2013 13th IEEE-RAS International Conference on
Humanoid Robots (Humanoids).
October 15 - 17, 2013. Atlanta, GA

Grounded Spatial Symbols for Task Planning Based on Experience

Kai Welke[1], Peter Kaiser[1], Alexey Kozlov[1], Nils Adermann[1], Tamim Asfour[1],
Mike Lewis[2], Mark Steedman[2]
[1]Institute for Anthropomatics, Karlsruhe Institute of Technology
Adenauerring 2, 76131 Karlsruhe, Germany
[2]School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh, EH8 9AB, Scotland, United Kingdom
{welke,asfour}@kit.edu, {steedman}@inf.ed.ac.uk

***Abstract*— Providing autonomous humanoid robots with the abilities to react in an adaptive and intelligent manner involves low level control and sensing as well as high level reasoning. However, the integration of both levels still remains challenging due to the representational gap between the continuous state space on the sensorimotor level and the discrete symbolic entities used in high level reasoning. In this work, we approach the problem of learning a representation of the space which is applicable on both levels. This representation is grounded on the sensorimotor level by means of exploration and on the language level by making use of common sense knowledge. We demonstrate how spatial knowledge can be extracted from these two sources of experience. Combining the resulting knowledge in a systematic way yields a solution to the grounding problem which has the potential to substantially decrease the learning effort.**

I. INTRODUCTION AND RELATED WORK

Establishing robotic systems that offer a level of autonomy suitable for real world applications requires bringing together expertise and approaches from a variety of different research fields. One of the most challenging problems therewith consists in integrating high level artificial intelligence (AI) with low level robot control.

The main challenge arises form the representational discontinuity between the continuous state spaces of robot control and the discrete symbolic representation used in most AI approaches. In order to fill-in this representational gap, the concept of object-action complexes (OACs) has been proposed as representation for all levels of the processing hierarchy ([1]). The OAC follows the affordance concept and tightly couples perception and action within a single representation. Applications of the OAC concept on several levels of the hierarchy including high level planning have been demonstrated ([2], [3]).

Bridging the gap between low level and high level processing requires defining a path from the continuous world to the symbolic representation and vice versa. While the OAC formalization takes into account the major processes within a cognitive architecture such as learning, predicting, and execution, the underlying structure in terms of appropriate state spaces needs to be defined in a problem specific way.

In this work, we focus on deriving representations of the spatial domain enabling the connection of high level planning with the sensori-motor level in humanoid robots. The need of such a representation can be visualized by considering the following action defined within a PDDL [4] domain specification:

```
(:action putdown
 :parameters ( ?x ?y ?z )
 :precondition (and (inHand ?x ?z)
                    (hand ?z)
                    (location ?y)
                    (graspable ?x))
 :effect (and (handEmpty ?z )
              (at ?x ?y)
              (not(inHand ?x ?z))))
```

The action `putdown` describes the process of putting an object `?x` held in the hand `?z` to a location `?y`. Two properties of this action render it a good example for the proposed work: First, the putdown action is required in several assistance tasks such as setting the table or stowing away. Second, the action has a direct reference to the spatial domain by means of the location `?y`. The spatial parameter `?y` appears in the binary predicate `at ?x ?y` which is necessary to describe the effect on the world state.

The goal now consists in establishing a representation for the parameter `?y` which is valid on the semantic as well as on the sensori-motor level. On the bottom-up path, this representation needs to support the observation of the world change triggered by the action in terms of the predicate `at ?x ?y`. On the top-down path, the execution of the action needs to be parameterized with the appropriate spatial location from the continuous domain in order to achieve this world change.

The simultaneous task and motion planning (STAMP) field tackles the problem of combining task planning and metric level. The goal consists in combining task and collision-free motion planning in a consistent fashion ([5],[6],[7],[8]). In contrast to STAMP, where full knowledge of the metrics and geometry as well as full knowledge of the task planning domain is assumed, our research focus lies on exploiting experience to improve the learning process of such representations.

