Incrementally Grounding Expressions for Spatial Relations between Objects

Tiago Mota[1] and **Mohan Sridharan**[2]
[1] The University of Auckland, NZ
[2] University of Birmingham, UK
tmot987@aucklanduni.ac.nz, m.sridharan@bham.ac.uk

Abstract

Recognizing, reasoning about, and providing understandable descriptions of spatial relations between objects is an important task for robots interacting with humans. This paper describes an architecture for incrementally learning and revising the grounding of spatial relations between objects. Answer Set Prolog, a declarative language, is used to represent and reason with incomplete knowledge that includes prepositional spatial relations between scene objects. A generic grounding of prepositions for spatial relations, human input (when available), and non-monotonic logical inference, are used to infer spatial relations between 3D point clouds in given scenes, incrementally acquiring a specialized metric grounding of the prepositions and the relative confidence associated with each grounding. The architecture is evaluated on a benchmark dataset of tabletop images and on complex simulated scenes of furniture.

1 Introduction

Robots[1] deployed to assist humans in complex domains have to reason with incomplete knowledge of domain objects and relations between them. Also, both sensing and actuation are unreliable on robots. These problems are partially offset by the robot's ability to sense and interact with the domain and humans, using the corresponding observations to revise the existing knowledge. Since humans may not have the time or expertise to provide comprehensive feedback, the robot can learn more effectively by referring to objects or events in terms of other known objects. For instance, in Figure 1a, asking "what is behind the cereal box?" directs the human's attention to the box of crisps. This paper focuses on reasoning with such spatial relations between objects, and on incrementally acquiring the *grounding* (i.e., meaning in the physical world) of words that describe these relations. The ability to accurately infer spatial relations improves performance in other tasks, e.g., scene understanding [Thippur *et al.*, 2015].

Spatial relations are often described using prepositions, i.e., words such as *above*, *below*, *behind*, and *in*. To reason with these prepositions, the robot needs a vocabulary and a grounding of these words, e.g., a mapping of these words to 3D regions or distances from reference points or objects. This grounding has to be revised over time in dynamic domains to account for factors such as sensing errors and changes in viewpoint. A robot with an incorrect grounding of spatial relations is likely to make decisions that are incorrect or suboptimal. The architecture described in this paper seeks to address these challenges and has the following characteristics:

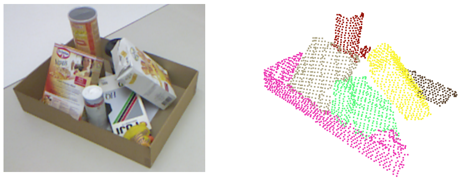

(a)
(b)

Figure 1: (a) Illustrative image of scene with objects; and (b) segmented version with 3D point clouds of objects in different colors.

- A declarative language is used to represent incomplete domain knowledge, which includes spatial relations between objects based on a generic (initial) grounding of prepositions in the 3D regions around objects.
- Non-monotonic logical inference with the existing knowledge, and human input (when available), are used to infer spatial relations between point clouds in new scenes, incrementally learning a specialized, histogram-based grounding of prepositions.
- Human input (when available) is also used to incrementally compute the relative accuracy of spatial relations inferred by the generic and specialized groundings, using the more reliable grounding for subsequent scenes.

In this paper, we consider (as input) 3D point clouds of objects in a scene, e.g., Figure 1b, and a generic grounding of prepositions for seven position-based and three distance-based relations. Learning corresponds to the incremental acquisition and revision of histograms as specialized grounding of these relations. We do not explicitly represent the uncertainty in processing visual input; any conclusion drawn with high probability is elevated to a logic statement with

[1] Terms "robot" and "agent" are used interchangeably.

complete certainty. Thus, our architecture enables robots to (a) infer spatial relations using a generic, manually-encoded grounding; (b) incrementally acquire a specialized grounding of spatial relations from a small number of examples; and (c) determine the relative confidence in each grounding and use the more reliable grounding for subsequent inference. We evaluate these capabilities on a benchmark dataset of tabletop objects and complex, simulated scenes of furniture.

