Perceptual Narratives of Space and Motion for Semantic Interpretation of Visual Data

Jakob Suchan[1(✉)], Mehul Bhatt[1], and Paulo E. Santos[2]

[1] Cognitive Systems, University of Bremen, Bremen, Germany
jsuchan@informatik.uni-bremen.de
[2] Centro Universitario da FEI, São Paulo, Brazil

Abstract. We propose a commonsense theory of *space* and *motion* for the high-level semantic interpretation of dynamic scenes. The theory provides primitives for commonsense representation and reasoning with *qualitative spatial relations, depth profiles*, and *spatio-temporal change*; these may be combined with probabilistic methods for modelling and hypothesising event and object relations. The proposed framework has been implemented as a general activity abstraction and reasoning engine, which we demonstrate by generating declaratively grounded visuo-spatial narratives of perceptual input from vision and depth sensors for a benchmark scenario.

Our long-term goal 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 control within the purview of cognitive vision, interaction, and control.

1 Introduction

Systems that monitor and interact with an environment populated by humans and other artefacts require a formal means for representing and reasoning about spatio-temporal, event and action based phenomena that are grounded to real public and private scenarios (e.g., logistical processes, activities of everyday living) of the environment being modelled. A fundamental requirement within such application domains is the need to explicitly represent and reason about dynamic spatial configurations or scenes and, for real world problems, integrated reasoning about space, actions, and change [1]. With these modelling primitives, the ability to perform *predictive* and *explanatory* analyses on the basis of sensory data is crucial for creating a useful intelligent function within such environments.

Commonsense, Space, Change. Qualitative Spatial & Temporal Representation and Reasoning (QSTR) provide a commonsensical interface to abstract and reason about quantitative spatial information [2]. *Qualitative spatial / temporal calculi* are relational-algebraic systems pertaining to one or more aspects of space such as *topology, orientation, direction, size* [3].

The integration of qualitative spatial representation and reasoning techniques within general commonsense reasoning frameworks in AI is an essential next-step

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 339–354, 2015.
DOI: 10.1007/978-3-319-16181-5_24

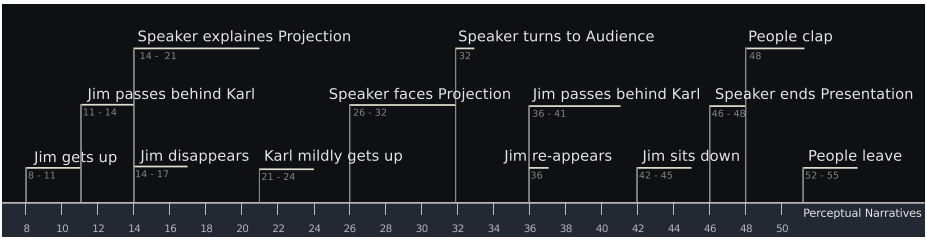


Fig. 1. Semantic Interpretation by Perceptual Narrativisation

for their applicability toward tasks such as spatial planning, spatio-temporal diagnosis and abnormality detection, event recognition and behaviour interpretation [4]. CLP(QS) [5] provides a framework for declarative spatial reasoning.

Perceptual Narratives [6] are declarative models of visual, auditory, haptic and other observations in the real world that are obtained via artificial sensors and / or human input. As an example, consider the *smart meeting cinematography* domain, where *perceptual narratives* as in Fig. 1 are generated based on perceived spatial change interpreted as interactions of humans in the environment. Such narratives explaining the ongoing activities are needed to anticipate changes in the environment, as well as to appropriately influence the real-time control of the camera system.

We suggest that the **semantic interpretation** of activities from video, depth (e.g., time-of-flight devices such as Kinect), and other forms of sensory input requires the representational and inferential mediation of qualitative abstractions of *space, action, and change* [1]. Generation of perceptual narratives, and their access via the declarative interface of logic programming facilitates the integration of the overall framework in bigger projects concerned with cognitive vision, robotics, hybrid-intelligent systems etc.

The particular focus and contributions of this paper are: (*a*) *Space and motion*: declaratively reasoning about qualitative spatial relations (e.g., topology, orientation), and motion in the context of everyday activities involving humans and artefacts (*b*) *Hybridisation*: integrating the qualitative theory with a probabilistic method for hypothesising object relations (*c*) *Semantic characterisation*: as a result of (a) and (b), generation of declarative narratives of perceptual RGB-D data that is obtained directly from people/object tracking algorithms.

2 Related Work

The core emphasis in activity and behaviour recognition has been on supervised learning algorithms requiring preprocessed (e.g., annotated) datasets from sensory streams. Unsupervised methods have received recent attention, with hybrid models integrating machine learning techniques with high-level structured representation and reasoning gaining recent momentum. The literature review below

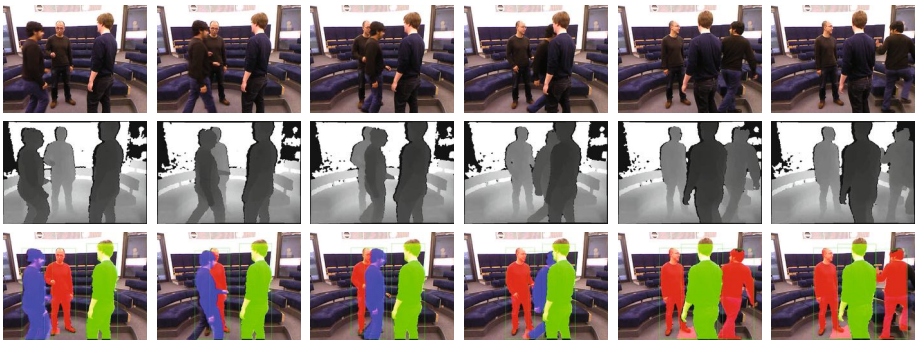


Fig. 2. Activity Sequence: *passing in-between people*, corresponding RGB and Depth profile data

concentrates on proposals concerned with the main aspects of the investigation reported in the present paper, namely, the high-level interpretation of events from the standpoint of Qualitative Spatial & Temporal Representation and Reasoning (QSTR). General reviews of work on activity and behaviour recognition can be found in [7–9].

