s interval relations [Allen, 1983] as described by [Vilain, 1982], i.e. these consist of relations between time *points*, *intervals*, and *point - interval*.

$$\mathcal{R}_{\text{point}} \equiv \{\bullet\text{before}\bullet, \bullet\text{after}\bullet, \bullet\text{equals}\bullet\}$$

$$\mathcal{R}_{\text{interval}} \equiv \{\text{before, after, during, contains, starts, started\_by, finishes, finished\_by, overlaps, overlapped\_by, meets, met\_by, equal}\}$$

$$\mathcal{R}_{\text{point-interval}} \equiv \{\bullet\text{before, after}\bullet, \bullet\text{starts, started\_by}\bullet, \bullet\text{during, contains}\bullet, \bullet\text{finishes, finished\_by}\bullet, \bullet\text{after, before}\bullet\}$$

The relations used for temporal representation of actions and events are the union of these three, i.e.  $\mathcal{R}_{\text{Time}} \equiv [\mathcal{R}_{\text{point}} \cup \mathcal{R}_{\text{interval}} \cup \mathcal{R}_{\text{point-interval}}]$ .

### 3.3 $\Sigma_{\text{Motion}}$ – Qualitative Spatial Dynamics

Spatial relations holding for perceived spatial objects change as an result of motion of the individuals in the scene. To account for this, we define motion relations by making qualitative distinctions of the changes in the parameters of the objects, i.e. the distance between two depth profiles and its size.

► **Relative Movement.** The relative movement of pairs of spatial objects is represented in terms of changes in the distance between two points representing the objects.



$\mathcal{R}_{\text{move}} \equiv \{\text{approaching, receding, static}\}$

► **Size Motion.** For representing changes in size of objects, we consider relations on each dimension (*horizontal*, *vertical*, and *depth*) separately. Changes on more than one of these parameters at the same time instant can be represented by combinations of the relations.

$\mathcal{R}_{\text{size}} \equiv \{\text{elongating, shortening, static}\}$

## 4 Image Schemas of the Moving Image

Table 1: *Image Schemas* identifiable in the literature (non-exhaustive list)

SPACE	ABOVE, ACROSS, COVERING, CONTACT, VERTICAL_ORIENTATION, LENGTH
MOTION	CONTAINMENT_PATH, PATH_GOAL, SOURCE_PATH_GOAL, BLOCKAGE, CENTER_PERIPHERY, CYCLE, CYCLIC_CLIMAX
FORCE	COMPULSION, COUNTERFORCE, DIVERSION, REMOVAL_OF_RESTRAINT / ENABLEMENT, ATTRACTION, LINK, SCALE
BALANCE	AXIS_BALANCE, POINT_BALANCE, TWIN_PAN_BALANCE, EQUILIBRIUM
TRANSFORMATION	LINEAR_PATH_FROM_MOVING_OBJECT, PATH_TO_ENDPOINT, PATH_TO_OBJECT_MASS, MULTIPLEX_TO_MASS, REFLEXIVE, ROTATION
OTHERS	SURFACE, FULL-EMPTY, MERGING, MATCHING, NEAR-FAR, MASS-COUNT, ITERATION, OBJECT SPLITTING, PART-WHOLE, SUPERIMPOSITION, PROCESS, COLLECTION

Image schemas have been a cornerstone in cognitive linguistics [Geeraerts and Cuyckens, 2007], and have also been investigated from the perspective of psycholinguistics, and language and cognitive development [Mandler, 1992, Mandler and Pagán Cánovas, 2014]. Image schemas, as embodied structures founded on experiences of interactions with the world, serve as the ideal framework for understanding and reasoning about perceived visuo-spatial dynamics, e.g., via generic conceptualisation of space, motion, force, balance, transformation, etc. Table 1 presents a non-exhaustive list of image schemas identifiable in the literature. We formalise image schemas on individuals, objects and actions of the domain, and ground them in the spatio-temporal dynamics, as defined in Section 3, that are underling the particular schema. As examples, we focus on the spatial entities PATH, CONTAINER, THING, the spatial relation CONTACT, and movement relations MOVE, INTO, OUT OF (these being regarded as highly important and foundational from the viewpoint of cognitive development [Mandler and Pagán Cánovas, 2014]).