2 Related Work

Existing approaches for grounding and interpreting the spatial relations between objects are broadly based on the use of manually encoded rules, or on training or learning algorithms. With manually-encoded rules, the vocabulary of spatial relations is grounded using *Qualitative Spatial Representations* (QSR) [Ye and Hua, 2013; Zampogiannis *et al.*, 2015; Elliott and Vries, 2015]. These approaches often approximate objects as points or establish rigid boundaries between spatial relations, and may not estimate spatial relations accurately. Moreover, the spatial relations are encoded in advance, but the grounding of these relations is likely to change over time in dynamic domains. Approaches that seek to learn the spatial relations or their grounding do so based on *Metric Spatial Representations* (MSR), i.e., measures such as angles and distances between objects. MSR have been used in approaches for different applications, e.g., for predicting the success of a robot's action in a previously unseen scenario [Fichtl *et al.*, 2015], and to learn relations between objects and generalize them to new objects [Mees *et al.*, 2017]. Other work has focused on choosing appropriate prepositions to describe the content of an image [Belz *et al.*, 2015]. In the context of human-robot interaction, existing systems have executed actions on objects and answered queries about spatial positions [Guadarrama *et al.*, 2013], compared QSR and MSR for scene understanding [Thippur *et al.*, 2015], and used MSR and a kd-tree to dynamically infer spatial relations [Ziaeetabar *et al.*, 2017]. However, these systems learn the grounding of spatial relations offline or in a separate training phase. In contrast, our approach starts with a hand-designed generic grounding, and incrementally and interactively learns a specialized grounding from experience and feedback.

In recent years, there has been considerable work on recognizing objects and inferring their spatial relationships from images and natural language expressions, e.g., for navigation and manipulation [Paul *et al.*, 2016; Pronobis and Rao, 2017; Shridhar and Hsu, 2017]. Since these system use neural (or deep) network architectures, they require a large number of training examples, learn the grounding offline, and are computationally expensive. Our architecture, on the other hand, combines the complementary strengths of non-monotonic logical inference, QSR, MSR, and interactive learning, to ground spatial relations from a small number of images and some human feedback.

3 Proposed Architecture

Figure 2 shows an overview of the key components of the architecture. We consider seven position-based prepositions (*in, above, below, front, behind, right, left*) and three distance-based prepositions (*touching, not-touching, far*), These prepositions are used to encode spatial relations between specific scene objects as logic statements in Answer Set Prolog (ASP), a declarative programming paradigm. The QSR module provides an initial, manually-encoded, generic grounding of spatial relations, which is used to extract spatial relations between pairs of 3D point clouds of each input scene (the "new observation"). Human feedback, when available, is also used to label the spatial relations between any pair of point clouds in a scene. Both the QSR-based output and human feedback are transmitted by the control node to the MSR module, which incrementally acquires and revises the MSR-based grounding of prepositions in the form of histograms. Assuming human feedback to be accurate, the control node also computes the relative trust in the QSR and MSR groundings. The more reliable grounding is used to extract logic statements representing spatial relations between scene objects in subsequent images; these are added to ASP program. Individual components are described below.

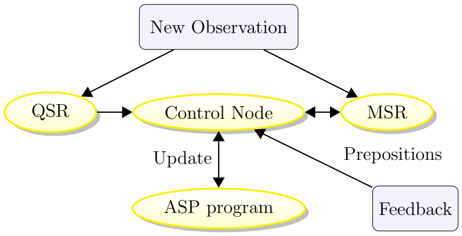


Figure 2: Proposed architecture.

Our architecture includes other modules, e.g., the 3D point cloud of a scene is sub-sampled and the Euclidean cluster extraction segmentation algorithm [Rusu, 2010][2] is used to segment the point cloud into objects. This algorithm is accurate for well-separated objects, and our system can recover from segmentation errors for occluded objects. These modules are not the focus of this work and are not discussed below.

3.1 Domain Representation in ASP

To represent and reason with incomplete knowledge, we use Answer Set Prolog (ASP), a declarative language that can represent recursive definitions, defaults, causal relations, special forms of self-reference, and language constructs that occur frequently in non-mathematical domains, and are difficult to express in classical logic formalisms. ASP is based on the stable model semantics [Gelfond and Kahl, 2014].

