Commonsense Scene Semantics for Cognitive Robotics

Towards Grounding Embodied Visuo-Locomotive Interactions

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

[1] University of Bremen, Germany. `www.spatial-reasoning.com`
EASE CRC. `http://ease-crc.org/`

[2] Örebro University, Sweden.
MPI Lab. `https://mpi.aass.oru.se`

Abstract

We present a commonsense, qualitative model for the semantic grounding of embodied visuo-spatial and locomotive interactions. The key contribution is an integrative methodology combining low-level visual processing with high-level, human-centred representations of space and motion rooted in artificial intelligence. We demonstrate practical applicability with examples involving object interactions, and indoor movement.

1. Introduction

Practical robotic technologies and autonomous systems in real-world settings are confronted with a range of situational and context-dependent challenges from the viewpoints of perception & sensemaking, high-level planning & decision-making, human-machine interaction etc. Very many research communities and sub-fields thereof addressing the range of challenges from different perpecsectives have flourished in the recent years: computer vision, artificial intelligence, cognitive systems, human-machine interaction, cognitive science, multi-agent systems, control & systems engineering to name a few. Driven by the need to achieve contextualised practical deployability in real-world non-mundane everyday situations involving living beings, there is now a clearly recognised need for integrative research that combines state of the art methods from these respective research areas. Towards this, the research presented in this paper addresses commonsense visuo-spatial scene interpretation in indoor robotics settings at the interface of vision, AI, and spatial cognition. The focus of this research is on activities of everyday living involving people, robots, movement, and human-machine interaction.

Interpreting Embodied Interaction:
On Grounded Visuo-Locomotive Perception

Visuo-locomotive perception denotes the capability to develop a conceptual mental model (e.g., consisting of abstract, commonsense representations) emanating from multi-sensory perceptions during embodied interactions and movement in a real world populated by static and dynamic entities and artefacts (e.g., moving objects, furniture). Visuo-locomotive perception in the context of cognitive robotics technologies and machine perception & interaction systems involves a complex interplay of high-level cognitive processes. These could, for instance, encompass capabilities such as explainable reasoning, learning, concept formation, sensory-motor control; from a technical standpoint of AI technologies, this requires the mediation of commonsense formalisms for reasoning about *space, events, actions, change*, and *interaction* [5].

▶ With visuo-locomotive cognition as the context, consider the task of semantic interpretation of multi-modal perceptual data (e.g., about human behaviour, the environment and its affordances), with objectives ranging from knowledge acquisition and data analyses to hypothesis formation, structured relational learning, learning by demonstration etc. Our research focusses on the processing and semantic interpretation of dynamic visuo-spatial imagery with a particular emphasis on the ability to abstract, reason, and learn commonsense knowledge that is semantically founded in qualitative spatial, temporal, and spatio-temporal relations and motion patterns. We propose that an ontological characterisation of human-activities — e.g., encompassing (embodied) spatio-temporal relations— serves as a bridge between high-level conceptual categories (e.g., pertaining to human-object interactions) on the one-hand, and low-level / quantitative sensory-motor data on the other.
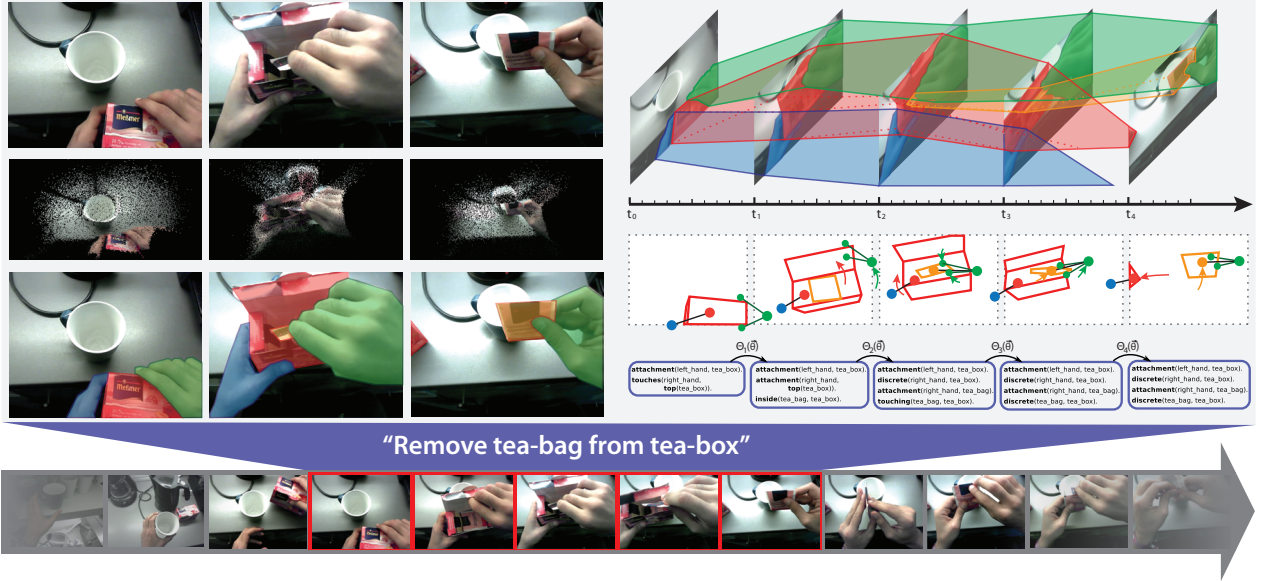
Figure 1: A Sample Activity – "Making a cup of tea" (egocentric view from a head-mounted RGB-D capture device)