2.1 Scene Interpretation

Research on scene interpretation has been largely based on probabilistic methods, motivated by the need to deal with sensor noise and image uncertainty [7], leaving aside the representation of general facts about the domain and the interplay between this representation and the actual interpretation of the scenes. Logic-based image interpretation, on the other hand, tackles the problem from the viewpoint of effective representation of general facts about the domain, as well as the generalisation of these facts to problems with infinite variables. Close to the topic of this paper, dos Santos et al. [10] presents a formalism for interpreting events such as *approaching*, *receding*, or *coalescing* from pairs of subsequent images obtained by a mobile robot's stereopair. Fernyhough et al. [11] proposed a technique for generating event models automatically based on qualitative reasoning and a statistical analysis of video input. This line of work has been further developed and has led to a range of related techniques broadly within the umbrella of the field of cognitive vision [12–14]. Dee et al. [14] proposes a method based on unsupervised clustering for building semantic scene models from video data using observed motion. Dubba et al. [12] presents a supervised learning framework to learn event models from large video datasets using inductive logic programming. Tran and Davis [15], and Morariu and Davis [16] present analogous results on the use of spatio-temporal relations within a first-order probabilistic language for the analysis of video sequences obtained in a parking lot. In a similar manner Song et al. [17] present a general framework for recognizing events in RGB-D data using probabilistic first-order logic and use it for tracking kitchen activities. Bohlken et al. [18] present work on a real-time

activity monitoring system defining activity concepts in an ontology which can be automatically transformed into a high-level scene interpretation system.

None of the works related to this paper, however, have considered a qualitative theory about space and motion as the basis to generate probabilistic interpretations of events. The present paper fulfills this gap by extending the qualitative theory proposed in [19] to account for the 3D space, while also combining it with interpretations of events from RGB-D data.

2.2 Cognitive Vision

The field of cognitive vision [20,21] has developed as an approach to enhance classical computer vision systems with cognitive abilities to obtain more robust vision systems, that are able to adapt to unforeseen changes, make sense of perceived data and show goal directed behavior. Vernon [20] 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 behavior, it should be able to adapt robustly to unforeseen changes of the visual environment, and it should be able to anticipate the occurrence of objects or events"* Vernon [20]

There are multiple approaches towards the goal of developing a cognitive vision system. A detailed research plan for the development of the field of cognitive vision systems can be found in the technical report of the ECVision (European Research Network for Cognitive Computer Vision Systems) [22]. Among others, a symbolic approach to model knowledge about spatio-temporal phenomena has gained attention [15,23–25]. Cohn et al. [26] present work towards a cognitive vision system built on qualitative spatial and temporal abstractions to ground high-level concepts in visually sensed data.

2.3 QSTR – Qualitative Spatial and Temporal Reasoning

Qualitative Spatial & Temporal Representation and Reasoning (QSTR) [27] 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. To represent the continuity of spatial change, Freksa [28] introduced the *conceptual neighborhoods*. Relations between two entities are conceptual neighbors if they can be directly transformed from one relation into the other by continuous change of the environment.

In the line of research about qualitative continuous spatial change, Galton [29–31] investigated movement on the basis of an integrated theory of space, time, objects, and position. Muller [32] defined continuous change using 4-dimensional regions in space-time. Hazarika and Cohn [33] build on this work but used an interval based approach to represent spatio-temporal primitives. In [34] Davis discusses the use of transition graphs for reasoning about continuous spatial change and applies them in physical reasoning problems.

Table 1. Spatial Relations and the Corresponding Motion Relations

Σ **Space**	
Topology	
	$discrete(p,q,t), partially\_overlapping(p,q,t),$ $proper\_part(p,q,t), proper\_part\_inverse(p,q,t), equal(p,q,t)$
Extrinsic Orientation (horizontal, vertical, and in depth)	
	$left(p,q,t), overlaps\_left(p,q,t), along\_left(p,q,t), horizontally\_equal(p,q,t),$ $overlaps\_right(p,q,t), along\_right(p,q,t), right(p,q,t)$
	$above(p,q,t), overlaps\_above(p,q,t), along\_above(p,q,t), vertically\_equal(p,q,t),$ $overlaps\_below(p,q,t), along\_below(p,q,t), below(p,q,t)$
	$closer(p,q,t), overlaps\_closer(p,q,t), along\_closer(p,q,t), distance\_equal(p,q,t),$ $overlaps\_further(p,q,t), along\_further(p,q,t), further(p,q,t)$
Σ **Motion**	
Movement	$approaching(p,q,t)$ and $receding(p,q,t)$
Size Motion	$elongating(x,p,t)$ and $shortening(x,p,t)$
Rate of Size Motion	$same\_rate(x,y,t)$, $faster(x,y,t)$,
Presence in the Scene	$appearing(p,t)$ and $disappearing(p,t)$