An ASP *program* (Π) has a *sorted signature* Σ and axioms. Σ includes *sorts* such as $object$, $location$, $color$, $shape$, and $step$ (for temporal reasoning); *statics*, i.e., domain attributes that do not change over time; and *fluents*, i.e., domain attributes whose values can be changed. In our case, the spatial relations are fluents such as:

$$\begin{aligned} &in(object, object), \quad above(object, object), \qquad (1)\\ &touching(object, object), \quad left(object, object). \end{aligned}$$

which are described in terms of their arguments' sorts. We choose the second argument of each such relation as the ref-

[2] Available at *www.pointclouds.org* for download.

erence object. In addition, predicate $holds(fluent, step)$ implies that a particular fluent holds true at a particular timestep.

The axioms of Π encode some rules to infer relations based on the spatial relations whose grounding is acquired:

$$\begin{aligned} holds(above(A,B),I) &\leftarrow holds(below(B,A),I). \\ holds(under(A,B),I) &\leftarrow holds(touch(A,B),I), \\ &\qquad holds(below(A,B),I). \end{aligned} \tag{2}$$

where the second axiom says that any object A that is *below* object B and touching it is considered to be *under* it. When action effects are to be modeled, the signature and axioms include *actions* with their preconditions and effects; a *history* of observations and executed actions is also considered. Since we do not currently need these capabilities, we do not describe them below. The ground literals in an *answer set* obtained by solving Π represent beliefs of an agent associated with Π. All reasoning (e.g., planning and inference) can be reduced to computing answer sets of Π [Gelfond and Kahl, 2014]. We use the SPARC system [Balai *et al.*, 2013] to compute answer set(s) of ASP programs.

The ASP-based representation of knowledge has some advantages. It supports concepts such as *default negation* (negation by failure) and *epistemic disjunction*. Unlike "$\neg$a", which implies that "*a is believed to be false*", "not a" only implies that "*a is not believed to be true*"; unlike "p $\vee$ $\neg$p" in propositional logic, "p or $\neg$p" is not tautological. Each literal can be true, false or unknown, i.e., the agent does not have to believe anything that it is not forced to believe. Also, unlike classical first-order logic, ASP supports non-monotonic logical reasoning, i.e., adding a statement can reduce the set of inferred consequences, aiding in the recovery from errors due to the incomplete knowledge. Modern ASP solvers support efficient reasoning in large knowledge bases, and are used by an international research community.

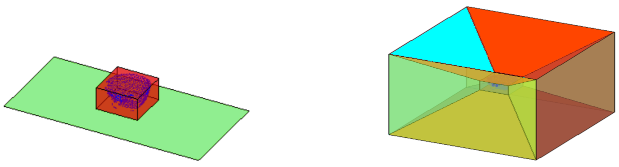
3.2 Qualitative Spatial Representation

Our QSR model is similar to that proposed by [Zampogiannis *et al.*, 2015]. For any given 3D point cloud, a bounding box containing it (i.e., convex cuboid around the object) is created—see Figure 3a. If this point cloud is considered the reference object, the space around this object is divided into non-overlapping pyramids representing the relations *left, right, front, behind, above* and *below*—see Figure 3b. In our implementation, the spatial relation of an object with respect to a reference object is determined by the non-overlapping pyramid around the reference that has most of the point cloud of the object. Also, any object with most of its point cloud located inside the bounding box of the reference object is said to be *in* the reference object. This definition of *in* can lead to errors, especially in domains with non-convex objects, e.g., a book that is actually *under* a large table may be classified (incorrectly) as being *in* the table because the bounding box of the table envelopes most of the point cloud of the book.

For ease of representation, our approach differs from [Zampogiannis *et al.*, 2015] in the definition of the distance-related prepositions: *touching, not-touching* and *far*. For a pair of point cloud clusters, the 10% closest distances between pairs of points drawn from the point clouds are computed, and the following criteria determine if the two objects are touching, not touching, or distinctly separated (i.e., far) from each other:

$$\begin{aligned} touching &\Rightarrow distance(10\%) \leq 0.01 \\ not\text{-}touching &\Rightarrow 0.01 < distance(10\%) < 1.0 \\ far &\Rightarrow distance(10\%) \geq 1.0 \end{aligned} \tag{3}$$

where distances are measured in meters, i.e., two objects are touching if the 10% closest distances are less than or equal to $1cm$. Although the generic, manually-encoded grounding based on the QSR model does not change over time, it is used by the robot to identify spatial relations between objects. This is based on the reasonable assumption that the robot has an initial idea of its camera's pose with respect to the scene. Next, we describe a specialized grounding of spatial relations that can be acquired over time.



