A Spatial Reasoning Framework for Commonsense Reasoning in Visually Intelligent Agents

Agnese Chiatti[1], Gianluca Bardaro[1], Enrico Motta[1] and Enrico Daga[1]

[1]*Knowledge Media Institute, The Open University, United Kingdom*

Abstract

Service robots are expected to reliably make sense of complex, fast-changing environments. From a cognitive standpoint, they need the appropriate reasoning capabilities and background knowledge required to exhibit human-like Visual Intelligence. In particular, our prior work has shown that i) commonsense reasoning is a necessary capability for Visual Intelligence and also that ii) commonsense reasoning crucially requires the ability to reason about spatial relations between objects in the world. In this paper, we first recap our approach to Visual Intelligence in robotics, which is based on a hybrid architecture integrating a deep learning component with commonsense reasoning. We then present a framework for spatial reasoning, which has been designed to support the commonsense reasoning component in our architecture. Differently from prior approaches to qualitative spatial reasoning in robotics, the proposed framework is robust to variations in the robot's viewpoint and object orientation. In the paper, we also show how this formally-defined framework can be operationalised in an off-the-shelf spatial database.

Keywords

spatial reasoning, commonsense reasoning, cognitive robotics, visual intelligence

1. Introduction

In all cases where it is inconvenient or even dangerous for us to intervene, there is an incentive to delegate tasks to *service robots*: e.g., under the extreme conditions imposed by space explorations [1], in hazardous manufacturing environments [2], or whenever social distance needs to be maintained [3]. Another compelling use case for service robots is autonomously monitoring office environments to prevent potential threats to the Health and Safety (H&S) of employees. For instance, a power cable dangling in a corridor constitutes a trip hazard. Similarly, a sweater left to dry on top of a heater may cause a fire. To tackle these tasks, at the Knowledge Media Institute (KMi), we are developing HanS [4], the Health and Safety robot inspector.

Before delegating complex tasks to robots, however, we need to ensure that they can reliably *make sense* of the stimuli coming from their sensors. Autonomous sensemaking remains an open challenge, because it requires not only to reconcile the high-volume and diverse data collected from real-world settings [5], but also to actually understand these data, going beyond mere pattern recognition [6, 7].

From a vision perspective, the problem of robot sensemaking becomes one of enhancing the *Visual Intelligence* of service robots, i.e., their ability to make sense of the environment through

AIC 2022, 8th International Workshop on Artificial Intelligence and Cognition, June 15-17, 2022, Örebro, Sweden.

agnese.chiatti@open.ac.uk (A. Chiatti); gianluca.bardaro@open.ac.uk (G. Bardaro); enrico.motta@open.ac.uk (E. Motta); enrico.daga@open.ac.uk (E. Daga)

their vision system and epistemic competences [8]. Naturally, several epistemic competences are required to build *Visually Intelligent Agents (VIA)*. In HanS' case, the first prerequisite to detect the fire threat posed by a sweater lying on top of a heater is to recognise the sweater and the heater in question. Moreover, HanS also needs *spatial reasoning* capabilities, to infer that the sweater is touching the heater. It also needs to know that sweaters are made of cloth and that a piece of cloth clogging an electric radiator can catch fire. The list goes on.

In [8], we identified a framework of *epistemic requirements*, i.e., knowledge properties and reasoning capabilities, which are needed to develop Visually Intelligent Agents. We gauged these requirements from cognitive theories that characterise the excellence of the human vision system [9, 6]. These include the capability to track objects over time, to consider the spatial configuration and functional parts that compose an object, as well as the knowledge of the physical and material properties of objects, to name just a few [8]. Cognitively-inspired requirements were also grounded in the types of classification errors that emerge when Deep Learning (DL) is applied to real-world robotic scenarios. This error analysis highlighted that misclassifications could in principle have been avoided, if the robot was capable of considering: (i) the canonical size of objects, e.g., that mugs are generally smaller than bins, as well as (ii) the typical *Qualitative Spatial Relations (QSR)* between objects. For instance, a fire extinguisher may be mistaken for a bottle due to its shape. However, the proximity of a fire extinguisher sign is a strong indication that the observed object is in fact a fire extinguisher. This element of *typicality* relates to the broader objective of developing AI systems which can reason about what is *plausible*, i.e., which exhibit commonsense [7, 10, 11, 12, 13]. Our most recent results [14] demonstrate that an architecture which leverages the awareness of the typical sizes and spatial relations of objects can significantly augment object recognition methods based on DL. In this context, the capability to reason about the spatial configuration of objects is one of the requirements that contributes to autonomous sensemaking. Thus, in this paper, we propose a framework for spatial reasoning to support visual commonsense reasoning. The proposed framework is generically conceived to support *mobile ground robotic* applications, i.e., robots that perceive the environment whilst navigating it and that operate in contact with the ground.

Differently from the previous frameworks that have been proposed to link the geometrical and perceptual data collected by a robot to formally-defined spatial concepts [15, 16], the proposed framework can account for variations in the robot's viewpoint and in the relative orientation of objects. Importantly, the formally-defined QSR in this framework are also mapped to the type of linguistic predicates used to describe *commonsense spatial relations* in English, which are discussed within seminal theories of spatial cognition [17, 18]. Operationally, we realised the proposed framework in a concrete architecture that capitalises on state-of-the-art Geographic Information Systems (GIS). Ultimately, we demonstrate how the implemented framework can be successfully applied to extract qualitative spatial relations in HanS' use-case scenario.

2. Related work

Broadly speaking, spatial relations can be represented qualitatively - e.g., A contains B - or quantitatively - e.g., the angle between A and B is θ [19]. Following [16], Qualitative Spatial Relations (QSR) can be further characterised as (i) *metric*, i.e., based on the metric distance

between objects (ii) *topological*, i.e., describing the neighbourhood of objects, or (iii) *directional*, i.e., relative to the axis directions in a reference coordinate system. The interested reader is referred to [20] for a foundational review of qualitative spatial representations. Compared to quantitative representations, qualitative representations are more similar to the kind of spatial predicates involved in natural language discourse. As a result, qualitative spatial representations are easier to interpret and aid Human-Robot Interaction [21, 22, 23]. Moreover, they are more similar to the kinds of spatial predicates available within linguistic Knowledge Bases (KB) [24], as well as within benchmark datasets for visual reasoning tasks, such as Visual Genome [25] and SpatialSense [26]. Thus, relying on qualitative representations has the potential to facilitate the repurposing of these resources in robotic contexts, especially given the paucity of comprehensive KBs for Visually Intelligent Agents [8].