3 A Theory of Space, and Motion

We present a theory of space and motion to represent spatio-temporal phenomena for activity interpretation. As basic entities of the theory we consider depth profiles (see Fig. 2), which are regions of space, with a depth structure (distance from the sensor). These depth profiles are obtained by the projections of detected individuals in the scene on the image plane, where each point of the projected region has an associated depth value. Based on the depth profile we make different abstractions to encounter different aspects of space, i.e. regions, points (centroid), bounding cuboids, oriented points, lines (object axis) etc. These relations are defined in terms of the following functions on the depth profiles attributes:

depth: $depth\ profile \times time\ point \rightarrow float$, gives an depth profiles average distance from the observer at a time instant;

depth_front: $depth\ profile \times time\ point \rightarrow float$, gives an depth profiles minimal distance from the observer at a time instant;

depth_back: $depth\ profile \times time\ point \rightarrow float$, gives an depth profiles maximal distance from the observer at a time instant;

centroid: $depth\ profile \times time\ point \rightarrow (integer, integer, integer)$, gives the x,y, and z coordinates of the depth profiles centre point

size: $dimension \times depth\ profile \times time\ point \rightarrow integer$, maps a dimension, a depth profile and a time point to the depth profile's size in the given dimension;

dist: $depth\ profile \times bounding\ box \times time\ point \rightarrow float$, maps two depth profiles and a time point to the angular distance separating the depth profiles centroids in that instant.

in_sight: $depth\ profile \times time\ point \rightarrow boolean$, maps a depth profile and a time point to the presence of the depth profile. A depth profile is present at a time point, as long as there is at least one pixel associated with the depth profile.

3.1 Σ Space – Qualitative Spatial Relations

The basic part of our spatial theory consists of spatial relations on pairs of depth profiles, which includes relations on *topology* and *extrinsic orientation* in terms of left, right, above, below relations and depth relations (distance of a depth profile from the Observer).

Topological Relations. We represent the connectedness of pairs of depth profiles by the relations of the region connection calculus [35] for the 2D bounding boxes, omitting the depth. We use the RCC5 [35] subset of the region connection calculus in a ternary version, which contains the relations $discrete(p, q, t)$, $partially\_overlapping(p, q, t)$, $proper\_part(p, q, t)$, $proper\_part\_inverse(p, q, t)$, and $equal(p, q, t)$, where the third argument represents the time point when the relation holds. As the topological relations are defined on the two dimensional image plane, they do not represent the connection of two physical objects but rather the connection of the projection of two physical objects [36]. Due to this fact, the topological relations combined with the depth of the objects can be used to model that one object occludes the other.

Extrinsic Orientation. We represent the extrinsic orientation (relative position) of two depth profiles, with respect to the observer's viewpoint, making distinctions on the *3D position* and the *size* of the depth profiles. To this end, we use the bounding cuboid of the perceived depth profile determined by its *width*, *height*, and *thickness*, given as *depth_front* and *depth_back*. Given that we have 3D objects, we end up with a set of relations that resemble Allen's interval algebra [37] for each dimension, i.e. *horizontal*, *vertical*, and *depth*. However, in terms of depth perception, the interval relations that happen "instantaneously" (namely, *meets*, *starts*, and *finishes*) are irrelevant.

$$closer(p, q, t) \leftrightarrow (depth\_back(p, t) < depth\_front(q, t)); \tag{1a}$$

$$overlaps\_closer(p, q, t) \leftrightarrow (depth\_front(p, t) < depth\_front(q, t)) \wedge (depth\_front(q, t) < depth\_back(p, t)); \tag{1b}$$

$$along\_closer(p, q, t) \leftrightarrow (depth\_front(p, t) < depth\_front(q, t)) \wedge (depth\_front(q, t) < depth\_back(p, t)) \wedge (depth\_back(q, t) < depth\_back(p, t)); \tag{1c}$$

$$depth\_equal(p, q, t) \leftrightarrow (|depth\_front(p, t) - depth\_front(q, t)| < 0) \wedge (|depth\_back(p, t) - depth\_back(q, t)| < 0). \tag{1d}$$

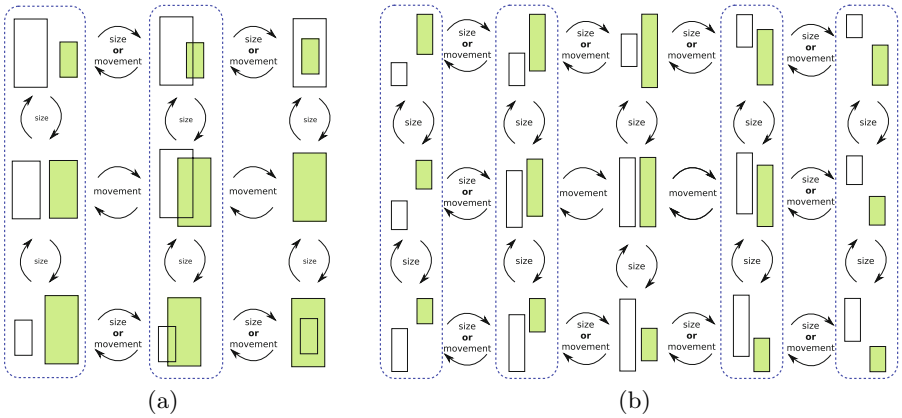


Fig. 3. Continuous Transitions between Spatial Relations on Topology and Extrinsic Orientation: topological and positional changes due to movement and transformation of the projected regions

Additionally we define the relations $further(p, q, t), overlaps\_further(p, q, t)$, and $along\_further(p, q, t)$ as inverse of the relations above. Accordingly to these relations on depth, we define the relations for the horizontal and the vertical dimension as listed in Table 1. To account for small diviations in the depth values, we apply a threshold μ represents the average error in the depth values.