(a)

(b)

Figure 3: (a) Bounding box for point cloud of a particular object; and (b) Pyramids delimiting space around the bounding box.

3.3 Metric Spatial Representation

MSR-based grounding of the spatial relations is also used to identify spatial relations between objects. Unlike the QSR-based grounding, the MSR model supports incremental updates from observations and human feedback.

Assume temporarily that the MSR module receives a pair of point cloud clusters corresponding to two objects, and the prepositions of the spatial relations between the objects, e.g., from QSR or humans. Our MSR module grounds each preposition using histograms, also referred to as "visual words", which are created by considering the point cloud data in a spherical coordinate system—each point is represented by its distance to a reference point and two angles (i) $\theta \in [0°, 180°]$; and (ii) $\varphi \in [-180°, 180°]$. On a robot, the coordinate frame for grounding is defined with respect to the robot's coordinate frame, its camera, and/or reference objects—information in one coordinate frame can be transformed to other coordinate frames. Also, sensor input processing introduces noise, but the non-monotonic logical reasoning and incremental learning modules of our architecture enable elegant recovery from errors due to noise.

We ground each of the seven position-based prepositions (*in, left, right, front, behind, above, below*) as 2D histograms of angles θ and φ, whereas each of the three distance-based prepositions (*touching, not-touching, far*) are ground using 1D histograms of the 10% closest distances between points in pairs of objects. Figures 4 and 5 show a distance and position histogram respectively. All histograms are normalized to ensure that large objects with many points do not have any undue influence on the grounding of relations.

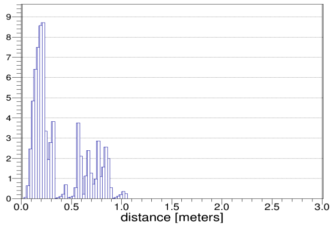


Figure 4: Example of 1D histogram grounding “not-touching”.

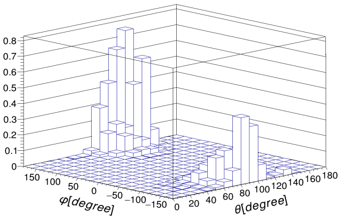


Figure 5: Example of 2D position histogram grounding “left”.

Any learned MSR-based grounding(s) are used on new scenes. For any given pair of point cloud clusters in a new scene, the corresponding 2D and 1D histograms (i.e., visual words) are constructed. The learned visual words that are most similar to the extracted visual words are used to assign the distance-based and position-based spatial relations between the corresponding scene objects, e.g., “$object_1$ is below $object_2$ and not touching it”. These inferred spatial relations are automatically translated to statements added to the ASP program, e.g., $below(obj_1, obj_2)$. Since axioms in the ASP program are applied recursively, each point cloud cluster only needs to be considered once.

The similarity between visual words is computed using the *intersection* measure for 1D (distance) histograms. For the 2D (position) histograms, we use the χ^2 measure, e.g., for any two histograms H and G:

$$D_{\chi^2}(H,G) = \sum_i \frac{|h_i - g_i|^2}{2(h_i + g_i)} \tag{4}$$

where h_i and g_i are bins in H and G respectively; larger values denote greater similarity. We use this measure for 2D histograms because the boundaries between the position-based relations are more difficult to define than those between distance-based relations. Once the spatial relations between a pair of point cloud clusters have been determined in a new scene, this information updates the learned visual words using a standard normalized histogram merging approach, i.e., the *MSR-based grounding is updated continuously*.

3.4 Combined Model and Other Relations

Recall that in addition to ASP-based inference using QSR and MSR groundings, spatial relations between point cloud clusters can also be determined by human feedback. While the QSR-based grounding remains unchanged and the MSR-based grounding changes as new scenes are processed, human input is assumed to be accurate, i.e., each human participant providing feedback is expected to be able to interpret spatial relations correctly. Since the QSR-based and MSR-based groundings may disagree on the relation between some pairs of objects, the control node initially assigns high (low) confidence to the QSR-based (MSR-based) grounding. The relative confidence in each grounding is then updated based on the number of times the output from the grounding matches human input—the more reliable grounding is used for subsequent scenes. Incorrect human annotation can thus affect the confidence in a grounding only if the number of such annotations is comparable to the number of correct annotations.