Semantic information in spatial representations has been exploited in semantic maps of the environment in the navigation and mapping field [9]. Such semantic maps usually describe topological relations between semantic places in

978-1-4799-2618-3/13/$31.00 ©2013 IEEE

the environment. Either these places are directly perceivable using appropriate models or detectors (e.g. [10]) or inferred using a known model of the binding from semantics to detectable places ([11], [12], [13], [14]).

In contrast to these approaches, the idea of exploiting experience in order to learn these semantic bindings of places stands at the core of our approach. Without the use of detectors for places, we make use of available common sense knowledge in order to establish the binding between semantics and the explored metrical representation. The extracted common sense knowledge is transferable from one environment to another and thus provides a consistent binding to the symbolic world. This transferability is essential for establishing task planning across large domains.

The extraction of spatial relations from natural language has been studied in application to understanding commands or directions to robots given in natural language (e.g. [15], [16], [17]). In contrast to approaches based on annotated corpora of command-executions or route instructions, or using knowledge bases like *Open Mind Common Sense* [18] explicitly created for artificial intelligence applications, we extract the relevant relations from large amounts of text written by humans for humans. The text mining techniques used in [19] and [20] to extract action-tool relations to disambiguate visual interpretations of kitchen actions are related.

II. THE SYSTEM CONCEPT

A. Conceptual assumptions

The focus of this work lies in the acquisition of grounded spatial representations from experience. Obviously this spatial knowledge is only a small fraction of the overall knowledge required to operate the system in an autonomous way. In order to clearly outline our approach, we assume prior knowledge to be present in several forms on the system. The prior knowledge assumed in this work includes

- Object knowledge: We assume extensive prior knowledge on objects in the world. This knowledge includes object models for recognition and localization as well as the associated class labels. More precisely, we know the models and class labels for common manipulable kitchen objects such as cups, plates, milk, or juice available from the KIT object model database [21].
- Action knowledge: We assume that the robotic platform is able to perform basic actions. We explicitly make use of the grasp and putdown actions during the exploration phase. Further, locomotion abilities are necessary in order to allow learning of larger scale spatial domains.

In order to apply the gathered knowledge in task planning, it is necessary to have full knowledge of the planning domain. In order to execute the plan, a sensori-motor representation of all involved actions and all predicates needs to be available. All non-locational constants, such as class labels of objects, need to be grounded on the sensori-motor level. The rules in terms of pre- and postconditions of actions have to be known, but can also be learned by exploration [22]. The missing piece, the combined sensori-motor representation of locations and the associated symbolic constant is learned by our approach.

B. System architecture

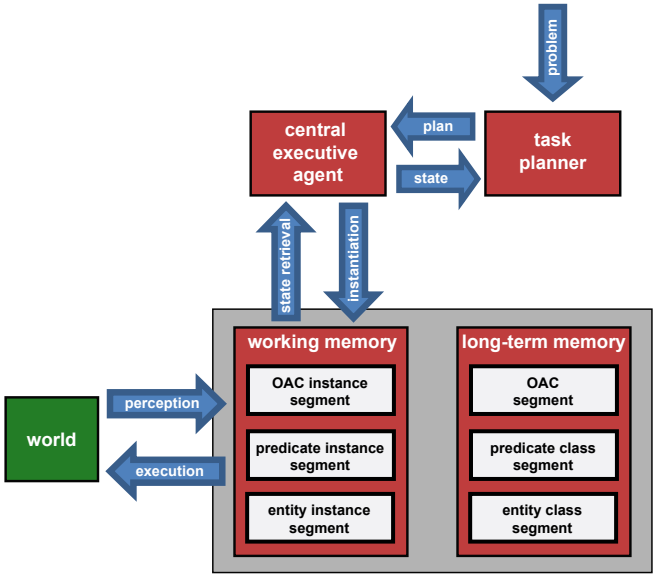


Fig. 1: The integration between the task planning level and the sensori-motor level is established by the central executive agent (CEA). For each plan element the CEA instantiates the appropriate OAC and associated predicates and entities. Entities correspond to constants on the planning level. The goal of this work consists in learning an entity of type `location` from experience.