3.2 Σ Motion – Qualitative Spatial Change

Spatial relations holding for perceived depth profiles change as an result of motion of the individuals in the scene (see Fig. 3). To account for this, we define motion relations by making qualitative distinctions of the changes in the depth profiles parameters, i.e. the distance between two depth profiles and its size. In each of the formulae presented below the timepoint t falls within the the open time interval (t1, t2). In this work, such time intervals are assumed to be very small; therefore, the predicates defined below are locally valid with respect to the time point t. We assume that this constraint is respected in this work but do not write it explicitly in the formulae for clarity. Further, we assume that there is a static relation between all relations to represent the case that the distance between two depth profiles stays the same, which is the case where the depth profile does not change in size or relative position.

Relative Movement. The relative movement of pairs of depth profiles is represented in terms of changes in the distance between their *centroids*. We represent these changes in terms of *approaching* and *receding* as defined below.

$$approaching(p, q, t) \leftrightarrow \exists t_1 t_2 (t_1 < t) \wedge (t < t_2) \wedge (dist(p, q, t_2) < dist(p, q, t_1)); \tag{2a}$$

$$receding(p, q, t) \leftrightarrow \exists t_1 t_2 (t_1 < t) \wedge (t < t_2) \wedge (dist(p, q, t_2) > dist(p, q, t_1)). \tag{2b}$$

Size-Motion. To represent size-motion of a single depth profile, we consider relations on changes in depth profiles *width*, *height* and *thickness* separately. Changes on more than one of these parameters at the same time instant can be represented by combinations of the relations below. In the relations below, the variable x is defined on the set of depth profiles attributes $x \in \{width, height, thickness\}$.

$$elongating(x, p, t) \leftrightarrow \exists t_1 t_2 (t_1 < t) \wedge (t < t_2) \wedge (size(x, p, t_1) < size(x, p, t_2)); \tag{3a}$$

$$shortening(x, p, t) \leftrightarrow \exists t_1 t_2 (t_1 < t) \wedge (t < t_2) \wedge (size(x, p, t_1) > size(x, p, t_2)). \tag{3b}$$

Ordering Relations on the Rate of Size-Motion. We need to define relations that state the rate of relative changes in the *width*, *height*, and *thickness* parameters of a depth profile. The relations introduced to account for these issues are defined below, where variables x and y are defined on the set of depth profile attributes ($x, y \in \{width, height, thickness\}$), and $\Delta(x)$ and $\Delta(y)$ denote the change in these parameters at the time point t which is defined on a short interval $[t_1, t_2]$ as described above.

$same\_rate(x, y, t)$ represents the case when attribute x changes "at the same rate" as y at a time point t (more formally, $\frac{\Delta(x)}{\Delta(y)} = 0$)

$faster(x, y, t)$ represents the case when attribute x changes "faster" than attribute y at a time point t (more formally, $\Delta(x) > \Delta(y)$)

Presence of depth profiles in the scene. The relations *appearing* and *disappearing* represent the events of an depth profile being present in the scene at time t that was not present in the scene at the previous time point, resp. not being present at time t but has been present at the previous time point.

$$appearing(p, t) \leftrightarrow \exists t_1 t_2 (t_1 < t) \wedge (t < t_2) \wedge \neg in\_sight(p, t_1) \wedge in\_sight(p, t_2); \tag{4a}$$

$$disappearing(p, t) \leftrightarrow \exists t_1 t_2 (t_1 < t) \wedge (t < t_2) \wedge in\_sight(p, t_1) \wedge \neg in\_sight(p, t_2). \tag{4b}$$

4 Spatial Change between Individuals in the Scene

To describe the observed scene in terms of spatio-temporal phenomena we combine the different aspects of the theory about *space* and *motion* providing a rich vocabulary about qualitative changes in the visual domain. This allows us to describe the ongoing interactions and operations between the physical entities represented by the depth profiles as well as on conceptual objects in the environment.

Individuals and objects in the scene. For the individuals and objects in the scene we assume that they have certain properties, i.e. we assume detected individuals to be rigid and non-opaque. Additionally we define abstract objects to represent the observer and the field of view of the sensing device. These objects are assumed to be non-moveable and for the field of view to have no physical object attached to it.

Visibility with Respect to the Observer. Topological relations of the depth profile's projection on the image plane, can be interpreted as visibility from the observers point of view [36] given, that the represented individuals are rigid and non-opaque. We use this fact to represent that one depth profile is occluded by another depth profile.

$$partially\_occluded(p, q, t) \leftarrow further(p, q, t) \wedge partiall\_overlapping(p, q, t). \tag{5a}$$

$$not\_occluded(p, q, t) \leftarrow discrete(p, q, t) \vee (closer(p, q, t) \wedge partially\_overlapping(p, q, t)). \tag{5b}$$

In the case of a full occlusion, the individual will not be detected any more, so this relation can only be hypothesised in the case of the disappearance of the individual.