Object shapes and sizes may also influence spatial relations depending on the viewpoint. However, since the MSR-based grounding is based on histograms of relative distances and angles, it can be used to infer spatial relations over a range of viewpoints. Also, the architecture has two mechanisms to limit and recover from errors. If the QSR-based grounding is applicable, e.g., viewpoint has not changed substantially, the system can use it to obtain an initial estimate of spatial relations and incrementally acquire the MSR-based grounding. If the QSR-based grounding is not applicable, it is still possible to acquire an MSR-based grounding from human input and use it for subsequent inference. Furthermore, the MSR-based grounding is obtained from a small number of images and is transferable, as described in Section 4.

There are some important caveats related to the proposed approach. First, the QSR-based grounding is assumed to be reasonably accurate initially; if this assumption does not hold and no human input is available, an inaccurate MSR-based grounding may be acquired, resulting in incorrect estimates of spatial relations. Second, human feedback improves the specialized grounding (MSR) and overall accuracy, but it is not essential for estimating spatial relations. Third, the encoded prepositions (with learned groundings) are translated to logic statements (i.e., observation literals) in an ASP program. These observations and the commonsense knowledge encoded in the ASP program limit possible relations between scene objects and help infer composite relations (e.g., *on, close to, next to* etc). For instance, the spatial relation *on* may be defined by the axiom:

$$on(O_1, O_2) \leftarrow above(O_1, O_2),\ touching(O_1, O_2) \tag{5}$$

which states that if object O_1 is above O_2 and touching it, then O_1 is on O_2. It is also possible to learn such axioms interactively, as demonstrated by some recent work [Sridharan and Meadows, 2017]. Finally, we currently assume that each pair of objects is related through one position-based and one distance-based spatial relation, but not all the prepositions are (or need to be) mutually exclusive.

4 Experimental Setup and Results

In this section, we describe the experimental setup and the results of experimental evaluation.

4.1 Experimental Setup

For experimental evaluation, we used the Table Object Scene Database (TOSD)[3] and simulated scenes. TOSD contains

[3] https://repo.acin.tuwien.ac.at/tmp/permanent/TOSD.zip

111 scenes for training and 131 scenes for testing—many scenes include complex object configurations (Figure 1a), while some scenes have only two objects (Figure 6a). Since TOSD includes segmentation labels but not spatial relation labels, we manually labeled 200 scenes. In addition, simulation scenes were generated with a real-time physics engine (Bullet physics library) by manually encoding the grounding of spatial relations. Different subsets of 21 household objects from the Yale-CMU-Berkeley dataset [Calli *et al.*, 2015], along with a table and a shelf, were used to create 1400 simulated scenes (200 for each preposition). An additional 25 labeled scenes for each preposition (175 total) were used for training. Experiments tested two hypotheses:

<u>H1</u> the proposed approach enables more effective use of human feedback;

<u>H2</u> the combination of the manually-encoded QSR grounding and the automatically-learned MSR grounding performs better than each grounding used individually.

The performance measure was the accuracy of the labels assigned to spatial relations. We also qualitatively evaluated the ability to identify and correct errors. Below, *all claims are statistically significant at the* 95% *significance level*.

4.2 Experimental Results

The first set of experiments was designed as follows, with the results summarized in Table 1:

1. Pairs of objects extracted from the training set of the TOSD were randomly divided into 10 subsets.
2. Seven pairs of objects from each subset were used to train the MSR-based grounding with human feedback. Each pair represents one of the position-based spatial relations (*in, left, right, front, behind, above, below*).
3. Seven pairs of objects from each subset labeled with human feedback, and 200 pairs with relations labeled using the QSR-based grounding, were used to train the MSR-based grounding.
4. The control node chose between QSR-based grounding and the MSR-based grounding trained using the QSR-based grounding and human feedback.

The three schemes ($\#2, \#3, \#4$ above) were evaluated on 200 object pairs in test scenes of varying complexity. Table 1 indicates that the MSR-based grounding acquired using the QSR-based grounding makes better use of human feedback than that acquired using just human feedback, which supports **<u>H1</u>**. Note that the same amount of human feedback is provided with scheme $\#2$ and scheme $\#3$. The difference is that the latter scheme bootstraps off the generic knowledge encoded in the QSR-based grounding. These results indicate that using prior knowledge, an appropriate representation for knowledge, experience, and human feedback improves performance. Also, the control node-based combination of the two groundings provides better accuracy than just using the MSR-based grounding.