The problem of representing spatial relations has been actively researched for decades, producing many theoretical frameworks for spatial reasoning [27, 20, 28, 29]. In Robotics, extensive efforts have been devoted to linking the robot sensor data and symbolic knowledge to the geometric maps modelling its environment [30, 31, 32]. These efforts have produced intermediate representational models also known as *semantic maps*, i.e., maps that contain, "in addition to spatial information about the environment, assignments of mapped features to entities of known classes" [30]. To combine the best of both worlds, futher approaches have been proposed [15, 33, 34] where semantic maps are also interpreted with respect to formal spatial theories.

In general, spatial relations are expressed between object pairs, where one of the two objects is considered as a *reference*, or *landmark*: e.g., bike near house. Young et al. [34] have used Ring Calculus to represent the closeness of objects. The authors of [33] have relied on ternary point calculus [35] to model directional relations with respect to both the robot's location and the location of the reference object. Thus, they reduced 3D object regions to point-like objects on the 2D plane. Moreover, they assumed that the robot's location does not change over time, and is always defined with respect to a tabletop. Differently from [33], Deeken and colleagues [15] represented directional relations by comparing 3D object regions through the halfspace-based model of [16]. However, this model is based on the assumption that the robot's viewpoint is always aligned both with the global coordinate system of the map and with the inherent orientation of the observed objects. Thus, it is not suitable to model the case of mobile ground robots. In real-world scenarios, as the robot moves, its viewpoint changes over time and the objects observed will be oriented differently. To address this issue, we propose to combine the robot's viewpoint and the orientation of the reference object within a *contextualised frame of reference*. This contextualised frame of reference allows us to define a contextualised 3D region, or *Contextualised Bounding Box*, which represents the location of the object with respect to both the robot's viewpoint and the frame of reference of a landmark. Crucially, the contextualised frame of reference and Bounding Box can be defined for any combination of robot and landmark location, thus ensuring that this framework can scale to many real-world robotic scenarios. Hence, consistently with the design methodology recommended by related studies of spatial ontology engineering [36, 37], we handle the ambiguity of language by situating spatial predicates with respect to a geometric frame of reference. However, differently from upper-level ontologies of space [27, 28, 29], which attempt to characterise how humans conceptualize spatial concepts through language, the proposed representation is tailored to the spatial reasoning components that mobile service robots need for visual sensemaking.

3. Proposed Framework

To define a spatial reasoning framework which satisfies the requirements of robot sensemaking, we extend the formal theory of spatial reasoning presented in [16]. Moreover, we map the obtained spatial relations to the commonsense predicates used to describe spatial relations between objects in English. These predicates are gathered from cognitive theories [17, 18]. By making an explicit link between formal AI theories and informal linguistic representations, we obtain a framework for spatial reasoning that supports commonsense reasoning in robotic scenarios.

Notation. In what follows, we model definitions as First Order Logic (FOL) statements. We represent logic variables through lowercase letters and constants through uppercase letters. We also use lowercase initials to denote functions, while uppercase initials symbolise predicates. For instance, *sReg* is a function, whereas *Above* is a predicate. Unless otherwise stated, free variables are universally quantified.

Spatial primitives. Our domain of discourse $\mathbb{D}$ is that of *spatial objects*, i.e., physical objects, "which have spatial extensions" [20]. From this perspective, a spatial object is represented in terms of the associated *spatial region*. In particular, we represent spatial regions as sets of *spatial points*, p. Let P be the set of all spatial points, then, for each spatial object $o \in \mathbb{D}$, we assume the existence of a function $sReg$ which, given o, returns the subset of P which includes all the points in the spatial region of o.

$$\textit{SpatialObj}(o) \Rightarrow \textit{sReg}(o) \subseteq P \tag{1}$$

$$\textit{SpatialObj}(o) \Rightarrow \textit{sReg}(o) \neq \varnothing \tag{2}$$

In particular, our focus is not on arbitrary collections of spatial points, but rather on one-piece regions [20], i.e., on sets of internally connected points:

$$\textit{SpatialObj}(o) \Rightarrow \textit{ProperSR}(\textit{sReg}(o)) \tag{3}$$

To provide a formal definition of the concept of *proper spatial region*, we need first to establish a *spatial frame of reference*.

Spatial frame of reference. A spatial object is characterised not only with respect to a spatial region but also in terms of a reference coordinate system, also known as *frame of reference*. A frame of reference consists of an origin point, O, and of a set of directed axes intersecting at the origin. In particular, modelling the 3D space requires three reference axes, X, Y, Z. Once we have defined a reference frame, we can interpret spatial points as *geometrical points*, i.e., as coordinate triples in $\mathbb{R}^3$. Let GP be the set of all geometrical points in the considered space:

$$GP = \{p | p = (x, y, z) \in \mathbb{R}^3\} \tag{4}$$

The identified frame of reference also has an associated *granularity*, i.e., an infinitesimally small constant $D > 0$ in $\mathbb{R}$, which defines the minimum distance for two geometrical points to be considered as distinct entities. Two geometrical points are then said to be *adjacent* iff their geometrical distance is equal to D. To compute the distance between two geometrical points, they have to be in the same frame of reference. Let $d(gp, gp')$ be a function which returns a real number indicating the geometric distance between points gp and gp'. Then:

$$\textit{Adj}(gp, gp') \Leftrightarrow d(gp, gp') = D \tag{5}$$

The definition of proper spatial region then follows from the notion of adjacency:

$$\textit{ProperSR}(sr) \Leftrightarrow \forall gp[gp \in sr \Rightarrow \textit{Conn}(gp, sr)] \tag{6}$$

$$\textit{Conn}(gp, sr) \Leftrightarrow \forall gp'[gp' \in sr \wedge gp' \neq gp] \Rightarrow \textit{ConnP}(gp, gp') \tag{7}$$

$$\textit{ConnP}(gp_1,gp_2) \Leftrightarrow \textit{Adj}(gp_1, gp_2) \vee \exists gp_3[\textit{Adj}(gp_1, gp_3) \wedge \textit{ConnP}(gp_3, gp_2)] \tag{8}$$

In our model, we assume that the *global spatial region*, GP, is a fully-connected set of points. Moreover, we assume that spatial regions can be approximated through 3D boxes. This simplifying assumption is consistent with standard practice in the literature [15, 16]. Bounding boxes can have an arbitrary orientation around the Z axis aligned with gravity, but their base is always parallel to the XY plane, as exemplified in Figure 1. In particular, we consider the minimum bounding box which best approximates the real volume occupied by an object and which is aligned with its *natural orientation* [8]. Let b be a set of geometrical points which contains the spatial region of o:

$$\textit{BoundBox}(b, o) \Leftrightarrow \textit{sReg}(o) \subseteq b \wedge b \subseteq GP \tag{9}$$

$$\textit{MinBBox}(b, o) \Leftrightarrow \textit{BoundBox}(b, o) \wedge \neg\exists b'[\textit{BoundBox}(b', o) \wedge b' \subset b] \tag{10}$$