While addressing a quite specific problem of symbol grounding for task planning in this work, we do not consider this problem isolated but within the context of a systematic way of coupling high level task planning with the sensori-motor level on humanoid robots. For this purpose, we developed and implemented an architecture that couples these levels. Relevant components within this architecture are illustrated in Fig. 1, where the focus lies on how sensori-motor representations are made available for planned task execution and plan monitoring. The major involved components are the task planner, the central executive agent (CEA) and the memory system. In our current implementation we make use of the PKS planner for STRIPS like task planning with the support of plan monitoring ([23], [24]). The CEA is the mediator between planning level and memory system. Its purpose is bidirectional: Bottom-up it translates the content of the working memory (WM) to a world state in PDDL format in order to enable planning and plan monitoring. Top-down the CEA translates plans from the PDDL domain to OAC representations. In a sequential manner it instantiates OACs from the long-term memory (LTM) in the WM according to the active plan element. Further, in order to allow state monitoring, it creates instances of required WM elements associated with the OAC. Such elements include

unary, binary, or n-nary predicates as required for the PDDL world state description and entities which correspond to constants on the planning level such as objects and locations. The underlying execution and perception mechanisms take care of keeping these constants and predicates consistent with the real world.

Taking into account this architecture the goal now consists in learning the representation of an entity class for locations which can be used as constant of type `location ?y` and can be used with the predicate `at ?x ?y`. While the class stored in LTM should be valid for different tasks and objects, the instantiation of this entity in the WM is specific for the current task and object.

III. ACQUIRING SPATIAL KNOWLEDGE FROM EXPERIENCE

The acquisition of spatial knowledge from experience has several advantages over resorting to manually generated spatial representations. The most important benefit lies in the improvement of the system's autonomy by establishing the required processes for acquiring such representations on the robot.

In this work, we exploit two different sources of experience: experience gathered through exploration on the robot system and experience available in common sense knowledge. The exploration on the robot yields embodied sensory-level representations that already encode the constraints of the platform such as the visibility of objects. Spatial information from common sense knowledge on the other hand is extracted from large text corpora and thus provides knowledge on the symbolic level. In the following, we introduce approaches that make available both sources of knowledge for the acquisition of spatial knowledge. We will show how to combine the gathered knowledge in order to acquire grounded spatial symbols in Section IV.

A. Spatial knowledge from exploration

The goal of the exploration consists in incrementally learning a spatial model of the environment with respect to the set of known objects. More precisely, the developed approach allows inferring common object locations based on object detection and localization results from multiple episodes. In order to keep the exploration effort low, the robot performs self-observation: While the robot is controlled through human interaction in our kitchen scenario, it records all encountered and manipulated objects with location, label, and current task.

1) Metric spatial representation: In order to represent the encountered objects, we employ probabilistic and continuous space representations, which are similar to those proposed by Stulp et al. in their concept of ARPlace [25]. More precisely, the object positions are described by a probability density function (PDF) in 3D space. This approach allows avoiding a prior space discretization, while simultaneously providing a natural way to incorporate object localization uncertainty. To represent the spatial distribution of an object class c, we use the Gaussian Mixture Model (GMM):

$$f_c(\vec{x}) = \sum_{i=1}^{N} w_i \mathcal{N}(\vec{x}; \vec{\mu_i}, \vec{\Sigma_i}), \qquad \forall w_i > 0 \tag{1}$$

The GMM has an important property of being a universal PDF approximator [26], which means that it can approximate any given distribution with an arbitrary precision. From a practical point-of-view, GMM is of particular interest because of its ability to cope with multi-modality and moderate storage requirements.

2) Learning common locations: Each time an object is recognized in the world, the spatial representation is updated. Initially, we start with an empty GMM η for each object class. The object position is modeled as a Gaussian $\mathcal{N}(\vec{x}; \vec{\mu_o}, \Sigma_o)$ encoding the localization uncertainty in 3D Cartesian space [27]. This Gaussian is added as a new component with a constant weight (e.g., 1) to the GMM: $\eta \leftarrow \eta \cup (1, \vec{\mu_o}, \Sigma_o)$ the corresponds to the object class. At the same time, the following three operations are applied to the existing components:

- **Aging** Since older observations are assumed to be less relevant than the recent ones, the weights of corresponding GMM components are reduced by multiplying with the discount coefficient $\gamma \in [0, 1]$:

$$\forall i \;\; w_i \leftarrow \gamma \cdot w_i \tag{2}$$

- **Pruning** Components with weights below the threshold W_{prune} are removed from the mixture:

$$\forall i: \;\; w_i < W_{prune} \quad \eta \leftarrow \eta \;\setminus\; (w_i, \vec{\mu_i}, \Sigma_i) \tag{3}$$

- **Merging** two components which are considered "similar" in terms of their Mahalanobis distance d are replaced with their *moment-preserving merge*:

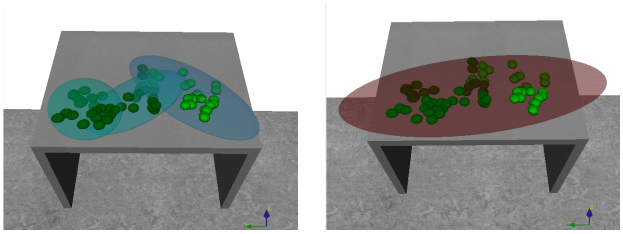
$$d(i,j) < D_{min}: \;\; \eta \leftarrow \eta \;\setminus\; (w_i, \vec{\mu_i}, \Sigma_i), (w_j, \vec{\mu_j}, \Sigma_j)$$
$$\eta \leftarrow \eta \;\cup\; (w_m, \vec{\mu_m}, \Sigma_m) \tag{4}$$

The calculation of $(w_m, \vec{\mu_m}, \Sigma_m)$ is performed according to [28].

The resulting representation encodes the spatial distribution of common object locations. The aging factor accounts for changes in the scene, while pruning and merging keep the representation compact. An example for common locations on the table is illustrated in Fig. 2.

3) Querying: To make spatial knowledge accessible, it should be provided at a suitable abstraction level for the task at hand. For this purpose, we implemented a query interface which allows for two types of generalization:

- **Spatial generalization** Spatial generalization allows combining several neighbored observation to a single cluster. For this purpose, three established GMM reduction algorithms were implemented (West [29], Runnalls [30], Williams [28]). The level of generalization can be adjusted by specifying the stop condition of the GMM reduction. The stop condition is either defined by a target number of clusters or by the maximum



(a) $M = 150$ (b) M = 400

Fig. 2: Different clusterings of the same position distribution achieved by setting *(a)* low and *(b)* high value of the deviation threshold M.

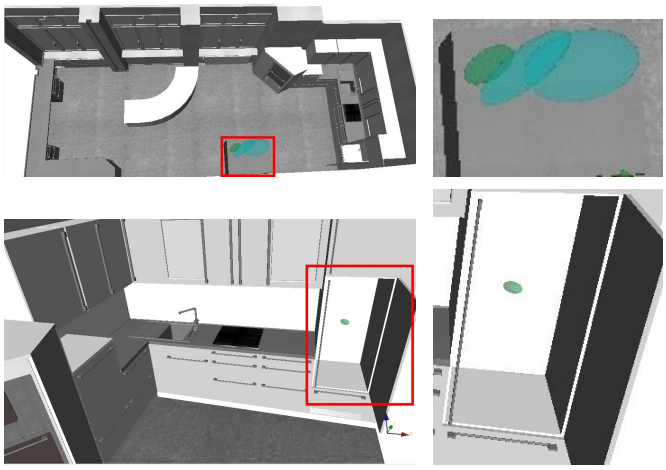


Fig. 3: Common places learned from four ARMAR-III kitchen demonstrations after spatial generalization.

deviation M within a cluster. An example of the spatial generalization is illustrated in Fig. 2.

- **Ontological generalization** In addition to positions of objects (e. g. Cup), places for abstract classes (e.g. Food) can be queried as well. This is achieved by using a simple class ontology with parent-child relations.

4) Common places in the kitchen domain: The learning algorithm and clustering approach described above were applied on object locations collected during the demonstrations of the humanoid robot ARMAR-III ([31], [32]) in the kitchen. In the scenario, the robot localized and manipulated objects in the fridge and on the table. Figure 3 illustrates a spatial generalization query on the representation resulting from four ARMAR-III demonstrations.