Visibility relations changes as a result of motion, either of the individuals in the scene or of the observer. As defined in [38] the space in the environment can be divided into separate regions based on the visibility relations of an object in these regions with respect to an occluding object and the observer. Which results in the three zones, the *Light Zone(LZ)*, the *Twilight Zone(TZ)*, and the *Shadow Zone(SZ)*. To move from one zone to an other the object can only move in a certain way. E.g. to get from the right *Light Zone* to the left *Light Zone*, without passing in front of the occluding object, the object has to pass the right *Twilight Zone*, the *Shadow Zone*, and the left *Twilight Zone*.

Movement Direction with Respect to the Observer. We represent relative movement of a depth profile with respect to the observer by introducing distinct objects for the observer as well as the borders of the cameras field of view.

$$moving\_closer(p, t) \leftarrow approaching(p, observer, t); \tag{6a}$$

$$moving\_further\_away(p, t) \leftarrow receding(p, observer, t); \tag{6b}$$

$$moving\_left(p, t) \leftarrow approaching(p, left\_border, t); \tag{6c}$$

$$moving\_right(p, t) \leftarrow approaching(p, right\_border, t). \tag{6d}$$

In this way we define the relations for: (1). *moving closer*: the depth profile moves towards the observer; (2). *moving further away*: the depth profile moves away from the observer; (3). *moving left / right*: the depth profile approaches the left / right border of the field of view.

5 Human Interactions Grounded in Spatial Change

The abstractions of space and motion described in the previous section reflect changes between individuals in the real world, that are consequences of interactions conducted in the environment (or possible noise). However, in many cases it is not possible to unambiguously map from the changes in the relations to interactions of objects in the world, thus we associate the predicates on spatial change with possible hypotheses on interactions. Towards this, interactions are declaratively defined by there spatio-temporal appearance in the scene, using a

3-layered hierarchical activity model grounded in the spatial change observed in the environment. The activity model consists of the activity, interactions, and operations.

- **Activity** defined by its goal and determined by the specific interaction sequence performed towards this specific goal
- **Interaction** goal driven interactions between individuals in the scene determined by the observed spatial operations involved in the interaction
- **Spatial Operation** elemental parts of an interaction defined by spatial and temporal relations on perceived individuals in the environment

We consider consecutive frames in which the same relation holds for a pair of depth profiles or for a single depth profile as intervals of space and motion, in the sense of Allen's intervals [37]. An Interaction is then defined by spatial operations carried out by individuals involved in the interaction. Spatial operations are the basic elements of an interaction and determine, how an interaction is carried out in the environment, in terms of perceivable change. Operations are defined based on the observed intervals of space and motion using Allen's interval algebra to model temporal relations between these intervals. E.g. the interaction passing behind is declaratively defined in logic programming as depicted in Eq. 7a-d.

$$interaction(passing\_behind, P, Q, I) :- \quad (7a)$$
$$interaction(passing, P, Q, I_1), observation(partially\_occluded, P, Q, I_2),$$
$$discrete\_time(during, I_1, I_2), discrete\_time(equal, I, I_2).$$

$$interaction(passing, P, Q, I) :- \quad (7b)$$
$$operation(changing\_sides, P, Q, I_1), operation(moves, P, I_2),$$
$$discrete\_time(during, I_1, I_2), discrete\_time(equal, I, I_1).$$

$$operation(changing\_sides, P, Q, interval(T2, T3)) :- \quad (7c)$$
$$observation(horizontal(left), P, Q, interval(T1, T2)),$$
$$observation(horizontal(right), P, Q, interval(T3, T4)),$$
$$discrete\_time(meets, interval(T1, T2), interval(T3, T4)).$$

$$operation(moves, P, I) :- observation(moving(\_), P, I). \quad (7d)$$

5.1 Hypotheses on Perceived Spatial Change

Hypotheses on interactions in the real world are generated based on the perceived spatial change represented by the qualitative abstractions of *space* and *motion* and the background knowledge described in the previous section. To make hypotheses on interactions in the environment, one has to take possible noise and faulty observations into account, as well as consistency constraints between concurrent interactions.

- **Uncertainty** due to limitations in the low-level sensing, or to occlusion by other individuals in the scene. E.g. noise, missing observations, and occlusion.

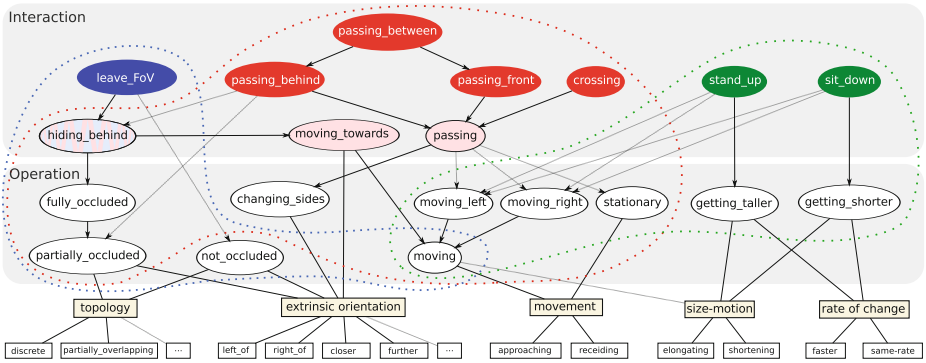


Fig. 4. Interaction taxonomy for the smart meeting domain

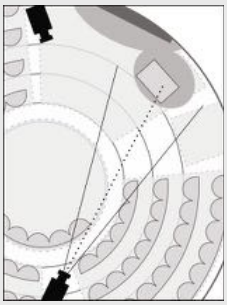
– **Consistency** in terms of concurrently performed interactions and the spatial operations contained in these interactions.