In this scenario, the environment navigated by a robot can also be modelled as a spatial region including an arbitrary number of objects, i.e., as a global spatial region. Consequently, the outer region of a spatial region, sr, is:

$$\textit{outReg}(sr) = \{gp | gp \in GP \wedge gp \notin sr\} \tag{11}$$

The frame of reference of the global region, F_g, is *extrinsic*, i.e., based on a reference point which is external to both an object and an observer. F_g remains fixed as the robot navigates the environment. Conversely, the robot's frame of reference, F_r, changes as the robot moves. Thus, it is *deitic*, relative to the observer's position. In [16, 15], all the spatial relations between objects are defined according to the same pre-defined frame of reference, whether it is an extrinsic, deitic or intrinsic one, i.e., inherent to a specific object. Unlike the latter spatial theories, linguistic spatial predicates implicitly refer both to (i) the location of the reference object, and to (ii) the observer's point of view [17]. Similarly, a robot would conclude that "A is on the left of B" based not only on the location of objects A and B within F_g, but also on F_r. From a different standpoint, A might appear on the right of B, or in front of it. To model such cases, we introduce the notion of *robot's viewpoint*, $F_{r'}$. Let C_o be the centroid of the spatial region representing object o. Then, $F_{r'}$ is obtained by rotating F_r along Z_r, by an angle α. Specifically, α is the angle between X_r and the imaginary line connecting the origin of F_r with C_o.

Let F_o of origin C_o and axes X_o, Y_o, Z_o be the intrinsic frame of reference of o, i.e., the frame of reference which is aligned with the orientation of o. Then, the contextualised frame of reference of the object, F_c, is the frame of reference of origin C_o whose axes have the same orientation of the axes defining the robot's viewpoint, $F_{r'}$ (Figure 1). Based on F_c, we can construct a *Contextualised Bounding Box (CBB)*, which is obtained by aligning the minimum bounding box with F_c. Let $\textit{rotZ}(b, \theta)$ be a function which returns the spatial region, sr, obtained by rotating an input bounding box along Z by an angle θ. Given a frame of reference F_c, then $\textit{yaw}(sr, F_c)$ returns the angle between the intrinsic frame of reference of sr and F_c, along Z. Then, given $\pi/2$:

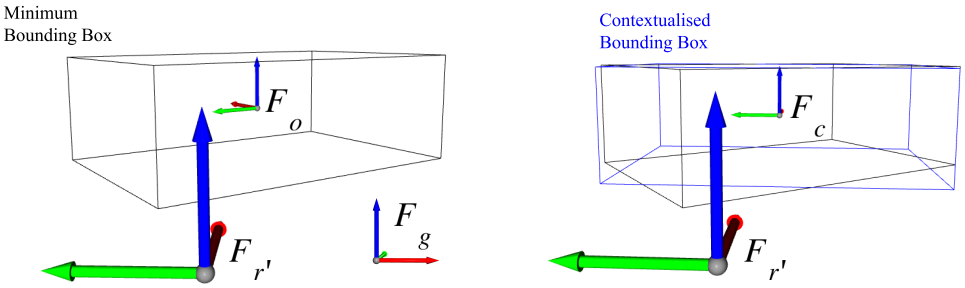


Figure 1: The robot's viewpoint, $F_{r'}$, consists of three axes $X_{r'}$ (in red), $Y_{r'}$ (in green) and $Z_{r'}$ (in blue). $F_{r'}$ may not coincide with the frame of reference of the global map, F_g, nor with the intrinsic frame of reference of the object, F_o. A spatial object is first modelled as the minimum 3D box bounding the object. Then, $F_{r'}$ is translated to the object's centroid to define a contextualised frame of reference F_c. Moreover, a Contextualised Bounding Box is generated (in blue), i.e., the bounding box which requires the minimum rotation along the Z axis to align the oriented bounding box with F_c.

$$\begin{aligned} \mathit{IsCBB}(\mathit{rotZ}(b,\theta),o) \Leftrightarrow \mathit{MinBBox}(b,o) \wedge \exists\theta[\mathit{mod}(\mathit{yaw}(\mathit{rotZ}(b,\theta),F_c),\pi/2) = 0 \wedge \\ \neg\exists\theta'[\mathit{mod}(\mathit{yaw}(\mathit{rotZ}(b,\theta'),F_c),\pi/2) = 0 \wedge \theta' < \theta]] \end{aligned} \quad (12)$$

Namely, to construct CBB, we select the minimum angle θ so that the value returned by the *yaw* function is divisible by $\pi/2$, i.e., the remainder of their division, *mod*, is zero. There are always four possible alignments of a bounding box, b, for which *mod* is zero. Thus, by selecting the minimum angle among these four, we apply the transformation which is least disruptive of the natural orientation of the object.

Metric spatial relations. Given two spatial objects o_1, o_2 and two geometrical points gp_1, gp_2 where $gp_1 \in o_1$ and $gp_2 \in o_2$, we define the distance between two geometrical points as the their Euclidean distance. Then, the distance between two spatial objects is defined as the global minimum of the pointwise distance function, d:

$$\begin{aligned} [\mathit{distance}(o_1,o_2) = d(gp_1,gp_2)] \Leftrightarrow gp_1 \in o_1 \wedge gp_2 \in o_2 \wedge \\ \forall gp_3, gp_4[gp_3 \in o_1 \wedge gp_4 \in o_2] \Rightarrow d(gp_3,gp_4) > d(gp_1,gp_2) \end{aligned} \quad (13)$$

A distance threshold, T, can be then introduced, to represent closeness between objects. That is, for a T greater than or equal to the frame granularity D defined earlier:

$$\mathit{IsClose}(o_1,o_2) \Leftrightarrow \mathit{distance}(o_1,o_2) \leqslant T \quad (14)$$

In particular, if the minimum distance equals D, then the two objects touch:

$$\mathit{Touches}(o_1,o_2) \Leftrightarrow \mathit{distance}(o_1,o_2) = D \quad (15)$$

Topological spatial relations Topological relations are spatial relations which are invariant under a topological isomorphism, i.e., a function $f : X \rightarrow Y$ which preserves neighbourhood relationships while mapping X to Y. Although a number of topological relations have been proposed [20], here we focus on the intersection and containment relations. As shown in the remainder of this Section, this minimal subset of relations, combined with metric and directional relations, is sufficient to cover all the spatial relations required in the scenario of interest. First,

based on our prior definitions, two spatial regions, sr, sr' intersect iff they have at least one geometrical point in common:

$$\textit{Int}(sr, sr') \Leftrightarrow \exists gp[gp \in sr \wedge gp \in sr'] \tag{16}$$

We also define the spatial region representing the intersection between two objects (i.e., the intersection between the associated spatial regions) as follows:

$$\textit{inter}(o1, o2) = \{gp | gp \in \textit{sReg}(o1) \wedge gp \in \textit{sReg}(o2)\} \tag{17}$$

Then, a special case of the intersection relation is the case where one spatial region completely contains the other:

$$\textit{ComplCont}(sr, sr') \Leftrightarrow \forall gp[gp \in sr' \Rightarrow gp \in sr] \tag{18}$$

Semantically, o contains o' completely iff all the geometrical points in the spatial region of o' are also members of the spatial region of o.