**CONTAINMENT** The CONTAINMENT schema denotes, that an object or an individual is inside of a container object.

```
containment(entity(E), container(C)) :- inside(E, C).
```

As an example consider the following description from the film domain described in Listing L1.

```
Irene OCCUPIES the RIGHT QUADRANT, WHILE The Driver OCCUPIES the LEFT QUADRANT.
```

In the movie example the ENTITY is a person in the film, namely The Driver, and the CONTAINER is a cinematographic object, the top-left quadrant, which is used to analyse the composition of the scene. We are defining the inside relation based on the involved individuals and objects, e.g. in this case we define the topological relationship between The Drivers face and the bottom-right quadrant.

```
inside(person(P), cinemat_object(quadrant(Q)) :-
  region(person(P), P_region),
  region(cinemat_object(quadrant(Q)), Q_region)
  topology(nttp, P_region, Q_region).
```

To decide on the words to use for describing the schema, we make distinctions on the involved entities and the spatial characteristics of the scene, e.g. we use the word 'occupies', when the person is taking up the whole space of the container, i.e. the size is bigger than a certain threshold.

```
phrase(containment(E, C), [E, 'occupy', C]) :-
  region(person(E), E_region),
  region(cinemat_object(quadrant(C)), C_region),
  threshold(C_region, C_tresh),
  size(bigger, E_region, C_tresh).
```

Similarly, we choose the word 'in', when the person is fully contained in the quadrant.

**PATH\_GOAL and SOURCE\_PATH\_GOAL** The PATH\_GOAL Image Schema is used to conceptualise the movement of an *object* or an *individual*, towards a *goal* location, on a particular *path*. In this case, the path is the directed movement towards the goal. The SOURCE\_PATH\_GOAL Schema builds on the PATH\_GOAL Schema by adding a source to it. Both Schemas are used to describe movement, however, in the first case, the source is not important, only the goal of the movement is of interest. Here we only describe the SOURCE\_PATH\_GOAL Schema in more detail, as the PATH Schema is the same, without the source in it.

```
source_path_goal(Trajector, Source, Path, Goal) :-
  entity(Trajector), location(Source), location(Goal),
  path(Path, Source, Goal),
  at_location(Trajector, Source, at_time(T_1)),
  at_location(Trajector, Goal, at_time(T_2)),
  move(Trajector, Path, between(T_1, T_2)).
```

In the way finding analysis one example of the SOURCE.PATH.GOAL schema is when a description of the path a subject was walking is generated.

Barbara WALKS FROM the EMERGENCY, THROUGH the ATRIUM LOBBY TO the BLUE ELEVATORS.

Another example is when a descriptions of a subjects eye movement is generated from the eye tracking experiment.

Barbaras eyes MOVE FROM the EMERGENCY SIGN, OVER the EXIT SIGN TO the ELEVATOR SIGN.

In both of these sentences there is a moving entity, the *trajector*, a *source* and a *goal* location, and a *path* connecting the source and the goal. In the first sentence it is Barbara who is moving, while in the second sentence Barbaras eyes are moving. Based on the different spatial entities involved in the movement, we need different definitions of locations, path, and the moving actions. In the way finding domain, a subject is at a location when the position of the person upon a 2-dimensional floorplan is inside the region denoting the location, e.g. a room, a corridor, or any spatial artefact describing a region in the floorplan.

```
at_location(Subject, Location) :-
  person(Subject), room(Location),
  position(Subject, S_pos), region(Location, L_reg),
  topology(ntpp, S_pos, Loc_reg).
```

Possible paths between the locations of a floorplan are represented by a topological route graph, on which the subject is walking.

```
move(person(Subject), Path) :-
  action(movement(walk), Subject, Path),
  movement(approaching, Subject, Goal).
```

For generating language, we have to take the type of the trajector into account, as well as the involved movement and the locations, e.g. the eyes are moving 'over' some objects, but Barbara moves 'through' the corridor.

**ATTRACTION** The ATTRACTION schema is expressing a force by which an entity is attracted.

```
attraction(Subject, Entity) :-
    entity(Subject), entity(Entity),
    force(attraction, Subject, Entity).
```

An example for ATTRACTION is the eye tracking experiment, when the attention of a subject is attracted by some object in the environment.