Hypotheses on interactions are arranged in a sequence, in the way, that the interactions and the corresponding spatial operations included in the interactions, best explain the observed spatial change. The probability for a certain interaction is then determined by the probability that the observation reflects the ongoing interactions in the environment, and the evidence that an observation provides for a certain interaction. For the use-case scenario presented in section 6 we use a causal network to evaluate the generated hypotheses given their grounding in observations on spatial change in the environment.

5.2 Perceptual Narratives of Human Activities

Sequences of hypothesized interactions are interpreted as perceptual narratives that describe the interactions performed in the environment with respect to the perceived spatial change. These narratives serve as a basis for reasoning in the sense of *explanation*, *prediction*, and *planning* for spatial control. As the perceptual narratives are grounded in the spatial change observed by the sensors, the narrative does not only reflect the performed interactions, but also states, how these interactions are performed in terms of the involved spatial operations.

Thus the narrative can be used to reason about the activity, the interactions within the activity, and the spatial change reflecting the interactions. And thereby help to explain incomplete or inconsistent observations, to reason about the most likely next steps towards the goal of the activity and thus predict upcoming spatial change, and to plan (spatial) control actions based on the aforementioned reasoning capabilities which is an important ability for dynamic control in smart environments.



Listing 1. Smart Meeting Cinematography
The smart meeting cinematography domain focusses on professional situations such as meetings and seminars. A basic task is to automatically produce dynamic recordings of interactive discussions, debates, presentations involving interacting people who use more than one communication modality such as hand-gestures (e.g., raising one's hand for a question, applause), voice and interruption, electronic apparatus (e.g., pressing of a button), movement (e.g., standing-up) and so forth. The scenario consists of people-tracking, gesture identification closed under a context-specific taxonomy, and also involves real-time dynamic collaborative co-ordination and self-control of pan-tilt-zoom (PTZ) cameras in a *sensing-planning-acting* loop. The long-term vision is to benchmark with respect to the capabilities of human-cinematographers, real-time video editors, surveillance personnel to record and semantically annotate individual and group activity (e.g., for summarisation, story-book format digital media and promo generation).

6 Use-Case: Smart Meeting Cinematography

We demonstrate the applicability of the theory of space and motion in the context of the meeting scenario (Listing 1). In this context, the basic interactions involved in the meeting process in Fig. 4 are considered. For the presented use-case, we assume that the camera is fixed in its position and orientation. Thus the changes observed in the relations are only due to object's motion (or noise in the sensor data).

Tracking and detection of Individuals. The particular hardware setup used in the meeting scenario consists of pan-tilt-zoom (PTZ) cameras, and depth sensors (Kinect), providing RGB-D data consisting of RGB images and corresponding depth information. Open source vision libraries, i.e. OpenCV and OpenNI are then used to detect and track individuals in the scene, which are perceived via their projection on the image plane of the sensor and their depth information. The thereby obtained depth profiles are 2.5 D regions of space, with a depth structure which gives the distance between the sensor and each pixel of the detected individuals.

Interactions in the smart meeting scenario. Interactions as performed in the meeting environment are modeled based on the spatial and temporal appearance of the interactions. For the meeting domain we take the interactions *enter_FoV*, *leave_FoV*, *passing_behind*, *passing_front*, *passing_between*, *crossing*, *stand_up*, and *sit_down* into account. Fig. 4 illustrates the taxonomy of these interactions and how they are defined based on the qualitative abstractions of space and motion. To generate the hypotheses on interactions in the environment we included a

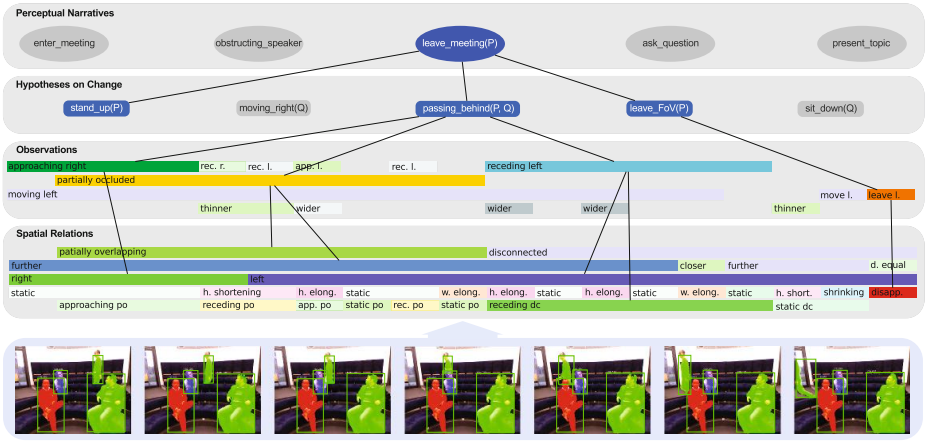


Fig. 5. Perceptual Narratives of Space, and Motion.

simple model of noise occurring in the sensing process and, of the consistency of concurrent interactions.