Directional spatial relations. Differently from metric and topological relations, directional spatial relations are dependent on the considered frame of reference. The spatial reasoning framework proposed by Deeken et al. [15] for robotic applications, which is based on the work in [16], models directional relations by partitioning the outer spatial region of an object into six halfspaces, i.e., one halfspace for each semi-axis of X, Y, Z. In particular, as in [15], halfspaces can be modelled as 3D extrusions, obtained by multiplying the extent of the object spatial region by a scaling factor $s \in \mathbb{R}$.

The coordinates of all geometrical points in the minimum bounding box are bound to a minimum and maximum value, e.g., x_{min} and x_{max}. Let X_o^+ and X_o^- be the positive and negative semi-axes of X_o in F_o. Then, we define a function, *hs*, which returns the halfspace of an input bounding box, given semi-axis, X_o^+, and frame of reference, F_o:

$$\begin{aligned} \textit{MinBBox}(mb_1, o_1) \Rightarrow \textit{hs}(mb_1, X_o^+, F_o) = \\ \{gp \in \textit{outReg}(mb_1) | gp = (x, y, z) \textit{ w.r.t } F_o, \\ x_{max} \leqslant x \leqslant x_{max} + x_{max} \cdot s, \\ y_{min} \leqslant y \leqslant y_{max}, \\ z_{min} \leqslant z \leqslant z_{max}\} \end{aligned} \tag{19}$$

Additional halfspaces can be similarly derived for the other semi-axes in F_o.Once these halfspaces have been defined, one can test whether a second object o_2 lies within any of the halfspaces of o_1. In particular, in the following, we consider "relaxed" (_r) spatial operators [16]. In other words, we infer directional relations by testing whether o_2 intersects the halfspaces of o_1. We use capital initials to represent predicates symbolising the *East*, *West*, *North*, *South*, *Above* and *Below* relations. Given a F_o which coincides with F_g, the relaxed definitions of *East*(o_2,o_1) is:

$$\textit{E\_r}(o_2, o_1, F_o) \Leftrightarrow \textit{MinBBox}(mb_1, o_1) \wedge \textit{Int}(\textit{hs}(mb_1, X_o^+, F_o), \textit{sReg}(o_2)) \tag{20}$$

The definitions of the remaining directional relations (i.e., *W_r*, *N_r*, *S_r*, *A_r*, *B_r*) are isomorphic to axiom 20 and are omitted for brevity. The model proposed in [15, 16] is based on the assumption that F_o is always aligned with F_g. However, this assumption does not hold in the case of mobile robot sensemaking. Indeed, the frame of reference of the robot, F_r is mobile, i.e., its origin and orientation change over time. Moreover, the natural orientation of objects may

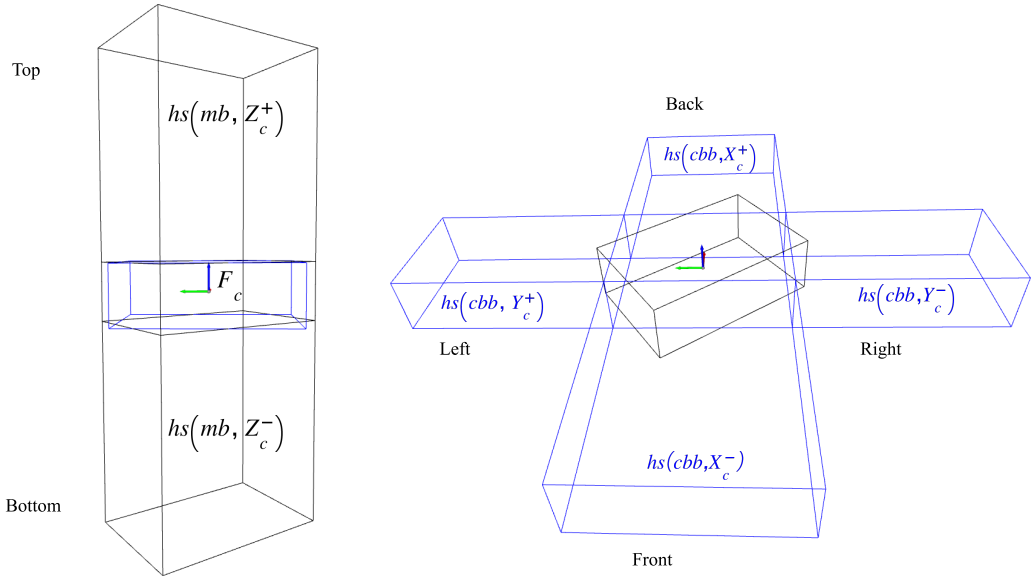


Figure 2: Halfspaces are generated by extruding the 3D bounding boxes. The top and bottom halfspaces are derived by extruding the minimum oriented bounding box along the Z axis (left-hand side of the Figure). The left, right, front and bottom halfspaces are instead extruded from the CBB (right-hand side of the Figure).

not be aligned with F_g. Thus, to produce a representational model which suits the case of robot sensemaking, we need to map directional relations to the contextualised frame of reference, F_c, which we defined earlier. Specifically:

$$\textit{Above}(o_2, o_1, F_c) \Leftrightarrow \textit{A\_r}(o_2, o_1, F_o) \tag{21}$$

$$\textit{Below}(o_2, o_1, F_c) \Leftrightarrow \textit{B\_r}(o_2, o_1, F_o) \tag{22}$$

Nonetheless, to model relations such as *RightOf* or *LeftOf*, we need to account for the robot's viewpoint. Thus, we apply the halfspace-based model to the Contextualised Bounding Box we have defined earlier (see Figure 2). By definition, CBB is aligned with the contextualised frame of reference, F_c, so the front halfspace of CBB, for instance, can be defined w.r.t. a given F_c as follows:

$$\begin{aligned} \textit{IsCBB}(cbb_1, o_1) \Rightarrow \textit{hs}(cbb_1, X_c^-, F_c) = \\ \{gp \in \textit{outReg}(cbb_1) | gp = (x, y, z) \; \textit{w.r.t.} F_c, \\ x'_{min} - x'_{min} \cdot s \leqslant x \leqslant x'_{min}, \\ y'_{min} \leqslant y \leqslant y'_{max}, \\ z'_{min} \leqslant z \leqslant z'_{max}\} \end{aligned} \tag{23}$$