Commonsense Scene Semantics:

Integrated Vision and Knowledge Representation

The starting point of the work presented in this paper is in formal commonsense representation and reasoning techniques developed in the field of Artificial Intelligence. Here, we address the key question:

How can everyday embodied activities (involving interaction and movement) be formally represented in terms of spatio-temporal relations and movement patterns (augmented by context-dependent knowledge about objects and environments) such that the representation enables robotic agents to execute everyday interaction tasks (involving manipulation and movement) appropriately?

We particularly focus on an ontological and formal characterisation of space and motion from a human-centered, commonsense formal modeling and computational reasoning viewpoint, i.e., *space-time*, as it is interpreted within the AI subdisciplines of knowledge representation and reasoning, and commonsense reasoning, and within spatial cognition & computation, and more broadly, within spatial information theory [1, 5, 6, 8, 9, 24].

▶ We build on state of the art methods for visual processing of RGB-D and point-cloud data for sensing the environment and the people within. In focus are 3D-SLAM data for extracting floor-plan structure based on plane detection in point-clouds, and people detection and skeleton tracking using Microsoft Kinect v2. Furthermore, we combine robot self-localisation and people tracking to localise observed people interactions in the global space of the environmental map.

We emphasise that the ontological and representational aspects of our research are strongly driven by computational considerations focussing on: **(a)**. developing general methods and tools for commonsense reasoning about space and motion categorically from the viewpoint of commonsense cognitive robotics in general, but human-object interactions occurring in the context of everyday activities in particular; **(b)**. founded on the established ontological model, developing models, algorithms and tools for reasoning about space and motion, and making them available as part of cognitive robotics platforms and architectures such as ROS. The running examples presented in the paper highlight the semantic question-answering capabilities that are directly possible based on our commonsense model directly in the context of *constraint logic programming*.

2. Commonsense, Space, Motion

Commonsense spatio-temporal relations and patterns (e.g. left-of, touching, part-of, during, approaching, collision) offer a human-centered and cognitively adequate formalism for logic-based automated reasoning about embodied spatio-temporal interactions involved in everyday activities such as *flipping a pancake*, *grasping a cup*, or *opening a tea box* [8, 26, 27, 30]. Consider Fig. 1, consisting of a sample human activity —"*making a cup of tea*"— as captured from an egocentric viewpoint with a head-mounted RGB-D capture device. From a commonsense viewpoint, the sequence of high-level steps typically involved in this activity, e.g., opening a tea-box, removing a tea-bag from the box and putting the tea-bag inside a tea-cup filled with water while holding the tea-cup, each qualitatively correspond to high-

Spatial Domain ($\mathcal{QS}$)	Formalisms	Spatial Relations ($\mathcal{R}$)	Entities ($\mathcal{E}$)
Mereotopology	RCC-5, RCC-8 [23]	disconnected (dc), external contact (ec), partial overlap (po), tangential proper part (tpp), non-tangential proper part (ntpp), proper part (pp), part of (p), discrete (dr), overlap (o), contact (c)	arbitrary rectangles, circles, polygons, cuboids, spheres
	Rectangle & Block algebra [16]	proceeds, meets, overlaps, starts, during, finishes, equals	axis-aligned rectangles and cuboids
Orientation	LR [25]	left, right, collinear, front, back, on	2D point, circle, polygon with 2D line
	OPRA [21]	facing towards, facing away, same direction, opposite direction	oriented points, 2D/3D vectors
Distance, Size	QDC [19]	adjacent, near, far, smaller, equi-sized, larger	rectangles, circles, polygons, cuboids, spheres
Dynamics, Motion	Space-Time Histories [17, 18]	moving: towards, away, parallel; growing / shrinking: vertically, horizontally; passing: in front, behind; splitting / merging; rotation: left, right, up, down, clockwise, counter-clockwise	rectangles, circles, polygons, cuboids, spheres

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

level spatial and temporal relationships between the agent and other involved objects. For instance, one may most easily identify relationships of contact and containment that hold across specific time-intervals. Here, parametrised manipulation or control actions ($\Theta_1(\boldsymbol{\theta})$, ...$\Theta_n(\boldsymbol{\theta})$) effectuate state transitions, which may be qualitatively modelled as changes in topological relationships amongst involved domain entities.