While walking THROUGH the HALLWAY, Barbaras attention is attracted by the OUTSIDE VIEW.

In this case the entity is Barbara's attention which is represented by the eye tracking data, and it is attracted by the force, the outside view applies on it. We define attraction by the fact, that the gaze position of Barbara has been on the outside for a substantial amount of time, however, this definition can be adapted to the needs of domain experts, e.g. architects who want to know what are the things that grab the attention of people in a building.

## 5 From Perceptual Narratives to Natural Language

The design and implementation of the natural language generation component has been driven by three key developmental goals: (1) ensuring support for, and uniformity with respect to the (deep) representational semantics of space and motion relations etc (Section 3); (2) development of modular, yet tightly integrated set of components that can be easily used within the state-of-the-art (constraint) logic programming family of KR methods; and (3) providing seamless integration capabilities within hybrid AI and computational cognition systems.

### System Overview (NL Generation)

The overall pipeline of the language generation component follows a standard natural language generation system architecture [Reiter and Dale, 2000, Bateman and Zock, 2003]. Figure 4 illustrates the system architecture encompassing the typical stages of content determination & result structuring, linguistic & syntactic realisation, and syntax tree & sentence generation.

**S1. Input – Interaction Description Schema** Interfacing with the language generator is possible with a generic (activity-theoretic) Interaction Description Schema (IDS) that is founded on the ontology of the (declarative) perceptual narrative, and a general set of constructs to introduce the domain-specific vocabulary. Instances of the IDS constitute the domain-specific input data for the generator.

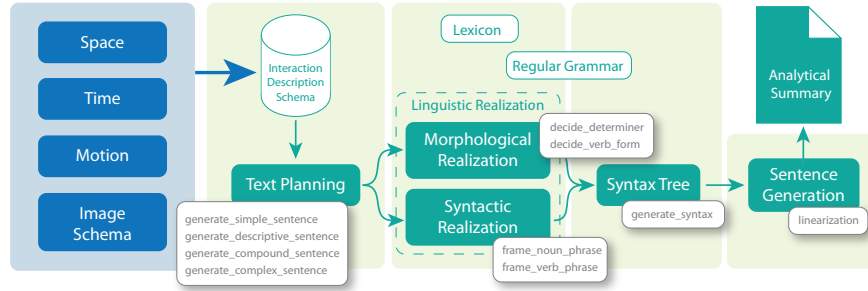


Fig. 4: From Perceptual Narratives to Natural Language

**S2. Syntax Tree and Sentence Generation** The generator consists of sub-modules concerned with input IDS instance to text planning, morphological & syntactic realisation, and syntax tree & sentence generation. Currently, the generator functions in a *single interaction* mode where each invocation of the system (with an input instance of the IDS) produces a single sentence in order to produce spatio-temporal domain-based text. The morphological and syntactic realisation module brings in assertions of detailed grammatical knowledge and the lexicon that needs to be encapsulated for morphological realisation; this encompasses aspects such as noun and verb categories, spatial relations and locations; part of speech identification is also performed at this stage, including determiner and adjective selection, selection of verb and tense etc. The parts of speech identified by the morph analyser taken together with context free grammar rules for simple, complex, and compound sentence constructions are used for syntactic realisation, and sentence generation.

### Language Generation (Done Declaratively)

Each aspect of generation process, be it at a factual level (grammar, lexicon, input data) or at a process level (realisation, syntax tree generation) is fully *declarative* (to the extent possible in logic programming) and *elaboration tolerant* (i.e., addition or removal of facts & rules, constraints etc does not break down the generation process). An important consequence of this level of declarativeness is that a query can work both ways: from input data to syntax tree to sentence, or from a sentence back to its syntax tree and linguistic decomposition wrt. to a specific lexicon.

### Empirical Evaluation of Language Generation

We tested the language generation component with data for 25 subjects, 500 IDS instances, and 53 domain facts (using an Intel Core i7-3630QM CPU @ 2.40GHz x 8). We generated summaries in simple/continuous present, past, future respectively for all IDS instances. Table (2): (a). average of 20 interactions, on an average 26.2 sentences / summary, with 17.6 tokens as the average length / sentence; (b) generated 100 sentences for simple, compound, and complex types reflecting the average sentence generation time.