These spatial constructs allow us to define the remaining directional relations:

$$\textit{RightOf}(o_2, o_1, F_c) \Leftrightarrow \textit{Int}(sr_2, \textit{hs}(cbb_1, Y_c^-, F_c)) \tag{24}$$

$$\textit{LeftOf}(o_2, o_1, F_c) \Leftrightarrow \textit{Int}(sr_2, \textit{hs}(cbb_1, Y_c^+, F_c)) \tag{25}$$

$$\textit{InFrontOf}(o_2, o_1, F_c) \Leftrightarrow \textit{Int}(\textit{sr}_2, \textit{hs}(\textit{cbb}_1, X_c^-, F_c)) \quad (26)$$

$$\textit{Behind}(o_2, o_1, F_c) \Leftrightarrow \textit{Int}(\textit{sr}_2, \textit{hs}(\textit{cbb}_1, X_c^+, F_c)) \quad (27)$$

For brevity, in axioms 24-27 we have omitted the predicate $\textit{IsCBB}(\textit{cbb}_1, o_1)$, which is always valid. Thanks to these newly-defined spatial concepts, we can now specify how the latter QSR align with linguistic spatial predicates. This mapping process is also known, in the qualitative spatial reasoning literature, as *qualification* [38].

Qualification. In English, objects are represented by nouns while the spatial relationships between objects are mainly represented through prepositions - e.g., on, next to, behind [17]. Spatial relations are also implied by using certain verbs (e.g., person wears shirt). However, almost invariably, these verbs can be reduced to a simplified form, followed by a preposition (e.g., person has shirt on). Hence, the canonical structure of a spatial sentence consists of three elements: (i) a *reference* object and (ii) a *figure* object, both expressed as noun phrases, as well as (iii) a spatial preposition. The reference object and the preposition, together, define the spatial region occupied by the figure object. As pointed out in [17], the object's *top* and *bottom* are defined as "the regions at the ends of whichever axis is vertical in the object's normal orientation" [17]. Thus, they are conceptually equivalent to the notion of top and bottom halfspaces we defined for the minimum oriented bounding box. Moreover, the object *front* is defined as the region at the end of the object's horizontal axis which also faces the observer. Conversely, the object *back* is located opposite to the observer along the same axis. Finally, the region at the end of any other horizontal axis can be called a *side*. As such, the geometric relations defined at axioms 21-22, and 24-27 are directly qualified through the *Above*, *Below*, *RightOf*, *LeftOf*, *InFrontOf* and *Behind* predicates. However, the *LeftOf* and *RightOf* relations can be further combined so that, given F_c:

$$\textit{Beside}(o_2, o_1, F_c) \Leftrightarrow \textit{RightOf}(o_2, o_1, F_c) \vee \textit{LeftOf}(o_2, o_1, F_c) \quad (28)$$

Furthermore, an object is said to be "near" another object if it is located in a region "extending up to some critical distance" [17]. This notion corresponds exactly to our definition of predicate *IsClose* (axiom 14).

An interesting case is that of the "on" preposition. One of the senses of "on" is semantically related to "above". However, while "above" typically implies absence of contact between the two objects, "on" strongly favours a contact reading [17]. Formally, we make this distinction by defining:

$$\textit{OnTopOf}(o_2, o_1, F_c) \Leftrightarrow \textit{Above}(o_2, o_1, F_c) \wedge \textit{Touches}(o_2, o_1) \quad (29)$$

Nonetheless, the "on" preposition can also be used to denote that the figure object is supported by the reference object. For instance, we say that a "clock is on the wall" although the two objects overlap horizontally. The phrase "clock on wall" also implies that the wall is adequately stable to support the clock. Indeed, if two objects differ in terms of size and mobility, we tend to consider the larger and more stable object as reference [17]. To differentiate these additional uses of "on", we define, for a given F_c:

$$\begin{aligned}\textit{LeansOn}(o_2, o_1, F_c) \Leftrightarrow \textit{Touches}(o_2, o_1) \wedge \neg\textit{Above}(o_2, o_1, F_c) \wedge \\ \neg\textit{Below}(o_2, o_1, F_c) \wedge \exists o_3[\textit{Touches}(o_2, o_3) \wedge \textit{Below}(o_3, o_2, F_c)]\end{aligned} \quad (30)$$

$$\textit{Touches}(o_2, o_1) \wedge \neg\textit{Above}(o_2, o_1, F_c) \wedge \neg\exists o_3 \textit{Touches}(o_3, o_2) \Rightarrow \textit{AffixedOn}(o_2, o_1, F_c) \quad (31)$$

Namely, whenever o_2 is supported by a reference object o_1 along the horizontal direction, it is typically said to be "leaning against" o_1: e.g., a ladder leaning against a wall. Furthermore, if the reference object o_1 provides the only support surface for o_2, o_2 is typically said to be *AffixedOn* o_1: e.g., a ladder which is affixed on the wall, above ground. Nonetheless, there may be cases where an object, o_2, is physically affixed to a surface, o_1, even though o_1 is not the only surface in contact with o_2: e.g., a ladder affixed at ground level. Hence, we used a single logic implication in Statement 31.

The 'in" preposition is polysemous [18]. First, "in" is generally used to imply that one object is "inside" another, or, based on our prior topological definitions, that one object is completely contained in the other (axiom 18). However, "in" is also used in cases where two objects only partially compenetrate each other. For instance, we would say that "a cat is in the box" even when the cat's tail is peeping from the box. To define this notion of partial containment, we define a function, *adjSRCard*, which, given two spatial regions, sr and sr', returns the cardinality of the set of points in sr' that are adjacent to points in sr:

$$\begin{aligned} \textit{adjSRCard}(sr, sr') = |\{gp' | gp' \in \textit{outReg}(sr) \wedge gp' \in sr' \\ \wedge \exists gp[gp \in sr \wedge \textit{Adj}(gp, gp')]\}| \end{aligned} \quad (32)$$

Hence, we can now define partial containment as follows:

$$\begin{aligned} \textit{PartIn}(o_1, o_2) \Leftrightarrow sr = \textit{inter}(\textit{sReg}(o_1), \textit{sReg}(o_2)) \wedge \\ \textit{adjSRCard}(sr, \textit{sReg}(o_1)) < \textit{adjSRCard}(sr, \textit{sReg}(o_2))] \end{aligned} \quad (33)$$