Resulting perceptual narrative. Using the described theory of space and motion and the interactions defined in the interaction taxonomy, combined with a simple naive bayes method for generating hypotheses, the system is able to generate the following narrative for the exemplary scene in Fig. 5.

$$
\left[\begin{array}{rl}
I_1 \equiv & interaction(stand\_up(P_4, interval(t_9, t_{13}))).\\
spatial\_operations(I_1) \equiv & getting\_taller(P_4, interval(t_9, t_{13})).\\
I_2 \equiv & interaction(passing\_behind(P_4, P_3, interval(t_{49}, t_{57}))).\\
spatial\_operations(I_2) \equiv & changing\_sides(P_4, P_3, interval(t_{52}, t_{53}))\wedge\\
& partially\_occluded(P_4, P_3, interval(t_{49}, t_{57}))\wedge\\
& moving\_left(P_4, interval(t_{45}, t_{65}))\wedge\\
& stationary(P_3, interval(t_1, t_{66})).\\
& \vdots\\
I_4 \equiv & interaction(leave\_FoV(P_4, interval(t_{66}, t_{66}))).\\
spatial\_operations(I_4) \equiv & moving\_towards(P_4, left\_border, interval(t_{65}, t_{65}))\wedge\\
& hiding\_behind(P_4, left\_border, interval(t_{66}, t_{66})).
\end{array}\right] \quad (8)
$$

Additionally to the interaction hypotheses, the narrative includes the spatial operations performed as a part of the interaction, and thereby reflect how the interactions are performed in the environment.

7 Conclusion and Outlook

Hypothesised object relations can be seen as building blocks to form complex interactions that are semantically interpreted as activities in the context of the

domain. As an example consider the sequence of observations in the meeting environment depicted in Fig. 5.

> Region P elongates vertically, region P approaches region Q from the right, region P partially overlaps with region Q while P being further away from the observer than Q, region P moves left, region P recedes from region Q at the left, region P gets disconnected from region Q, region P disappears at the left border of the field of view

These observations can be explained by the means of a percaptual narrative in terms of interactions in the real world performed in the meeting situation.

> Person P stands up, passes behind person Q while moving towards the exit and leaves the room.

Toward the generation of (declaratively grounded) perceptual narratives [6] such as the above, we developed and implemented a commonsense theory of qualitative *space* and *motion* for abstracting and reasoning about dynamic scenes. We defined combined relations capturing different spatial modalities in the context of a benchmark domain, namely the smart meeting cinematography scenario of the ROTUNDE initiative [39]. As a proof of concept, we integrated our proposed theory with a basic probabilistic reasoning method to generate hypotheses on interactions performed in the smart meeting scenario based on the combined model of *space* and *motion*. The smart meeting cinematography scenario serves as a challenging benchmark to investigate narrative based high-level cognitive interpretation of everyday interactions. Work is in progress to release certain aspects (pertaining to space, motion, real-time high-level control) emanating from the narrative model via the interface of constraint logic programming (e.g., as a Prolog based library of space–motion). Perceptual narrative based scene interpretation will be used for cognitive camera control consisting of interpreting the observations, to identify important information, and plan control actions based on the spatial requirements and constraints of scene. Work towards this end includes the integration of multiple camera viewpoints, where the system has to reason about perspective changes and visibility based on qualitative spatio-temporal abstractions.

References

1. Bhatt, M.: Reasoning about space, actions and change: a paradigm for applications of spatial reasoning. In: Qualitative Spatial Representation and Reasoning: Trends and Future Directions. IGI Global, USA (2012)
2. Cohn, A.G., Renz, J.: Qualitative spatial reasoning. In van Harmelen, F., Lifschitz, V., Porter, B., (eds.) Handbook of Knowledge Representation. Elsevier (2007)
3. Ligozat, G.: Qualitative Spatial and Temporal Reasoning. Wiley, ISTE (2013)
4. Bhatt, M., Guesgen, H., Wölfl, S., Hazarika, S.: Qualitative spatial and temporal reasoning: Emerging applications, trends, and directions. Spatial Cognition & Computation **11**, 1–14 (2011)