B. Spatial relations from human knowledge

Besides learning from the robot's own experience, we would like to gain information on spatial relations from human knowledge. Human knowledge could tell the robot that *milk* is usually kept in refrigerators. Hence, there is a certain probability that a spatial cluster containing positions of *milk* is a refrigerator. In this section we propose a method to infer a set of likely locations for a given object.

1) Extracting spatial relations from text: Spatial relations are linguistically expressed using spatial prepositions:

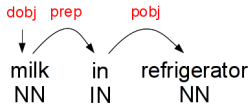


Fig. 4: A syntactic ngram containing two content-words and a preposition. The words are equipped with a part-of-speech tag and a dependency label[1].

- *The milk is **in** the refrigerator*
- *Take a knife **from** the drawer.*

In this work, we propose to determine the conditional probability of a location given an object based on the number of matching prepositional relations in a text corpus. We are aiming for prepositions like *in* and *on*, but do not predefine a set of valid prepositions.

Let N_{obj} be the frequency of occurrence of obj in prepositional contexts and let $N_{obj,loc}$ be the number of those prepositional contexts where obj and loc occurr together. The conditional probability $P(loc|obj)$ can then be approximated as follows:

$$P(loc|obj) = \frac{P(obj, loc)}{P(obj)} \approx \frac{N_{obj,loc}}{N_{obj}} \tag{5}$$

Working on the whole vocabulary of the corpus makes the values of the conditional probability difficult to compare. As we know the set L of possible locations in the kitchen from the planning domain specification, we can formulate the restricted conditional probability:

$$P_L(loc|obj) = \frac{P_L(obj, loc)}{P_L(obj)} \approx \frac{N_{obj,loc}}{\sum_{l \in L} N_{obj,l}} \tag{6}$$

These formulas imply the assumption that a text corpus is a suitable foundation for estimating $P(loc|obj)$. See section III-B.4 for a discussion.

2) The Text Corpus: In this paper we propose to extract spatial relations from the *Google Books Ngrams Corpus* [33], in the following referred to as the *Google Corpus*. This corpus contains a representation of 3.5 million English books with a total size of about 345 billion words. It does not contain the raw text. Several preprocessing steps have been applied to the sentences:

1) Parsing into dependency trees
2) Extracting *syntactic ngrams*, i.e. n content-words long subpaths of the dependency trees (see Fig. 4)
3) Counting the frequency of occurrence of each syntactic ngram

We are using the corpus in its *arcs*-variant, which only includes syntactic ngrams with two content-words plus possible non-content-words like prepositions or conjunctions. Overall, the preprocessing makes the Google Corpus convenient for conducting analysis on the frequency of grammatical structures.

In the Google Corpus, each syntactic ngram is stored in a distinct line. The information that is relevant in this paper is

[1] *NN* - noun, *IN* - preposition
dobj - direct object, *prep* - preposition, *pobj* - prepositional object

TABLE I: Restricted conditional probability $P_L(loc|obj)$. Darker colors indicate higher probabilities, omitted probabilities are zero.

	cellar	counter	cupboard	dishwasher	drawer	freezer	microwave	oven	refrigerator/fridge	shelf	table
beer	0.0763	0.0518	0.0095						0.6045		0.2579
bread	0.0033	0.0235	0.0515		0.0017	0.0065		0.2566	0.0046	0.0181	0.6343
cereal			0.4045							0.1685	0.4270
coffee	0.0076	0.1458	0.0108				0.0120	0.0089		0.0203	0.7945
cup		0.0728	0.0337	0.0029	0.0066	0.0025	0.0172	0.0278	0.0059	0.0405	0.7901
dough		0.1293				0.0376		0.2674	0.3834		0.1823
juice	0.0190	0.0306				0.0146	0.0146	0.0889	0.6064		0.2259
knife		0.0723			0.2752			0.0089	0.0036	0.0197	0.6203
meat	0.0194	0.0180	0.0143			0.0680	0.0132	0.1180	0.1309	0.0028	0.6154
milk	0.0275	0.0299	0.0141			0.0255	0.0319	0.1054	0.3832	0.0238	0.3586
pot	0.0103	0.1154	0.0291					0.1195	0.0139	0.0734	0.6385
wine	0.4128	0.0061	0.0144					0.0112	0.0387	0.0050	0.5119

the ngram itself and its frequency of occurrence. The entry for the exemplary ngram in Fig. 4 looks as follows:

milk/NN/dobj/0 in/IN/prep/1 refrigerator/NN/pobj/2 160

The syntactic ngram's path consists of three nodes, each containing the following relevant fields:

- The word that the node represents in the original sentence (e.g. *milk*).
- The Penn-Treebank part-of-speach tag [34] for this word (e.g. *NN*).
- The basic Stanford-dependencies label [35] for the node's grammatical function (e.g. *dobj*).

The final number is the frequency of occurrence of the syntactic ngram. This exemplary ngram occurred 160 times in the Google Corpus.

3) Extracting Relations from the Corpus: We are interested in extracting prepositional relations between objects and locations. Referring to the above exemplary line from the Google Corpus, we are looking for lines that match the following pattern:[2]

$$\textit{[object]}/\text{NN}/\bullet/\bullet\ \bullet/\bullet/\text{prep}/\bullet\ \textit{[location]}/\text{NN}/\bullet/\bullet. \quad (7)$$

After searching and accumulating prepositional contexts with regard to pattern (7), the probability of a location given an object can be approximated using (6).

4) Evaluation: Table I shows the restricted conditional probabilities $P_L(loc|obj)$ as defined in (6) of a set of objects (y-axis) given a predefined set of possible locations (x-axis). The table shows that the proposed method of extracting prepositional contexts from a text corpus is able to infer reasonable values for $P_L(loc|obj)$. Exemplary conclusions that can be drawn from the results include:

- Refrigerators are a likely location for beer, juice and milk.
- Cups and coffee may be found on tables.
- Apart from the oven, bread could be on a table or in a cupboard.

IV. GROUNDED SPATIAL SYMBOLS

In this section we will outline how symbols for locations can be obtained which are grounded in language as well as in the continuous domain. For this purpose, the two sources of experience introduced in the previous section, exploration and common sense knowledge, are combined. In the following we will show how the predicate `at ?x fridge` can be inferred from experience. Since we assumed to have a representation of all involved objects on the sensory as well as on the symbolic level, we know all enumerations of the parameters `?x` and the associated object models. The problem of evaluating the above predicate then boils down to inferring a grounded representation of the location constant `fridge`. Note that the constant `fridge` does not refer to an object but to the support locations offered by the fridge in terms of the `at` predicate as defined in our PDDL definition of the putdown action in section I.

The greedy acquisition of a representation of the fridge locational constant grounded in the spatial domain would require providing a large set of spatial locations corresponding to the fridge together with the symbolic tag. Collecting this data on the robot either requires the involvement of a teacher, or the evaluation of the symbolic binding via higher level inference (e.g. by exploiting knowledge such as temperature in the fridge) in order to assert the validity of the symbolic tag. Such grounding processes are slow and costly in terms of resources. By exploiting experience, a good prior for such constants can be achieved, substantially decreasing the grounding effort.

The complete chain of acquiring a grounded representation of `fridge` from experience is illustrated in Fig. 5. The only prior knowledge at the start of the process consists in a common reference frame for spatial locations that defines the space of all possible locations. Through self-observation and by applying common sense knowledge, we calculate a prior for fridge support locations. The final grounding step again involves means of ascertaining the gained representation similar to the greedy approach but with a prior that eases the grounding process. In the following, the single steps towards this prior are discussed in detail.