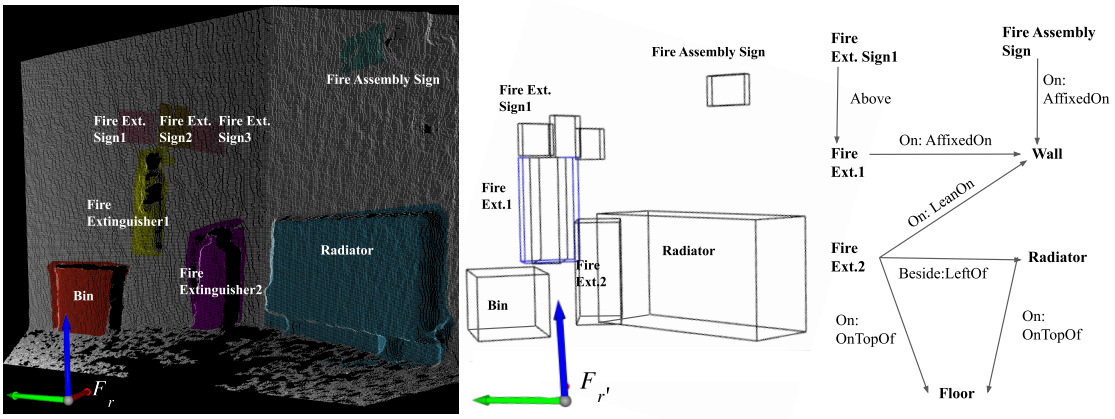
Namely, o_1 is partially contained in o_2 iff the number of points in o_1 that are adjacent to the intersection region of o_1 and o_2 is strictly smaller than the number of points in o_2 that are adjacent to the same intersection region.

4. Coverage Study and Framework Implementation

In this Section, we show how our framework for spatial reasoning can be implemented in practice. Specifically, we show that, once a set of basic spatial concepts has been derived through GIS operators, our framework provides a method to combine these basic spatial concepts to model the commonsense spatial predicates of [17].

Figure 3a shows an example of RGB-Depth (RGB-D) data collected through HanS' Orbbec Astra Pro monocular camera. At each time frame, t, the distance between the robot's pose and the surfaces reached by the laser in the depth sensor is measured. These data are also known as *depth images*, and can be converted to collections of 3D geometrical points in the considered frame of reference, i.e., to *PointClouds*. Consistently with [15], we store the object regions and labels in the semantic map within a spatial database, implemented in PostgreSQL. By linking these data to a spatial database, we can capitalise on the PostGIS engine and on the SFCGAL backend, that provide a series of query operators for spatial reasoning in the 3D space. Objects are stored in the PostGIS database using a minimum oriented polyhedron derived by applying the Convex Hull algorithm on the segmented PointCloud.

Furthermore, we complete the spatial database with 3D polygons representing the walls. To minimise the errors propagated from extracting the wall surfaces automatically, we developed a



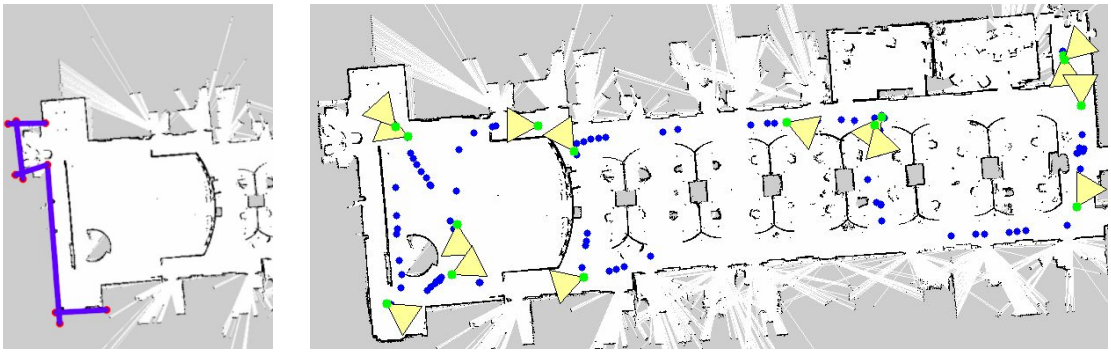
(a) (b) (c)

Figure 3: Example of operational workflow: (a) the PointCloud representing the observed scene is segmented and annotated with object categories. Then, (b) the minimum oriented bounding boxes and CBBs (in blue) are constructed. Lastly, (c) a set of QSR in figure-reference form is derived. In Figures 3b, 3c we show a subset of the bounding boxes and QSR representing the scene, for readability.

Graphic User Interface for annotating the wall edges on the 2D floorplan of the target environment (Figure 4a). For each added edge, a record is automatically added to the spatial database, indicating a wall surface. Namely, edges are extruded on the vertical axis by a fixed height, w_h, e.g., 4 meters in the case of our lab.

Once the spatial database has been populated, the contextualised bounding boxes and halfspaces introduced in the previous section can be derived (Figure 3b), by capitalising on PostGIS operators. Specifically, the mapping of spatial concepts to GIS operators is shown in Table 1a. Although neither PostGIS nor SFCGAL support 3D containment tests, we circumvent this limitation by comparing the volume of objects with the volume of their intersection region, through *ST_Volume*. Namely, if the volume of the intersection region equals the volume of the smaller object, e.g., o', then $ComplCont(o, o')$. To compute only QSR which are in figure-reference form, i.e., aligned with natural language [17], we sort objects by volume in descending order. Then, we only compute the QSR between one object and the objects which are larger than it. For instance, in Figure 3c, the QSR *fire extinguisher 1 AffixedOn wall* is extracted while the redundant *wall Behind fire extinguisher1* is avoided. In this way, we also reduce the computational load of extracting QSR for all pairwise object combinations.

In sum, PostGIS ensures a full coverage of the basic building blocks of our spatial framework. Then, the commonsense relations defined in Section 3 can be seen as a combination of these building blocks (Table 1b). The next step is evaluating how accurately these commonsense QSR can be extracted from robot-collected images.



(a) (b)

Figure 4: (a) The annotation tool, with walls marked as purple edges. (b) The complete robot route is traced with blue dots, while the evaluated points of interest are highlighted in green. The yellow triangles show the field of view of the robot.

5. Framework Evaluation

For evaluation purposes, we sampled 15 frames from a broader collection of RGB and Depth data that was previously collected during one of HanS' scouting routines. The broader overall dataset, which was collected through a Turtlebot mounting an Orbbec Astra Pro camera and consists of 1414 object regions, is publicly available, in addition to the implemented code, at https://github.com/kmi-robots/spatial-KB/tree/test. Images were collected from varying robot viewpoints and in unconstrained conditions of clutter. As a result, the object regions in the populated spatial database (Section 4) can partially occlude one another. From the overall set, we selected 15 frames, along the robot's route, that maximise the number of objects in the robot's field of view (Figure 4b). We then completed the selected sample with dense annotations of the spatial relations depicted in each scene. Consistently with our automated protocol for QSR extraction (Section 4), we only annotated QSR that are in figure-reference form. Overall, our evaluation sample is worth 268 spatial relations.