Embodied interactions, such as those involved in Fig. 1, may be grounded using a holistic model for the commonsense , qualitative representation of space, time, and motion (Table 1). In general, qualitative, multi-modal, multi-domain[1] representations of spatial, temporal, and spatio-temporal relations and patterns, and their mutual transitions can provide a mapping and mediating level between human-understandable natural language instructions and formal narrative semantics on the one hand [8, 13], and symbol grounding, quantitative trajectories, and low-level primitives for robot motion control on the other. By spatio-linguistically grounding complex sensory-motor trajectory data (e.g., from human-behaviour studies) to a formal framework of space and motion, generalized (activity-based) qualitative reasoning about dynamic scenes, spatial relations, and motion trajectories denoting single and multi-object path & motion predicates can be supported [14]. For instance, such predicates can be abstracted within a region based 4D space-time framework [3, 4, 18], object interactions [10, 11], and spatio-temporal narrative knowledge [12, 13, 29]. An adequate qualitative spatio-temporal representation can therefore connect with low-level constraint-based movement control systems of robots [2], and also help grounding symbolic descriptions of actions and objects to be manipulated (e.g., natural language instructions such as cooking recipes [28]) in the robots perception.

---

[1]Multi-modal in this context refers to more than one aspect of space, e.g., topology, orientation, direction, distance, shape. Multi-domain denotes a mixed domain ontology involving points, line-segments, polygons, and regions of space, time, and space-time [18]. Refer Table 1.

3. Visuo-Locomotive Interactions: A Commonsense Characterisation

3.1. Objects and Interactions in Space-Time

Activities and interactions are described based on visuo-spatial domain-objects $\mathcal{O} = \{o_1, o_2, ..., o_i\}$ representing the visual elements in the scene, e.g., *people* and objects. The **Qualitative Spatio-Temporal Ontology** ($\mathcal{QS}$) is characterised by the basic spatial and temporal entities ($\mathcal{E}$) that can be used as abstract representations of domain-objects and the relational spatio-temporal structure ($\mathcal{R}$) that characterises the qualitative spatio-temporal relationships amongst the entities in ($\mathcal{E}$). Towards this, domain-objects ($\mathcal{O}$) are represented by their spatial and temporal properties, and abstracted using the following basic *spatial entities*:

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

and the temporal entities:

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

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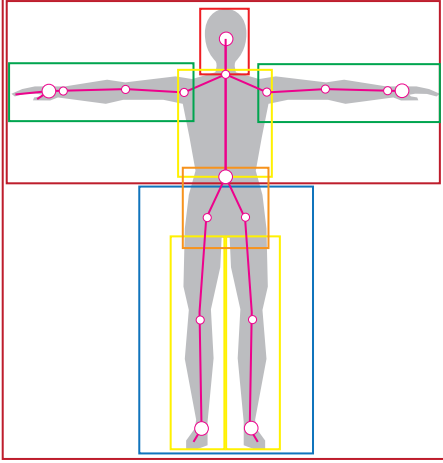
```
person(full_body,
       [upper_body, lower_body]).

person(upper_body,
       [head, left_arm, right_arm, ...]).
...

body_part(left_upper_arm,
          joint(shoulder_left),
          joint(elbow_left)).

body_part(left_forearm,
          joint(elbow_left),
          joint(elbow_left)).
...

joint(spine_base, joint(id(0)).
joint(spine_mid, joint(id(1))).
joint(neck, id(2)).
joint(head, id(3)).
...
joint(thumb_right, id(24)).
```

Figure 2: Declarative Model of Human-Body Posture

Declarative Model of Visuo-Locomotive Interactions
Based on the qualitative spatio-temporal ontology ($\mathcal{QS}$), human interactions in the environment are represented using a declarative model of visuo-locomotive interactions, encompassing dynamics of *humans*, *objects*, and the *environmental characteristics*.

- *Human Body Pose* The human body is represented using a declarative model of the body-structure (see Fig. 2), within this model we ground the human body-pose in 3D-data of skeleton joints and body-parts obtained from RGB-D sensing.

- *Semantics of the Environment* The semantic structure of the environment is represented using a topological map corresponding to the floor-plan of the environment extracted from 3D point-clouds obtained from 3D-SLAM data.

Using these models, visuo-locomotive interactions, involving humans, robots, and objects can be declaratively ab-

stracted by the spatio-temporal characteristics of the involved domain-objects, and may be used for high-level interpretation and reasoning about scene dynamics.

3.2. Spatio-Temporal Characteristics of Human Activities

The *space-time histories* (sth) used to abstract the dynamics of human activities are based on basic spatio-temporal entities obtained from the sensed data, corresponding to the declarative model of visuo-locomotive interactions. To extract these entities, we define functions for the specific spatio-temporal properties of domain-objects. I.e., the following functions are used for static spatial properties.

- position: $\mathcal{O} \times \mathcal{T} \to \mathbb{R} \times \mathbb{R} \times \mathbb{R}$, gives the 3D position (x,y,z) of an object o at a time-point t;

- *size*: $\mathcal{O} \times \mathcal{T} \to \mathbb{R}$, gives the size of an object o at a time-point t;

- *distance*: $\mathcal{O} \times \mathcal{O} \times \mathcal{T} \to \mathbb{R}$, gives the distance between two objects o_1 and o_2 at a time-point t;

- *angle*: $\mathcal{O} \times \mathcal{O} \times \mathcal{T} \to \mathbb{R}$, gives the angle between two objects o_1 and o_2 at a time-point t;

To account for changes in the spatial properties of domain-objects we use the following functions for dynamic spatio-temporal properties.