The second set of experiments was designed as follows, with the results summarized in Table 2:

	Accuracy of labels over test set of 200 object pairs		
Training sets	MSR (feedback)	MSR (QSR + feedback)	Combined model
Sets 1	65%	77%	84%
Sets 2	82%	80%	94%
Sets 3	68%	80%	85%
Sets 4	66%	83%	87%
Sets 5	65%	74%	82%
Sets 6	68%	77%	86%
Sets 7	64%	87%	90%
Sets 8	64%	84%	91%
Sets 9	62%	82%	87%
Sets 10	52%	72%	81%
Mean	65%	79%	87%
Std Dev	7.2%	4.6%	8.3%

Table 1: Comparison of (a) MSR grounding trained with just human feedback; (b) MSR grounding trained with 200 pairs labeled by the QSR grounding and seven pairs labeled with human feedback; and (c) the combination of MSR grounding, trained as in (b), and QSR-based grounding with the choice made by the control node.

1. Pairs of objects extracted from the training set of the TOSD were randomly divided into five subsets.
2. A MSR-based grounding was acquired using QSR-based labels for four out of the five subsets ($\approx$ 2000 pairs) in each run.
3. The use of the control node to chose between the MSR-based grounding (trained as above) and the QSR-based grounding, was also considered.

The two different schemes ($\#2, \#3$) were evaluated on a set of 200 object pairs in scenes of varying complexity—ground truth, once again, was obtained manually. Table 2 indicates that the control-node based combination of the groundings estimates spatial relations more accurately than using either grounding individually, which supports hypothesis **<u>H2</u>**.

Next, we acquired MSR-based groundings from different amounts of human feedback (with no QSR)—one, 15, and 25 training sets, each with seven object pairs in simulated scenes. These groundings were tested on 1400 object pairs from simulated (test) scenes of varying complexity. Table 3 shows that spatial relations are estimated accurately even when a small number of labeled samples are used to acquire the grounding.

Next, we conducted experiments similar to those for Table 1, but with a larger number of simulated scenes. The MSR-based grounding acquired using just human input had an accuracy of 95.9%, whereas the grounding obtained using human input and the QSR-based grounding had an accuracy of 97.2%. These results are similar to those with the TOSD.

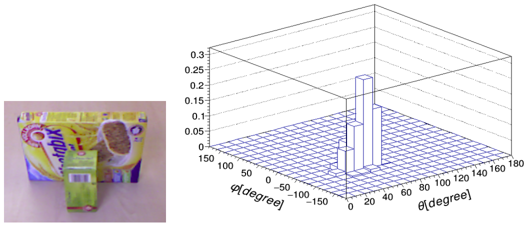
Further analysis indicates that most errors from the control node-based combination of the groundings correspond to truly ambiguous spatial relations, e.g., a scene in which object

	Accuracy of labels over test set of 200 object pairs		
Training sets	QSR only	MSR trained by QSR	Combined model
Sets 1+2+3+4	70%	62%	96%
Sets 1+2+3+5	70%	62%	96%
Sets 1+2+4+5	70%	60%	95%
Sets 1+3+4+5	70%	60%	96%
Sets 2+3+4+5	70%	60%	96%
Mean	70%	61%	96%
Std Dev	0	1.1%	0.5%

Table 2: Comparison of (a) QSR-based grounding; (b) MSR-based grounding from ≈ 2000 pairs labeled with QSR-based grounding (no human feedback); and (c) using the control node to combine MSR-based grounding, as trained in (b), and QSR-based grounding.

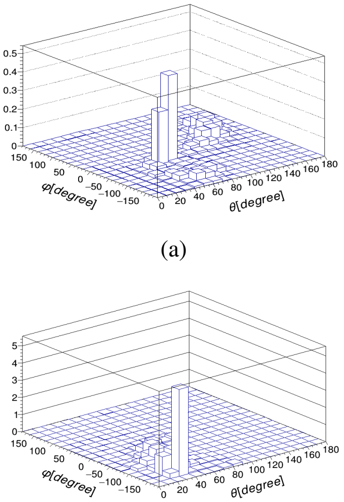
Model	Accuracy of labels over test set of 1400 object pairs
QSR	61.9%
MSR after 1 training set	96.1%
MSR after 15 training sets	98.5%
MSR after 25 training sets	98.6%

Table 3: QSR-based grounding compared with MSR-based groundings obtained using different amounts of human feedback.