As summarised in Table 2, we tracked the Accuracy, Precision, Recall, and F1 metrics for each type of commonsense QSR in the considered set. For each spatial predicate, we compute evaluation metrics in binary terms: i.e., we assess whether a ground truth QSR was extracted or not, through our system. Overall, the proposed framework allowed us to model and correctly extract the majority of spatial relations depicted in the considered sample. Namely, 225 of the 268 ground truth QSR (84%) were successfully extracted. The *Near* and *Beside* predicates were excluded from our evaluation. Indeed, in the case of natural scenes, *Near* can be seen as a superclass of the QSR under evaluation, with the only exception of the *Above* relation, which does not entail closeness - e.g., *the sky is above*. Similarly, the study of the *Beside* predicate is subsumed by the evaluation of the more specific *LeftOf* and *RightOf* relations.

The measured F1 scores are equal to or greater than 80% for the majority of QSR types. The extraction of the *LeansOn*, *Below* and *Behind* relations was relatively more challenging. In the case of *LeansOn*, despite the high ratio of true positives (97% Recall), a higher number of *LeansOn* relations was generated compared to the ground truth. A visual inspection of the results

Input	GIS operators	Output
Convex Hull	*ST_OrientedEnvelope*, *ST_ZMin*, *ST_ZMax*, *ST_Extrude*	Min Oriented BBox
Min Orient BBox, Robot heading	*ST_Rotate*, *ST_Angle*, *ST_Centroid*	CBB
Min Orient BBox, s	*ST_Extrude*	Top/Bottom Halfspaces
CBB, s	*ST_Extrude*	L/R/Front/Back Halfspaces
Min Orient BBox	*ST_Volume*	Reference object set
Min Orient BBox	*ST_3DDWithin*	*IsClose*, *Touches*
Min Orient BBox	*ST_3DIntersects*	*Intersects (Int)*
Min Orient BBox	*ST_3DIntersection*	*inter*
Min Orient BBox	*ST_3DIntersection*, *ST_Volume*	*ComplCont*
Min Orient BBox Halfspaces	*ST_3DIntersection* *ST_Volume*	*Left/RightOf* *Above*, *Below* *InFrontOf*, *Behind*
Min Orient BBox	*ST_Scale*, *ST_Volume* *ST_Intersection*	*adjSRCard*

(a)

QSR	Follows from
Beside	*RightOf*, *LeftOf*
OnTopOf	*Touches*, *Above*
LeansOn	*Touches*, *Above* *Below*
AffixedOn	*Touches*, *Above*
Inside	*ComplCont*
PartIn	*inter*, *adjSRCard*
Near	*isClose*

(b)

Table 1
(a) Coverage of spatial notions through PostGIS operators. (b) The basic spatial relations covered by PostGIS are combined to derive more complex QSR.

revealed that these false positives were mainly caused by segmentation issues. In particular, a subset of object regions also include points of the occluding objects, as a result of deriving PointClouds from 2D-segmented masks. Moreover, a portion of true positives for the *Below* class was missed, due to approximating object regions as rectangular boxes. For instance, in the case of a desktop computer below a desk, the bounding box representing the desk also includes the hollow space between the legs. Hence, objects lying under the desktop do not intersect the bottom halfspace. The lowest accuracy score is associated to the *Behind* relation. We can ascribe this result to the fact that PointClouds were derived from individual depth images, without reconstructing regions behind the surfaces which are reached by the laser sensor. Additional causes of errors that were discovered from visually inspecting the results include: (i) inaccurate sensor measurements, yielding noisy object regions, as well as (ii) misalignments between wall annotations on the 2D map and the object-wall distance measured through the depth sensor. In sum, many resulting errors are related to the problem of accurately modelling object regions in real-world environments, rather than to the system's ability to infer spatial relations from object regions. Indeed, despite the challenges posed by this realistic robotic scenario, the proposed framework ensured an average F1 score of 83,1% across the evaluated relation types.

Our further experiments on the complete image set [14] show that realising a spatial reasoning module that adheres to the proposed framework significantly enhances the robot's ability to

	Avg.	*OnTopOf*	*LeansOn*	*AffixedOn*	*LeftOf*	*RightOf*	*Above*	*Below*	*InFrontOf*	*Behind*
Accuracy	71.6	88.3	65.3	74.2	66.7	75.0	81.5	63.6	73.0	56.5
Precision	85.3	94.6	66.7	85.2	90.9	100.	84.6	87.5	81.8	76.5
Recall	82.5	93.0	97.0	85.2	71.4	75.0	95.6	70.0	87.1	68.4
F1	83.1	93.8	79.0	85.2	80.0	85.7	89.8	77.8	84.4	72.2

Table 2
Evaluation of the extracted QSR, with metrics in percentages.

recognise objects, particularly when spatial awareness is coupled with the knowledge of object sizes.

6. Conclusion and Future Work

In this paper, we have presented a framework for spatial reasoning which satisfies the requirements of robot sensemaking in real-world scenarios. Differently from prior approaches to qualitative spatial reasoning in robotics, this framework is robust to variations in the robot's viewpoint and object orientation, thus ensuring scalability to many application scenarios. Crucially, this framework contributes a cognitively-inspired conceptual layer on top of geometrical spatial operators, to model commonsense spatial predicates. The resulting linguistic predicates facilitate the integration of background spatial knowledge from external resources. As such, the proposed framework contributes to the broader objective of developing Visually Intelligent Agents, which can reliably assist us with our daily tasks. Conveniently, the proposed framework can be fully implemented with state-of-the-art GIS technologies. Moreover, it led to the accurate extraction of 84% of the spatial relations from real-world images collected by a robot in the context of autonomous Health and Safety monitoring.

Because our framework was built on previous works that have modelled QSR with crisp spatial definitions, it does not capture uncertainty. In [14] we have introduced a definition of typicality of the QSR, to accompany spatial relations with belief scores based on background knowledge. However, this representational frame could be further extended by introducing fuzzy-logic statements. Moreover, the proposed QSR are designed to model situations at individual time frames. For instance, a clock may be affixed to the wall at time $t1$ and lie on the floor at time $t2$. Therefore, in our future work, we also aim at examining the spatio-temporal evolution of these relations.

References

[1] P. Nilsson, S. Haesaert, R. Thakker, K. Otsu, C.-I. Vasile, A.-A. Agha-Mohammadi, R. M. Murray, A. D. Ames, Toward specification-guided active mars exploration for cooperative robot teams, Robotics: Science and Systems (RSS) (2018).

[2] H. Liu, L. Wang, Remote human–robot collaboration: A cyber–physical system application for hazard manufacturing environment, Journal of manufacturing systems 54 (2020) 24–34.

[3] G. Yang, H. Lv, Z. Zhang, L. Yang, J. Deng, S. You, J. Du, H. Yang, Keep healthcare workers safe: application of teleoperated robot in isolation ward for covid-19 prevention and control, Chinese Journal of Mechanical Engineering 33 (2020) 1–4.