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

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

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

These functions are used to obtain basic spatial entities, e.g. *points, lines, regions*, from the sensor data. Spatio-temporal relationships ($\mathcal{R}$) between the basic entities in $\mathcal{E}$ may be characterised with respect to arbitrary spatial and spatio-temporal domains such as *mereotopology, orientation, distance, size, motion, rotation* (see Table 1 for a list of considered spatio-temporal abstractions). E.g, let $D_1, \ldots, D_n$ be spatial domains (e.g. the domain of axis-aligned rectangles). A spatial relation r of arity n ($0 < n$) is defined as:

$$r \subseteq D_1 \times \cdots \times D_n.$$

The spatio-temporal dynamics of the scene can then be represented based on the relations holding between the objects in the scene, and the changes with in them.
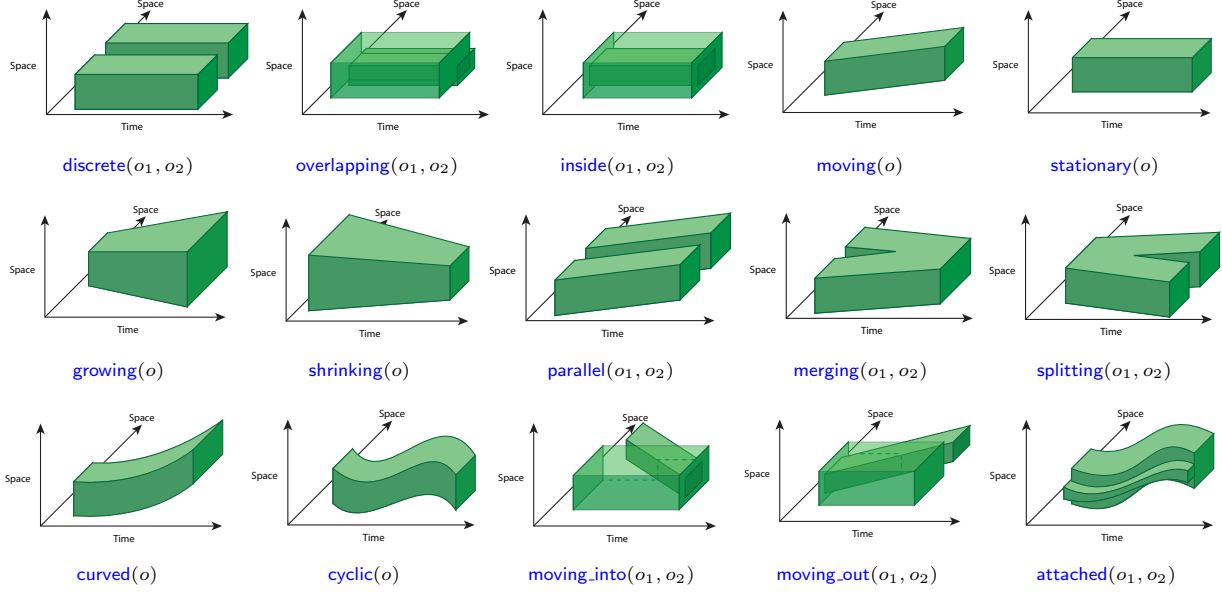
discrete(o_1, o_2)   overlapping(o_1, o_2)   inside(o_1, o_2)   moving(o)   stationary(o)

growing(o)   shrinking(o)   parallel(o_1, o_2)   merging(o_1, o_2)   splitting(o_1, o_2)

curved(o)   cyclic(o)   moving_into(o_1, o_2)   moving_out(o_1, o_2)   attached(o_1, o_2)

Figure 3: Commonsense Spatial Reasoning with Space-Time Histories Representing Dynamics in Everyday Human Activities

Interaction (Θ)	Description
$pick\_up(P, O)$	a person P picks up an object O.
$put\_down(P, O)$	a person P puts down an object O.
$reach\_for(P, O)$	a person P is reaches for an object O.
$pass\_over(P_1, P_2, O)$	a person P_1 passes an object O to another person P_2.

Interaction (Θ)	Description
$moves\_into(P, FS)$	a person P enters a floor-plan structure FS.
$passes(P, FS)$	a person P passes through a floor-plan structure FS.

Table 2: Sample Interactions Involved in Everyday Human Activities: Human Object Interactions and Peoples Locomotive Behaviour

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

$$\begin{aligned} \text{holds-in}(\text{approaching}(o_i, o_j), \delta) \supset{} & \text{during}(t_i, \delta) \wedge \text{during}(t_j, \delta) \wedge \\ & \text{before}(t_i, t_j) \wedge (distance(o_i, o_j, t_i) > distance(o_i, o_j, t_j)). \end{aligned} \tag{1}$$

Interactions. Interactions $\Theta = \{\theta_1, \theta_2, ..., \theta_i\}$ describe processes that change the spatio-temporal configuration of objects in the scene, at a specific time; these are defined by the involved spatio-temporal dynamics in terms of changes in the status of space-time histories caused by the interaction, i.e. the description consists of (dynamic) spatio-temporal relations of the involved space-time histories, before, during and after the interaction (See Table 2 for exemplary interactions). We use occurs-at(θ, t), and occurs-in(θ, δ) to denote that an interaction θ occurred at a *time point* t or in an *interval* δ, e.g., a person reaching for an object can be defined as follows.