Table 2: Time (in ms) for (a) summaries, (b) sentences

<b>Tense</b>	<b>Avg. Min. Max.</b>			<b>Type</b>	<b>Time</b>
simple	77.8	70	96	simple	0,52
continous	84.48	73	99	compound	1,23
				complex	1,32

(a)

(b)

## 6 DISCUSSION AND RELATED WORK

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 [Vernon, 2006, 2008, Dubba et al., 2011, Bhatt et al., 2013b, Spranger et al., 2014, Dubba et al., 2015]. There has also been an increased interest from the computer vision community to synergise with cognitively motivated methods for language grounding and inference with visual imagery [Karpathy and Fei-Fei, 2015, Yu et al., 2015]. This paper has not attempted to present advances in basic computer vision research; in general, this is not the agenda of our research even outside the scope of this paper. The low-level visual processing algorithms that we utilise are founded in state-of-the-art outcomes from the computer vision community for detection and tracking of *people, objects, and motion* [Canny, 1986, Lucas and Kanade, 1981, Viola and Jones, 2001, Dalal and Triggs, 2005].<sup>6</sup> On the language front, the number of research projects addressing natural language generation systems [Reiter and Dale, 2000, Bateman and Zock, 2003] is overwhelming; there exist a plethora of projects and initiatives focussing on language generation in general or specific contexts, candidate examples being the works in the context of *weather report* generation [Goldberg et al., 1994, Sripada et al., 2014], *Pollen* forecasts [Turner et al., 2006].<sup>7</sup> Our focus on the (declarative) language generation component of the framework of this paper (Section 5) has been on the use of “deep semantics” for space and motion, and to have a unified framework –with each aspect of the embodied perception grounding framework– fully implemented within constraint logic programming.

Our research is motivated by computational cognitive systems concerned with interpreting multimodal dynamic perceptual input; in this context, we believe that it is essential to build systematic methods and tools for embodied visuo-spatial conception, formalisation, and computation with primitives of space and motion. Toward this, this paper has developed a declarative framework for embodied grounding and natural language based analytical summarisation of the moving image; the implemented model

<sup>6</sup> For instance, we analyse motion in a scene sparse and dense optical flow [Lucas and Kanade, 1981, Farnebäck, 2003], detecting faces using cascades of features [Viola and Jones, 2001], detecting humans using histograms of oriented gradients [Dalal and Triggs, 2005].

<sup>7</sup> We have been unable to locate a fitting & comparable spatio-temporal feature sensitive language generation module for open-source usage. We will disseminate our language generation component as an open-source PROLOG library.

consists of modularly built components for logic-based representation and reasoning about qualitative and linguistically motivated abstractions about space, motion, and image schemas. Our 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 sensemaking [Bhatt et al., 2013a, Bhatt, 2013] of dynamic visuo-spatial imagery is central.

### **Acknowledgements**

We acknowledge the contributions of DesignSpace members Saurabh Goyal, Giulio Carducci, John Sutton, and Vasiliki Kondyli in supporting developmental, design, experimentation, and expert (qualitative) analysis tasks.