[4] E. Bastianelli, G. Bardaro, I. Tiddi, E. Motta, Meet hans, the heath & safety autonomous inspector., in: International Semantic Web Conference, Demo track, 2018.
[5] M. B. Alatise, G. P. Hancke, A review on challenges of autonomous mobile robot and sensor fusion methods, IEEE Access 8 (2020) 39830–39846.
[6] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, S. J. Gershman, Building machines that learn and think like people, Behavioral and Brain Sciences 40 (2017).
[7] E. Davis, G. Marcus, Commonsense reasoning and commonsense knowledge in artificial intelligence, Communications of the ACM 58 (2015) 92–103.
[8] A. Chiatti, E. Motta, E. Daga, Towards a Framework for Visual Intelligence in Service Robotics: Epistemic Requirements and Gap Analysis, in: Proceedings of KR 2020- Special session on KR & Robotics, IJCAI, 2020, pp. 905–916.
[9] D. D. Hoffman, Visual intelligence: How we create what we see, WW Norton & Company, 2000.
[10] H. Levesque, Common Sense, the Turing Test, and the Quest for Real AI, The MIT Press, 2017.
[11] P. J. Hayes, The Second Naive Physics Manifesto. Formal theories of the common sense world, Ablex Publishing Corporation, 1988.
[12] J. McCarthy, et al., Programs with common sense, RLE and MIT computation center, 1960.
[13] A. Newell, The knowledge level, Artificial intelligence 18 (1982) 87–127.
[14] A. Chiatti, E. Motta, E. Daga, Robots with commonsense: Improving object recognition through size and spatial awareness, in: To appear in Proceedings of the AAAI 2022 Spring Symposium on Machine Learning and Knowledge Engineering for Hybrid Intelligence (AAAI-MAKE 2022), CEUR, 2022.
[15] H. Deeken, T. Wiemann, J. Hertzberg, Grounding semantic maps in spatial databases, Robotics and Autonomous Systems 105 (2018) 146–165.
[16] A. Borrmann, E. Rank, Query Support for BIMs using Semantic and Spatial Conditions, in: Handbook of Research on Building Information Modeling and Construction Informatics: Concepts and Technologies, IGI Global, 2010, pp. 405–450.
[17] B. Landau, R. Jackendoff, "What" and "where" in spatial language and spatial cognition, Behavioral and Brain Sciences 16 (1993) 217–238.
[18] A. Herskovits, Language and spatial cognition, volume 12, Cambridge University Press, 1986.
[19] A. Thippur, C. Burbridge, L. Kunze, M. Alberti, J. Folkesson, P. Jensfelt, N. Hawes, A comparison of qualitative and metric spatial relation models for scene understanding, in: 29th AAAI Conference and the 27th Innovative Applications of Artificial Intelligence Conference (IAAI), volume 2, AI Access Foundation, 2015, pp. 1632–1640.
[20] A. G. Cohn, J. Renz, Qualitative Spatial Representation and Reasoning, in: Foundations of Artificial Intelligence, volume 3 of *Handbook of Knowledge Representation*, Elsevier, 2008, pp. 551–596.
[21] G. Sarthou, R. Alami, A. Clodic, Semantic Spatial Representation: a unique representation of an environment based on an ontology for robotic applications, in: Combined Workshop on Spatial Language Understanding (SpLU) and Grounded Communication for Robotics (RoboNLP)., 2019, p. 12.
[22] E. A. Sisbot, J. H. Connell, Where is My Stuff? An Interactive System for Spatial Relations,

arXiv:1909.06331 [cs] (2019). URL: http://arxiv.org/abs/1909.06331, arXiv: 1909.06331.
[23] A. Thippur, J. A. Stork, P. Jensfelt, Non-parametric spatial context structure learning for autonomous understanding of human environments, in: 2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), 2017, pp. 1317–1324.
[24] S. Storks, Q. Gao, J. Y. Chai, Recent advances in natural language inference: A survey of benchmarks, resources, and approaches, arXiv preprint arXiv:1904.01172 (2019).
[25] R. Krishna, Y. Zhu, O. Groth, J. Johnson, et al., Visual genome: Connecting language and vision using crowdsourced dense image annotations, International journal of computer vision 123 (2017) 32–73.
[26] K. Yang, O. Russakovsky, J. Deng, SpatialSense: An Adversarially Crowdsourced Benchmark for Spatial Relation Recognition, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Seoul, Korea (South), 2019, pp. 2051–2060.
[27] J. A. Bateman, J. Hois, R. Ross, T. Tenbrink, A linguistic ontology of space for natural language processing, Artificial Intelligence 174 (2010) 1027–1071.
[28] P. Grenon, B. Smith, Snap and span: Towards dynamic spatial ontology, Spatial cognition and computation 4 (2004) 69–104.
[29] A. C. Varzi, Spatial Reasoning and Ontology: Parts, Wholes, and Locations, in: M. Aiello, I. Pratt-Hartmann, J. Van Benthem (Eds.), Handbook of Spatial Logics, Springer Netherlands, Dordrecht, 2007, pp. 945–1038.
[30] A. Nüchter, J. Hertzberg, Towards semantic maps for mobile robots, Robotics and Autonomous Systems 56 (2008) 915–926.
[31] S. Coradeschi, A. Saffiotti, An introduction to the anchoring problem, Robotics and autonomous systems 43 (2003) 85–96.
[32] I. Kostavelis, A. Gasteratos, Semantic mapping for mobile robotics tasks: A survey, Robotics and Autonomous Systems 66 (2015) 86–103.
[33] L. Kunze, C. Burbridge, M. Alberti, A. Thippur, J. Folkesson, P. Jensfelt, N. Hawes, Combining top-down spatial reasoning and bottom-up object class recognition for scene understanding, in: 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2014, pp. 2910–2915.
[34] J. Young, L. Kunze, V. Basile, E. Cabrio, N. Hawes, B. Caputo, Semantic web-mining and deep vision for lifelong object discovery, in: 2017 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2017, pp. 2774–2779.
[35] R. Moratz, M. Ragni, Qualitative spatial reasoning about relative point position, Journal of Visual Languages & Computing 19 (2008) 75–98.
[36] J. Bateman, Situating spatial language and the role of ontology: Issues and outlook, Language and Linguistics Compass 4 (2010) 639–664.
[37] C. Eschenbach, Geometric structures of frames of reference and natural language semantics, Spatial Cognition and Computation 1 (1999) 329–348.
[38] G. D. Felice, P. Fogliaroni, J. O. Wallgrün, A hybrid geometric-qualitative spatial reasoning system and its application in gis, in: International Conference on Spatial Information Theory, Springer, 2011, pp. 188–209.