$$\begin{aligned} \text{occurs-in}(\text{reach\_for}(o_i, o_j), \delta) \supset{} & \text{person}(o_i) \wedge \\ & \text{holds-in}(approaching(body\_part(hand, o_i), o_j), \delta_i) \wedge \\ & \text{holds-in}(touches(body\_part(hand, o_i), o_j), \delta_j) \wedge \\ & \text{meets}(\delta_i, \delta_j) \wedge \text{starts}(\delta_i, \delta) \wedge \text{ends}(\delta_j, \delta). \end{aligned} \tag{2}$$

These definitions can be used to represent and reason about people interactions involving people movement in the environment, as well as fine-grained activities based on body pose data.

4. Application: Grounding Visuo-Locomotive Interactions

We demonstrate the above model for grounding everyday activities in perceptual data obtained from RGB-D sensing. The model has been implemented within (Prolog based) constraint logic programming based on formalisations of qualitative space in CLP(QS) [7].
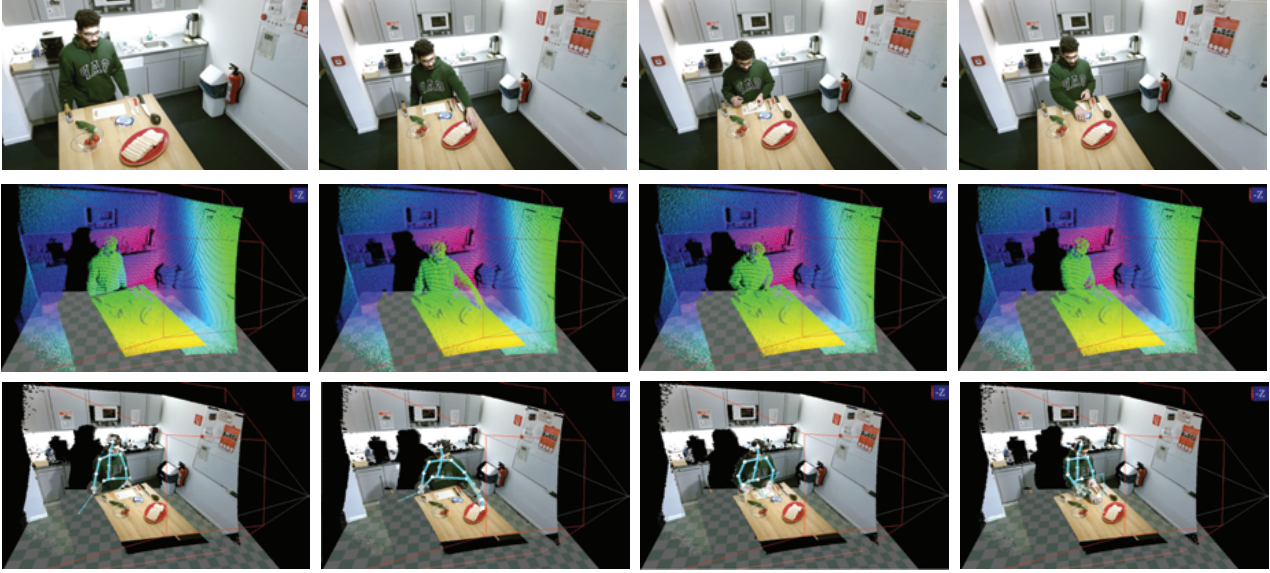
Figure 4: RGB-D data of Human Activities with Corresponding Skeleton Data

Human Activity Data

RGB-D Data (video, depth, body skeleton): We collect data using Microsoft Kinect v2 which provides RGB and depth data. The RGB stream has a resolution of 1920x1080 pixel at 30 Hz and the depth sensor has a resolution of 512x424 pixels at 30 Hz. Skeleton tracking can track up to 6 persons with 25 joints for each person. Further we use the point-cloud data to detect objects on the table using table-top object segmentation, based on plane detection to detect the tabletop and clustering of points above the table. For the purpose of this paper simple colour measures are used to distinguish the objects in the scene.

3D SLAM Data (3d maps, point-clouds, floor-plan structure): We collect 3D-SLAM data using Real-Time Appearance-Based Mapping (RTAB-Map) [20], which directly integrates with the Robot Operating System (ROS) [22] for self localisation and mapping under real-time constraints. In particular, for semantic grounding presented in this paper, we use the point-cloud data of the 3D maps obtained from RTAB-Map to extract floor-plan structures by, 1) detection of vertical planes as candidate wall-segments, 2) pre-processing of the wall-segments using clustering and line-fitting, and 3) extraction of room structures based on extracted wall-segments and lines.

Plane Detection. Planes in the point-cloud data are detected using a region-growing approach based on the normals of points. To extract candidate wall-segments, we select planes, that are likely to be part of a wall, i.e., horizontal, and sufficiently high or connected to the ceiling. These planes are then abstracted as geometrical entities, specified by their *position*, *size*, and *orientation* (given by the normal), which are used for further analysis.