(a) (b)

Figure 6: (a) Image from TOSD dataset; (b) Histogram generated from the image using the smaller box as the reference object.



(b)

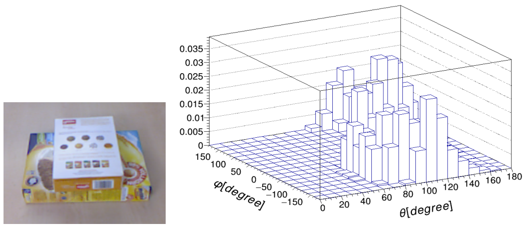
Figure 7: Histograms representing learned MSR groundings for: (a) $above$; and (b) $behind$.

A can be considered to be to the "left" or "behind" object B. Multiple labels are acceptable in such cases, and we just need to let the inference system allow multiple answers. In other cases, e.g., when each grounding is used individually, errors are due to the grounding being (or becoming) inaccurate—even in these cases, results do not depend on the order in which the training and test data are provided.

We also evaluated the ability to identify and correct errors. For the TOSD image in Figure 6a, the MSR-based grounding incorrectly stated that the larger box was *above* the smaller one. We compared the learned visual words for this label and correct label ("behind") with the histogram extracted from the object pair in the image. The χ^2 measure between the learned and observed visual words was 0.325 for *above* and 0.319 for *behind*. Even the QSR-based grounding detected 349 points in the *above* region and 23 in the *behind* region. The error was thus due to the incorrect input provided by the QSR-based grounding to the MSR-based grounding. We then visually compared the 2D histograms between the two objects—Figure 6b—with the MSR-based grounding for *above* and *behind*—Figure 7. The extracted histogram was more similar to the grounding for *above*—under standard viewpoints and orientations, $\theta > 90°$ for $above$, but many points corresponded to $\theta \approx 60°$ in this case. To correct this error, we processed an image that actually contained an instance of the $above$ relation—Figure 8a. The θ values in the revised histogram for $above$ were mostly $\in [90°, 120°]$—Figure 8b. The MSR-based grounding then provided the correct spatial relation between the objects in Figure 6a—the χ^2 similarity scores were 0.319 for *behind* and 0.088 for *above*.

5 Conclusions

To truly assist humans in complex domains, robots need the ability to recognize, reason about, and provide understandable descriptions of spatial relations between objects. Our architecture uses Answer Set Prolog to represent and reason with incomplete domain knowledge, which includes spatial relations computed using a generic qualitative grounding of these relations (QSR). These inferred relations and human input (when available) are used to incrementally acquire a more specialized quantitative grounding of spatial relations (MSR). Also, a relative measure of confidence in the two groundings is computed to enable the use of the more reliable grounding for inferring spatial relations in the subsequent scenes.



(a) (b)

Figure 8: (a) Image with one object *above* another; and (b) revised 2D histogram for *above*.

Experimental evaluation demonstrates the ability to reliably estimate spatial relations in a benchmark dataset of complex tabletop images and simulated scenes of furniture, even with a small number of labeled training samples. Future work will consider more drastic changes in factors such as viewpoint and scale, and explore the acquisition of action models that include the learned spatial relations. Furthermore, we will include modules for scene understanding and explore the interplay between reasoning and learning on a mobile robot collaborating with humans in complex indoor domains.

Acknowledgements

This work was supported in part by the Asian Office of Aerospace Research and Development award FA2386-16-1-4071, and the US Office of Naval Research Science of Autonomy award N00014-17-1-2434. All opinions and conclusions described in this paper are those of the authors.

References

[Balai *et al.*, 2013] Evgenii Balai, Michael Gelfond, and Yuanlin Zhang. Towards Answer Set Programming with Sorts. In *International Conference on Logic Programming and Nonmonotonic Reasoning*, September 2013.