5. Bhatt, M., Lee, J.H., Schultz, C.: CLP(QS): a declarative spatial reasoning framework. In: Egenhofer, M., Giudice, N., Moratz, R., Worboys, M. (eds.) COSIT 2011. LNCS, vol. 6899, pp. 210–230. Springer, Heidelberg (2011)
6. Bhatt, M., Suchan, J., Schultz, C.: Cognitive interpretation of everyday activities - toward perceptual narrative based visuo-spatial scene interpretation. In: Finlayson, M., Fisseni, B., Lwe, B., Meister, J.C., (eds.) Computational Models of Narrative (CMN) (2013)
7. Lavee, G., Rivlin, E., Rudzsky, M.: Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews **39**, 489–504 (2009)
8. Gonzlez, J., Moeslund, T.B., Wang, L., (eds.) Special issue on Semantic Understanding of Human Behaviors in Image Sequences. In: Computer Vision and Image Understanding. vol. 116, pp. 305–472. Elsevier (2012)
9. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing **28**, 976–990 (2010)
10. dos Santos, M., de Brito, R.C., Park, H.H., Santos, P.: Logic-based interpretation of geometrically observable changes occurring in dynamic scenes. Applied Intelligence **31**, 161–179 (2009)
11. Fernyhough, J.H., Cohn, A.G., Hogg, D.: Constructing qualitative event models automatically from video input. Image Vision Comput. **18**, 81–103 (2000)
12. Dubba, K.S.R., Cohn, A.G., Hogg, D.C.: Event model learning from complex videos using ILP. In: ECAI, pp. 93–98 (2010)
13. Sridhar, M., Cohn, A.G., Hogg, D.C.: Unsupervised learning of event classes from video. In: AAAI (2010)
14. Dee, H.M., Cohn, A.G., Hogg, D.C.: Building semantic scene models from unconstrained video. Computer Vision and Image Understanding **116**, 446–456 (2012)
15. Tran, S.D., Davis, L.S.: Event modeling and recognition using markov logic networks. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 610–623. Springer, Heidelberg (2008)
16. Morariu, V., Davis, L.: Multi-agent event recognition in structured scenarios. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2011)
17. Song, Y.C., Kautz, H., Allen, J., Swift, M., Li, Y., Luo, J., Zhang, C.: A markov logic framework for recognizing complex events from multimodal data. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction. ICMI 2013, pp. 141–148. ACM, New York (2013)
18. Bohlken, W., Neumann, B., Hotz, L., Koopmann, P.: Ontology-based realtime activity monitoring using beam search. In: Crowley, J.L., Draper, B.A., Thonnat, M. (eds.) ICVS 2011. LNCS, vol. 6962, pp. 112–121. Springer, Heidelberg (2011)
19. Santos, P.: Reasoning about depth and motion from an observer's viewpoint. Spatial Cognition and Computation **7**, 133–178 (2007)
20. Vernon, D.: Cognitive vision: The case for embodied perception. Image Vision Comput. **26**, 127–140 (2008)
21. Vernon, D.: The Space of Cognitive Vision. In: Christensen, H.I., Nagel, H.-H. (eds.) Cognitive Vision Systems. LNCS, vol. 3948, pp. 7–24. Springer, Heidelberg (2006)
22. Auer et al.: A research roadmap of cognitive vision. Technical Report v5, ECVISION (2005). www.ecvision.org
23. Sridhar, M., Cohn, A.G., Hogg, D.C.: Learning functional object-categories from a relational spatio-temporal representation. In: ECAI, pp. 606–610 (2008)

24. Dubba, K.S.R., Cohn, A.G., Hogg, D.C.: Event model learning from complex videos using ilp. In: Proc. ECAI. Volume 215 of Frontiers in Artificial Intelligence and Applications, pp. 93–98. IOS Press (2010)
25. Gupta, A., Srinivasan, P., Shi, J., Davis, L.S.: Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In: IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2009, pp. 2012–2019 (2009)
26. Cohn, A.G., Hogg, D.C., Bennett, B., Devin, V., Galata, A., Magee, D.R., Needham, C.J., Santos, P.: Cognitive vision: integrating symbolic qualitative representations with computer vision. In: Christensen, H.I., Nagel, H.-H. (eds.) Cognitive Vision Systems. LNCS, vol. 3948, pp. 221–246. Springer, Heidelberg (2006)
27. Cohn, A., Hazarika, S.: Qualitative spatial representation and reasoning: An overview. Fundam. Inf. **46**, 1–29 (2001)
28. Freksa, C.: Conceptual neighborhood and its role in temporal and spatial reasoning. In: Singh, M., Travé-Massuyès, L. (eds.) Decision Support Systems and Qualitative Reasoning, pp. 181–187. North-Holland, Amsterdam (1991)
29. Galton, A.: Towards an integrated logic of space, time and motion. In: IJCAI, pp. 1550–1557 (1993)
30. Galton, A.: Towards a qualitative theory of movement. In: Frank, A.U., Kuhn, W. (eds.) Spatial Information Theory - A Theoretical Basis for GIS (COSIT'95), pp. 377–396. Springer, Heidelberg (1995)
31. Galton, A.: Qualitative Spatial Change. Oxford University Press (2000)
32. Muller, P.: A qualitative theory of motion based on spatio-temporal primitives. In: Cohn, A.G., Schubert, L.K., Shapiro, S.C., (eds.) Proceedings of the Sixth International Conference on Principles of Knowledge Representation and Reasoning (KR'98), pp. 131–143. Morgan Kaufmann, Trento, June 2–5, 1998
33. Hazarika, S.M., Cohn, A.G.: Abducing qualitative spatio-temporal histories from partial observations. In: KR, pp. 14–25 (2002)
34. Davis, E.: Qualitative reasoning and spatio-temporal continuity. In: Hazarika, S.M. (ed.) Qualitative Spatio-Temporal Representation and Reasoning: Trends and Future Directions, pp. 97–146. IGI Global, Hershey (2012)
35. Cohn, A.G., Bennett, B., Gooday, J., Gotts, N.M.: Qualitative spatial representation and reasoning with the region connection calculus. Geoinformatica **1**, 275–316 (1997)
36. Randell, D., Witkowski, M., Shanahan, M.: From images to bodies: Modeling and exploiting spatial occlusion and motion parallax. In: Proc. of IJCAI, Seattle, U.S., pp. 57–63 (2001)
37. Allen, J.F.: Maintaining knowledge about temporal intervals. Commun. ACM **26**, 832–843 (1983)
38. Tassoni, S., Fogliaroni, P., Bhatt, M., Felice, G.D.: Toward a Qualitative 3D Visibility Model. In: 25th International Workshop on Qualitative Reasoning, co-located with the IJCAI-11 Conference, Barcelona, Spain (2011)
39. Bhatt, M., Suchan, J., Freksa, C.: ROTUNDE - A Smart Meeting Cinematography Initiative. In: Bhatt, M., Guesgen, H., Cook, D. (eds.) Proceedings of the AAAI-2013 Workshop on Space, Time, and Ambient Intelligence (STAMI). AAAI Press, Washington (2013)