Clustering and Line-Fitting. The detected wall-segments are grouped in a two-stage clustering process using density-based clustering (DBSCAN) [15], in the first step we cluster wall-segments based on their 2D orientation, in the second step, we align all wall-segments based on the average orientation of the cluster they are in and cluster the wall-segments based on the distance between the parallel lines determined by the wall-segments. We use least square linear regression to fit lines to the resulting wall clusters, which provide the structure of the environment.

Extracting Room Structures. Candidate structures for rectangular rooms and corridors are determined by the intersection points of the lines fitted to the wall clusters by considering each intersection point as a possible corner of a room or a corridor. The actual rooms and corridors are then selected based on the corresponding wall segments, projected to the lines.

Ex 1. Human-Object Interactions

Sample Activity: "Making a Sandwich". The activity of making a sandwich is characterised with respect to the interactions between a human and its environment, i.e. objects the human uses in the process of preparing the sandwich. Each of these interactions is defined by its spatio-temporal characteristics, in terms of changes in the spatial arrangement in the scene (as described in Sec. 3). As an result we obtain a sequence of interactions performed within the track of the particular instance of the activity, grounded in
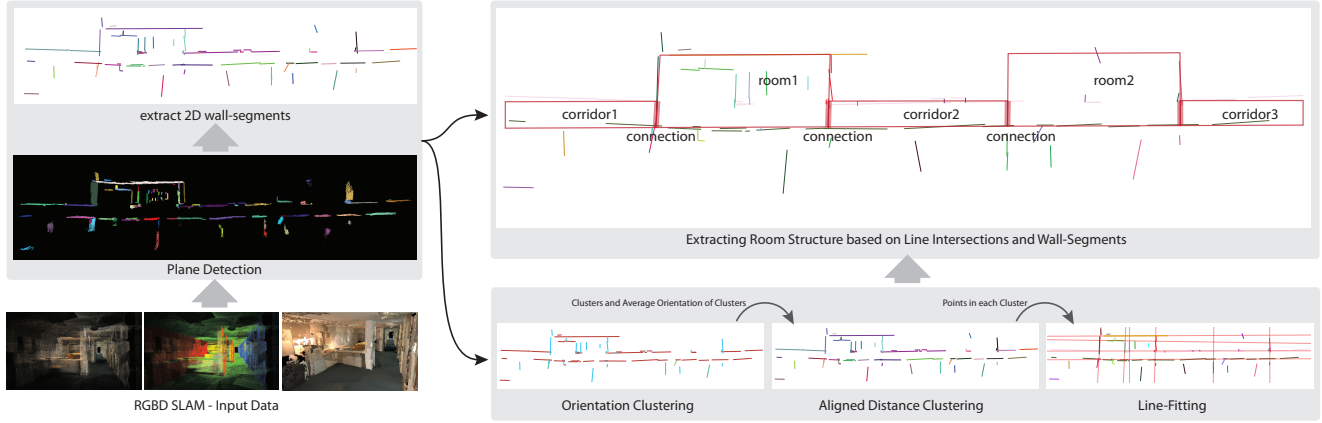
Figure 5: Extracting Floor-Plan Semantics from 3D SLAM Data

the spatio-temporal dynamics of the scenario. As an example consider the sequence depicted in Fig. 4, the interactions in this sequence can be described as follows:

"Person1 *reaches* for the bread, *picks up* a slice of bread, and *moves* the hand together with the bread *back* to the chopping board."

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

```
at(joint(id(0), person(id(1))),
    tracking_status(2),
    pos_3d(point(0.230083,-0.0138919,2.05633)),
    time_point(2168577589)).

at(joint(id(1), person(id(1))),
    tracking_status(2),
    pos_3d(point(0.228348,0.275798,1.98048)),
    time_point(2168577589)).
...
at(object(id(0)), type(bread),
    pos_3d(point(0.223528,0.500194,1.92038)),
    time_point(2168577589)).
...
```

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

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

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

```
Interaction = reach_for(person(id(1)), object(bread)),
Interval = interval(t1, t3),
Grounding =
  [holds_in(
   approaching(
   body_part(right_hand, person(id(1))), object(bread)),
   interval(t1,t2)),
   holds_in(
   touching(
   body_part(right_hand, person(id(1))), object(bread)),
   interval(t2,t3)];
```

```
Interaction = pick_up(person(id(1)), object(bread)),
Interval = interval(t4, t6),
Grounding =
  [occurs_at(
   grasp(
   body_part(right_hand, person(id(1))), object(bread)),
   timepoint(t4),
   holds_in(
   attached(
   body_part(right_hand, person(id(1))), object(bread)),
   interval(t5,t8)),
   holds_in(
   move_up(body_part(right_hand, person(id(1)))),
   interval(t5,t6))];
...
```

In particular, the interaction $reach\_for(person(id(1)),$ $object(bread))$ occurring between time-point $t1$ and $t3$ is composed of the spatio-temporal pattern of $approaching$, stating that the right hand of person 1 is approaching the bread during time-interval $t1$ to $t2$, and the pattern $touching$, stating that the right hand of person 1 is touching the bread during time-interval $t2$ to $t3$. Similarly the interaction $pick\_up(person(id(1)), object(bread))$ is composed of $grasping$, $attachment$ and $upwards\ movement$, with the difference, that grasping itself is an interaction, that can be further grounded in movement dynamics. This kind of declarative grounding can be used, e.g., for relational learning by demonstration, etc.