[Belz *et al.*, 2015] Anja Belz, Adrian Muscat, Maxime Aberton, and Sami Benjelloun. Describing Spatial Relationships between Objects in Images in English and French. In *Workshop on Vision and Language*, pages 104–113, 2015.

[Calli *et al.*, 2015] Berk Calli, Aaron Wallsman, Arjun Singfh, and Siddhartha S. Srinivasa. Benchmarking in Manipulation Research. *IEEE Robotics and Automation Magazine*, (September):36–52, 2015.

[Elliott and Vries, 2015] Desmond Elliott and Arjen P De Vries. Describing Images using Inferred Visual Dependency Representations. In *Annual Meeting of the Association for Computational Linguistics*, pages 42–52, 2015.

[Fichtl *et al.*, 2015] Severin Fichtl, Dirk Kraft, Norbert Krüger, and Frank Guerin. Using Relational Histogram Features and Action Labelled Data to Learn Preconditions for Means-End Actions. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (Workshop on Sensorimotor Contingencies for Robotics)*, 2015.

[Gelfond and Kahl, 2014] Michael Gelfond and Yulia Kahl. *Knowledge Representation, Reasoning and the Design of Intelligent Agents*. Cambridge University Press, 2014.

[Guadarrama *et al.*, 2013] Sergio Guadarrama, Lorenzo Riano, Dave Golland, Daniel Gouhring, Yangqing Jia, Dan Klein, Pieter Abbeel, and Trevor Darrell. Grounding Spatial Relations for Human-robot Interaction. In *International Conference on Intelligent Robots and Systems*, pages 1640–1647, 2013.

[Mees *et al.*, 2017] Oier Mees, Nichola Abdo, Mladen Mazuran, and Wolfram Burgard. Metric Learning for Generalizing Spatial Relations to New Objects. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3175–3182, Vancouver, Canada, September 24-28, 2017.

[Paul *et al.*, 2016] Rohan Paul, Jacob Arkin, Nicholas Roy, and Thomas Howard. Efficient Grounding of Abstract Spatial Concepts for Natural Language Interaction with Robot Manipulators. In *Robotics: Science and Systems*, Ann Arbor, USA, June 18-22, 2016.

[Pronobis and Rao, 2017] Andrzej Pronobis and Rajesh Rao. Learning Deep Generative Spatial Models for Mobile Robots. In *RSS Workshop on Spatial-Semantic Representations in Robotics*, Cambridge, USA, July 16, 2017.

[Rusu, 2010] Radu Bogdan Rusu. Semantic 3D Object Maps for Everyday Manipulation in Human Living Environments. *KI - Kunstliche Intelligenz*, 24(4):345–348, 2010.

[Shridhar and Hsu, 2017] Mohit Shridhar and David Hsu. Grounding Spatio-Semantic Referring Expressions for Human-Robot Interaction. In *RSS Workshop on Spatial-Semantic Representations in Robotics*, July 2017.

[Sridharan and Meadows, 2017] Mohan Sridharan and Ben Meadows. A Combined Architecture for Discovering Affordances, Causal Laws, and Executability Conditions. In *International Conference on Advances in Cognitive Systems*, May 2017.

[Thippur *et al.*, 2015] Akshaya Thippur, Chris Burbridge, Lars Kunze, Marina Alberti, John Folkesson, Patric Jensfelt, and Nick Hawes. A Comparison of Qualitative and Metric Spatial Relation Models for Scene Understanding. In *AAAI Conference*, pages 1632–1640, 2015.

[Ye and Hua, 2013] Jun Ye and Kien A Hua. Exploiting Depth Camera for 3D Spatial Relationship Interpretation. In *ACM Conference on Multimedia Systems*, pages 151–161, 2013.

[Zampogiannis *et al.*, 2015] Konstantinos Zampogiannis, Yezhou Yang, Cornelia Ferm, and Yiannis Aloimonos. Learning the Spatial Semantics of Manipulation Actions through Preposition Grounding. In *International Conference on Robotics and Automation*, pages 1389–1396, May 2015.

[Ziaeetabar *et al.*, 2017] Fatemeh Ziaeetabar, Eren Erdal Aksoy, Florentin Wörgötter, and Minija Tamosiunaite. Semantic Analysis of Manipulation Actions Using Spatial Relations. In *International Conference on Robotics and Automation*, pages 4612–4619, 2017.