## Bibliography

- F. L. Aldama. The Science of Storytelling: Perspectives from Cognitive Science, Neuroscience, and the Humanities. *Projections*, 9(1):80–95, 2015. doi: doi:10.3167/proj.2015.090106.
- J. F. Allen. Maintaining knowledge about temporal intervals. *Commun. ACM*, 26(11): 832–843, 1983. ISSN 0001-0782.
- J. Bateman and M. Zock. Natural language generation. *Oxford handbook of computational linguistics*, pages 284–304, 2003.
- J. A. Bateman. Situating spatial language and the role of ontology: Issues and outlook. *Language and Linguistics Compass*, 4(8):639–664, 2010. doi: 10.1111/j.1749-818X.2010.00226.x.
- M. Bhatt. Reasoning about Space, Actions and Change: A Paradigm for Applications of Spatial Reasoning. In *Qualitative Spatial Representation and Reasoning: Trends and Future Directions*. IGI Global, USA, 2012. ISBN ISBN13: 9781616928681.
- M. Bhatt. Between Sense and Sensibility: Declarative narrativisation of mental models as a basis and benchmark for visuo-spatial cognition and computation focussed collaborative cognitive systems. *CoRR*, abs/1307.3040, 2013.
- M. Bhatt and J. O. Wallgrün. Geospatial narratives and their spatio-temporal dynamics: Commonsense reasoning for high-level analyses in geographic information systems. *ISPRS Int. J. Geo-Information*, 3(1):166–205, 2014. doi: 10.3390/ijgi3010166. URL <http://dx.doi.org/10.3390/ijgi3010166>.
- M. Bhatt, J. H. Lee, and C. P. L. Schultz. CLP(QS): A declarative spatial reasoning framework. In *Spatial Information Theory - 10th International Conference, COSIT 2011, Belfast, ME, USA, September 12-16, 2011. Proceedings*, volume 6899 of *Lecture Notes in Computer Science*, pages 210–230. Springer, 2011. doi: 10.1007/978-3-642-23196-4\_12. URL [http://dx.doi.org/10.1007/978-3-642-23196-4\\_12](http://dx.doi.org/10.1007/978-3-642-23196-4_12).
- M. Bhatt, C. Schultz, and C. Freksa. The ‘Space’ in Spatial Assistance Systems: Conception, Formalisation and Computation. In T. Tenbrink, J. Wiener, and C. Claramunt, editors, *Representing space in cognition: Interrelations of behavior, language, and formal models. Series: Explorations in Language and Space*, Explorations in Language and Space. 978-0-19-967991-1, Oxford University Press, 2013a. ISBN 9780199679911.
- M. Bhatt, J. Suchan, and C. P. L. Schultz. Cognitive interpretation of everyday activities - toward perceptual narrative based visuo-spatial scene interpretation. In M. A. Finlayson, B. Fisseni, B. Löwe, and J. C. Meister, editors, *2013 Workshop on Computational Models of Narrative, CMN 2013, August 4-6, 2013, Hamburg, Germany*, volume 32 of *OASICS*, pages 24–29. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2013b. ISBN 978-3-939897-57-6. doi: 10.4230/OASICS.CMN.2013.24.
- M. Bhatt, C. P. L. Schultz, and M. Thosar. Computing narratives of cognitive user experience for building design analysis: KR for industry scale computer-aided architecture design. In C. Baral, G. D. Giacomo, and T. Eiter, editors, *Principles of*