Ex 2. Visuo-Locomotive Interactions

Sample Activity: "Indoor Robot Navigation". Robots having to behave and navigate in environments populated by humans have to understand human activities and interactions and have to behave accordingly. In this context, high-level abstractions of human everyday activities and the semantics of the environment can be used to guide robot decision-making to account for humans moving in the environment.

As an example consider a robot that has to move from $room1$ to $room2$, during this movement the robot has to pass through the corridor $corridor2$. The structure of the environment is represented as follows:

```
floorplan_structure(id(room1), type(room)).
geom(id(room1), polygon([
    point(33.04074928940648, 47.78135448806469),
    point(41.95226523762394, 53.36407052934558),
    point(44.20648040233633, 49.48777998147939),
    point(35.17204967399538, 43.83961776145081)
])).

floorplan_structure(id(corr1), type(corridor)).

geom(id(corr1), polygon([
    point(34.17204967399538, 42.83961776145081),
 ...
])).

floorplan_structure(id(room2), type(room)).
geom(id(room2), polygon([ ... ])).

...
```

Based on the extracted semantic floor-plan structure and the detected people in the environment, the robot can make decisions using high-level navigation rules, e.g. defining how to behave in narrow passages, such as the corridor when there are people in the corridor. E.g. we can define a rule that a robot can only enter a corridor (the robot can execute the control action $enter(Floorplan\_Structure, T)$), if there is no person in the corridor, or the person is passing the corridor in the same direction as the robot. For this example we use a simple action language for planing the actions of the robot, in this context the predicate $poss\_at(\theta, t)$ defines the spatio-temporal configuration, in which the control action θ can be executed.

```
poss_at(
  enter(fp_struct(id(FS_ID), type(corridor))),
  timepoint(T)) :-

    not(holds_at(
      inside(person(P_ID), fp_struct(FS_ID)),
      timepoint(T))).

poss_at(
  enter(fp_struct(id(FS_ID), type(corridor))),
  timepoint(T)) :-

    holds_at(
      inside(person(P_ID), fp_struct(id(FS_ID), _)),
      timepoint(T)),

    occurs_in(
      passing(person(P_ID), fp_struct(id(FS_ID), _), Dir1),
      interval(I)),

    time(during(T, I)),

    trajectory(person(P_ID), P_Path, interval(I)),
    planned_trajectory(R_Path),

    movement_direction(P_Path, Dir1),
    movement_direction(R_Path, Dir2),
    same_direction(Dir1, Dir2).
```

The above rules state that the robot can only execute the control action $enter(fp\_struct(id(FS\_ID), type(corridor)))$ if one of the two rules is true. The first rule simply states that the robot can enter the corridor if there is no person in the corridor. The second rule states, that if there is a person inside the corridor, and the person is passing through the corridor, the robot can enter the corridor, if the trajectory of the person passing through the corridor and the planed path of the robot are passing the corridor in the same direction.

In this way, high-level semantic behaviour descriptions can be used for guiding low-level robot controls, such as path planning, etc.

5. Summary and Outlook

Visuo-locomotive sensemaking for practical cognitive robotics in contextualised settings with humans, mobility, and human-machine interaction is a complex endeavour requiring integrative methodologies combining the state of the art from several research areas such as vision, AI, HCI, and engineering. Our research emphasis the particular merit of combining visual processing with semantic knowledge representation and reasoning techniques rooted in artificial intelligence, particularly commonsense reasoning. We have presented a declarative commonsense model for grounding embodied visuo-locomotive interactions; the proposed model —consisting of a formal characterisation of space, time, space-time, and motion patterns— is geared towards knowledge representation and reasoning capabilities such as commonsense abstraction, learning, reasoning, embodied simulation. With a particular focus on the representation of space-time histories and motion patterns, the model is illustrated with select RGB-D datasets corresponding to representative activities from a larger dataset of everyday activities. Immediate next steps involve integration with state of the art robot control platforms such as ROS; this will be accomplished via integration into the ExpCog commonsense cognition robotics platform for experimental / simulation purposes, and within openEASE as a state of the art cognition-enabled control of robotic control platform for real robots.[2]

Acknowledgements

---

[2]ExpCog – http://www.commonsenserobotics.org
openEASE – http://ease.informatik.uni-bremen.de/openease/

References

[1] M. Aiello, I. E. Pratt-Hartmann, and J. F. v. Benthem. *Handbook of Spatial Logics*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007. 2

[2] G. Bartels, I. Kresse, and M. Beetz. Constraint-based movement representation grounded in geometric features. In *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*, Atlanta, Georgia, USA, October 15–17 2013. 3

[3] B. Bennett, A. G. Cohn, P. Torrini, and S. M. Hazarika. Describing rigid body motions in a qualitative theory of spatial regions. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on on Innovative Applications of Artificial Intelligence*, pages 503–509, 2000. 3

[4] B. Bennett, A. G. Cohn, P. Torrini, and S. M. Hazarika. A foundation for region-based qualitative geometry. In *Proceedings of the 14th European Conference on Artificial Intelligence*, pages 204–208, 2000. 3

[5] 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. 1, 2

[6] M. Bhatt, H. Guesgen, S. Wölfl, and S. Hazarika. Qualitative spatial and temporal reasoning: Emerging applications, trends, and directions. *Spatial Cognition & Computation*, 11(1):1–14, 2011. 2

[7] 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*, pages 210–230, 2011. 5

[8] 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*. 978-0-19-967991-1, Oxford University Press, 2013. 2, 3

[9] A. G. Cohn and J. Renz. Qualitative spatial reasoning. In F. van Harmelen, V. Lifschitz, and B. Porter, editors, *Handbook of Knowledge Representation*. Elsevier, 2007. 2

[10] E. Davis. Pouring liquids: A study in commonsense physical reasoning. *Artif. Intell.*, 172(12-13):1540–1578, 2008. 3

[11] E. Davis. How does a box work? a study in the qualitative dynamics of solid objects. *Artif. Intell.*, 175(1):299–345, 2011. 3

[12] E. Davis. Qualitative spatial reasoning in interpreting text and narrative. *Spatial Cognition & Computation*, 13(4):264–294, 2013. 3

[13] M. Eppe and M. Bhatt. Narrative based postdictive reasoning for cognitive robotics. In *COMMONSENSE 2013: 11th International Symposium on Logical Formalizations of Commonsense Reasoning*, 2013. 3

[14] C. Eschenbach and K. Schill. Studying spatial cognition - a report on the dfg workshop on "the representation of motion". *KI*, 13(3):57–58, 1999. 3

[15] M. Ester, H. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In E. Simoudis, J. Han, and U. M. Fayyad, editors, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA*, pages 226–231. AAAI Press, 1996. 6

[16] H. W. Guesgen. *Spatial reasoning based on Allen's temporal logic*. Technical Report TR-89-049. International Computer Science Institute Berkeley, 1989. 3

[17] P. J. Hayes. Naive physics I: ontology for liquids. In J. R. Hubbs and R. C. Moore, editors, *Formal Theories of the Commonsense World*. Ablex Publishing Corporation, Norwood, NJ, 1985. 3

[18] S. M. Hazarika. *Qualitative Spatial Change : Space-Time Histories and Continuity*. PhD thesis, The University of Leeds, School of Computing, 2005. Supervisor - Anthony Cohn. 3

[19] D. Hernández, E. Clementini, and P. Di Felice. *Qualitative distances*. Springer, 1995. 3

[20] M. Labbe and F. Michaud. Online Global Loop Closure Detection for Large-Scale Multi-Session Graph-Based SLAM. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2661–2666, Sept 2014. 6

[21] R. Moratz. Representing relative direction as a binary relation of oriented points. In *ECAI*, pages 407–411, 2006. 3

[22] M. Quigley, K. Conley, B. P. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng. Ros: an open-source robot operating system. In *ICRA Workshop on Open Source Software*, 2009. 6

[23] D. A. Randell, Z. Cui, and A. G. Cohn. A spatial logic based on regions and connection. *KR*, 92:165–176, 1992. 3

[24] J. Renz and B. Nebel. Qualitative spatial reasoning using constraint calculi. In *Handbook of Spatial Logics*, pages 161–215. 2007. 2

[25] A. Scivos and B. Nebel. The Finest of its Class: The Natural, Point-Based Ternary Calculus LR for Qualitative Spatial Reasoning. In *C. Freksa et al. (2005), Spatial Cognition IV. Reasoning, Action, Interaction: International Conference Spatial Cognition. Lecture Notes in Computer Science Vol. 3343, Springer, Berlin Heidelberg*, volume 3343, pages 283–303, 2004. 3

[26] M. Spranger, J. Suchan, and M. Bhatt. Robust Natural Language Processing - Combining Reasoning, Cognitive Semantics and Construction Grammar for Spatial Language. In *IJCAI 2016: 25th International Joint Conference on Artificial Intelligence*. AAAI Press, 2016. 2

[27] M. Spranger, J. Suchan, M. Bhatt, and M. Eppe. Grounding dynamic spatial relations for embodied (robot) interaction. In *PRICAI 2014: Trends in Artificial Intelligence - 13th Pacific Rim International Conference on Artificial Intelligence, Gold Coast, QLD, Australia, December 1-5, 2014. Proceedings*, pages 958–971, 2014. 2

[28] S. Tellex. *Natural language and spatial reasoning*. PhD thesis, Massachusetts Institute of Technology, 2010. 3

[29] A. Tyler and V. Evans. *The semantics of English prepositions: spatial scenes, embodied meaning and cognition*. Cambridge University Press, Cambridge, 2003. 3

[30] F. Worgotter, E. E. Aksoy, N. Kruger, J. Piater, A. Ude, and M. Tamosiunaite. A simple ontology of manipulation actions based on hand-object relations. 2012. 2