- Knowledge Representation and Reasoning: Proceedings of the Fourteenth International Conference, KR 2014, Vienna, Austria, July 20-24, 2014*. AAAI Press, 2014. ISBN 978-1-57735-657-8.
- R. Cama. *Evidence-Based Healthcare Design*. Wiley, 2009. ISBN 9780470149423.
- J. Canny. A computational approach to edge detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-8(6):679–698, Nov 1986. ISSN 0162-8828. doi: 10.1109/TPAMI.1986.4767851.
- M. Coegnarts and P. Kravanja. Embodied Visual Meaning: Image Schemas in Film. *Projections*, 6(2):84–101, 2012. doi: doi:10.3167/proj.2012.060206.
- A. Cohn and S. Hazarika. Qualitative spatial representation and reasoning: An overview. *Fundam. Inf.*, 46(1-2):1–29, 2001. ISSN 0169-2968.
- A. Cohn, B. Bennett, J. Gooday, and N. Gotts. Representing and reasoning with qualitative spatial relations about regions. In O. Stock, editor, *Spatial and Temporal Reasoning*, pages 97–134. Kluwer Academic Publishers, Dordrecht, 1997.
- N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893 vol. 1, June 2005. doi: 10.1109/CVPR.2005.177.
- K. S. R. Dubba, M. Bhatt, F. Dylla, D. C. Hogg, and A. G. Cohn. Interleaved inductive-abductive reasoning for learning complex event models. In *Inductive Logic Programming - 21st International Conference, ILP 2011, Windsor Great Park, UK, July 31 - August 3, 2011, Revised Selected Papers*, volume 7207 of *Lecture Notes in Computer Science*, pages 113–129. Springer, 2011. doi: 10.1007/978-3-642-31951-8\_14.
- K. S. R. Dubba, A. G. Cohn, D. C. Hogg, M. Bhatt, and F. Dylla. Learning relational event models from video. *J. Artif. Intell. Res. (JAIR)*, 53:41–90, 2015. doi: 10.1613/jair.4395. URL <http://dx.doi.org/10.1613/jair.4395>.
- G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Proceedings of the 13th Scandinavian Conference on Image Analysis, SCIA'03*, pages 363–370, Berlin, Heidelberg, 2003. Springer-Verlag. ISBN 3-540-40601-8. URL <http://dl.acm.org/citation.cfm?id=1763974.1764031>.
- C. Freksa. Spatial cognition: An AI perspective. In *Proceedings of the 16th European Conference on Artificial Intelligence, ECAI'2004, including Prestigious Applicants of Intelligent Systems, PAIS 2004, Valencia, Spain, August 22-27, 2004*, pages 1122–1128. IOS Press, 2004.
- A. Galton. Towards an integrated logic of space, time and motion. In *IJCAI*, pages 1550–1557, 1993.
- A. Galton. Towards a qualitative theory of movement. In A. U. Frank and W. Kuhn, editors, *Spatial Information Theory - A Theoretical Basis for GIS (COSIT'95)*, pages 377–396. Springer, Berlin, Heidelberg, 1995.
- A. Galton. *Qualitative Spatial Change*. Oxford University Press, 2000. ISBN 0198233973.
- D. Geeraerts and H. Cuyckens. *The Oxford Handbook of Cognitive Linguistics*. Oxford Handbooks. Oxford University Press, USA, 2007. ISBN 9780198032885.
- E. Goldberg, N. Driedger, and R. I. Kittredge. Using natural-language processing to produce weather forecasts. *IEEE Expert: Intelligent Systems and Their Applications*, 9(2):45–53, Apr. 1994. ISSN 0885-9000. doi: 10.1109/64.294135.

- D. Hamilton and D. Watkins. *Evidence-Based Design for Multiple Building Types*. Wiley, 2009. ISBN 9780470129340.
- S. M. Hazarika and A. G. Cohn. Abducing qualitative spatio-temporal histories from partial observations. In *KR*, pages 14–25, 2002.
- M. Johnson. *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. Philosophy, psychology, cognitive sciences. University of Chicago Press, 1990. ISBN 9780226403182.
- A. Karpathy and L. Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Columbus, Boston, USA*. IEEE, 2015.
- G. Lakoff. *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. Cognitive science, linguistics, philosophy. University of Chicago Press, 1990. ISBN 9780226468044.
- B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. pages 674–679, 1981.
- J. M. Mandler. How to build a baby: Ii. conceptual primitives. *Psychological Review*, pages 587–604, 1992.
- J. M. Mandler and C. Pagán Cánovas. On Defining Image Schemas. *Language and Cognition*, 6:510–532, 12 2014. ISSN 1866-9859. doi: 10.1017/langcog.2014.14.
- K. Mix, L. Smith, and M. Gasser. *The Spatial Foundations of Cognition and Language: Thinking Through Space*. Explorations in Language and Space. OUP Oxford, 2009. ISBN 9780199553242.
- P. Muller. A qualitative theory of motion based on spatio-temporal primitives. In A. G. Cohn, L. K. Schubert, and S. C. Shapiro, editors, *Proceedings of the Sixth International Conference on Principles of Knowledge Representation and Reasoning (KR'98), Trento, Italy, June 2-5, 1998*, pages 131–143. Morgan Kaufmann, 1998.
- T. Nannicelli and P. Taberham. Contemporary cognitive media theory. In T. Nannicelli and P. Taberham, editors, *Cognitive Media Theory*, AFI Film Readers. Routledge, 2014. ISBN 978-0-415-62986-7.
- E. Reiter and R. Dale. *Building Natural Language Generation Systems*. Cambridge University Press, Cambridge, U.K., 2000.
- C. P. L. Schultz and M. Bhatt. Towards a declarative spatial reasoning system. In *ECAI 2012 - 20th European Conference on Artificial Intelligence. Including Prestigious Applications of Artificial Intelligence (PAIS-2012) System Demonstrations Track, Montpellier, France, August 27-31, 2012*, volume 242 of *Frontiers in Artificial Intelligence and Applications*, pages 925–926. IOS Press, 2012. doi: 10.3233/978-1-61499-098-7-925.
- V. Sobchack. *Carnal Thoughts: Embodiment and Moving Image Culture*. University of California Press, November 2004. ISBN 0520241290.
- M. Spranger, J. Suchan, M. Bhatt, and M. Epe. Grounding dynamic spatial relations for embodied (robot) interaction. In *PRICAI 2014: Trends in Artificial Intelligence - 13th Pacific Rim International Conference on Artificial Intelligence, Gold Coast, QLD, Australia, December 1-5, 2014. Proceedings*, volume 8862, pages 958–971. Springer, 2014. doi: 10.1007/978-3-319-13560-1\_83.
- S. Sripada, N. Burnett, R. Turner, J. Mastin, and D. Evans. A case study: Nlg meeting weather industry demand for quality and quantity of textual weather forecasts.

- In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 1–5, Philadelphia, Pennsylvania, U.S.A., June 2014. Association for Computational Linguistics.
- J. Suchan, M. Bhatt, and P. E. Santos. Perceptual narratives of space and motion for semantic interpretation of visual data. In L. de Agapito, M. M. Bronstein, and C. Rother, editors, *Computer Vision - ECCV 2014 Workshops - Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part II*, volume 8926 of *Lecture Notes in Computer Science*, pages 339–354. Springer, 2014. doi: 10.1007/978-3-319-16181-5\_24. URL [http://dx.doi.org/10.1007/978-3-319-16181-5\\_24](http://dx.doi.org/10.1007/978-3-319-16181-5_24).
- R. Turner, S. Sripada, E. Reiter, and I. P. Davy. Generating Spatio-temporal Descriptions in Pollen Forecasts. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, EACL '06, pages 163–166, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.
- B. Tversky. Narratives of Space, Time, and Life. *Mind & Language*, 19(4):380–392, 2004. doi: 10.1111/j.0268-1064.2004.00264.x.
- B. Tversky. Visuospatial Reasoning. In K. J. Holyoak and R. G. Morrison, editors, *The Cambridge handbook of thinking and reasoning*, chapter 10, pages 209–240. Cambridge University Press, NY, 2005. doi: 10.2277/0521531012.
- D. Vernon. The space of cognitive vision. In H. I. Christensen and H.-H. Nagel, editors, *Cognitive Vision Systems*, volume 3948 of *Lecture Notes in Computer Science*, pages 7–24. Springer, 2006. ISBN 978-3-540-33971-7.
- D. Vernon. Cognitive vision: The case for embodied perception. *Image Vision Comput.*, 26(1):127–140, 2008.
- M. B. Vilain. A system for reasoning about time. In D. L. Waltz, editor, *Proceedings of the National Conference on Artificial Intelligence. Pittsburgh, PA, August 18-20, 1982.*, pages 197–201. AAAI Press, 1982. URL <http://www.aaai.org/Library/AAAI/1982/aaai82-047.php>.
- P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), with CD-ROM, 8-14 December 2001, Kauai, HI, USA*, pages 511–518. IEEE Computer Society, 2001. doi: 10.1109/CVPR.2001.990517. URL <http://doi.ieeecomputersociety.org/10.1109/CVPR.2001.990517>.
- P. Walega, M. Bhatt, and C. Schultz. ASPMT(QS): Non-Monotonic Spatial Reasoning with Answer Set Programming Modulo Theories. In *LPNMR: Logic Programming and Nonmonotonic Reasoning - 13th International Conference*, 2015.
- D. Waller and L. Nadel. *Handbook of Spatial Cognition*. American Psychological Association (APA), 2013. ISBN 978-1-4338-1204-0.
- H. Yu, N. Siddharth, A. Barbu, and J. M. Siskind. A Compositional Framework for Grounding Language Inference, Generation, and Acquisition in Video. *J. Artif. Intell. Res. (JAIR)*, 52:601–713, 2015. doi: 10.1613/jair.4556.