A. Exploration of grounded support locations

The first step in our approach consists in determining grounded support locations in the environment by exploration. This exploration is realized by means of self-observation as introduced in Section III-A. The resulting

[2] "•" denotes a wildcard

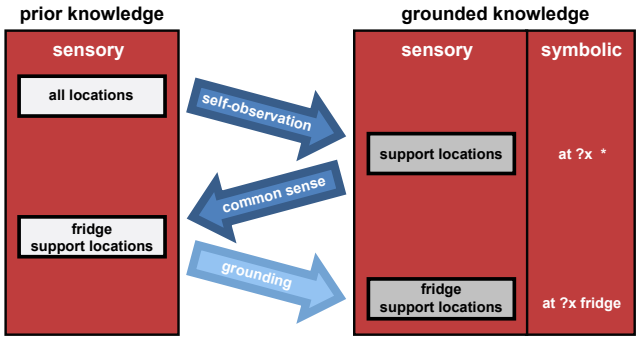


Fig. 5: The proposed approach for acquiring grounded spatial symbols combines two sources of experience: Self-observation of the robot during the execution of kitchen tasks yields grounded representations of support locations. These support locations are associated with semantic symbols by exploiting common sense knowledge.

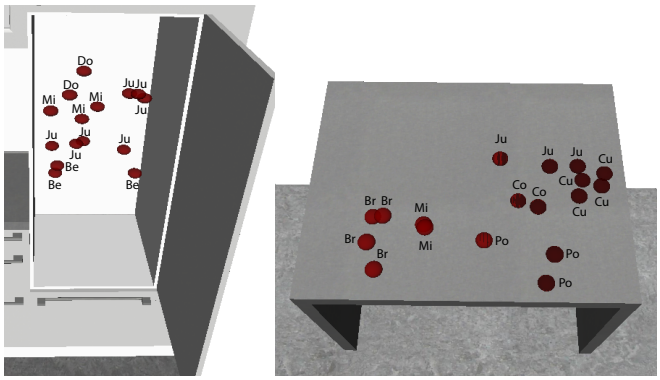


Fig. 6: Simulation of the exploration phase with the following objects: **be**er, **br**ead, **co**ffee, **cu**p, **do**ugh, **ju**ice, **mi**lk, **po**t. The pickup and putdown poses were chosen at random in the fridge and on the table.

locations collected during the execution of tasks on the robot are stored using the proposed metric spatial representation. Thereby, each experienced location is accompanied with the label of the occupying object and the action that has been executed on the object. In order to extract support locations from this data, only these locations are considered which correspond to the actions pickup or putdown. Since the support locations are collected during self-observation, the resulting representation naturally includes reachability constraints of the experimental platform. With this representation, we acquired a grounded concept of support locations for the covered objects `?x` denoted with `at ?x *` in Fig. 5. An example of such a representation based on simulated data is illustrated in Fig. 6.

B. Common sense knowledge for symbol binding

In the second step of our approach, we employ common sense knowledge in order to establish a symbolic binding for support locations within the representation explored in the previous step. In our example, the goal consists in inferring a good prior for `at ?x fridge` based on the explored `at ?x *`. The major challenge here consists in establishing a representation of the location constant `fridge` that is grounded in the sensory as well as in the language domain. Here we exploit that spatial relations between objects and locations are part of human common sense knowledge and are accessible on the linguistic level in terms of prepositional structures as detailed in Section III-B. For each location stored in the representation from step one we can conclude the probability of belonging to a fridge location by accessing the associated object class label and querying the common sense knowledge encoded in Table I. By assuming that locational symbols stay constant in the local neighborhood (e.g. neighbored locations of a fridge location are also likely to be fridge locations), evidence for the symbol can be propagated from location to location. This can be efficiently implemented by making use of the spatial generalization query introduced in Section III-A.3. For each resulting soft cluster, the associated likelihood of belonging to the location `fridge` is collected over all object locations belonging to the cluster by means of a linear opinion pool. Figure 7 illustrates the result of this process for the simulated data from Fig. 6. Using an appropriate deviation threshold for spatial generalization yields two clusters: one in the fridge and one on the table. The probability of the fridge cluster belonging to the constant `fridge` is calculated with the above procedure and amounts to $P(Fridge|O1) = 0.53$. The same approach for the table cluster yields $P(Fridge|O2) = 0.15$. The application of common sense knowledge in this way allows exploiting negative examples. For instance, clusters that contain a high frequency of cups are not likely to be located in the fridge according to Table I.

Associating the explored objects with location symbols by exploiting common sense knowledge yields the representation we were seeking: a representation of the locational constant `fridge` which is grounded within the metric spatial representation. This representation can be used to establish the world state in terms of the predicate `at ?x fridge`. For objects `?x` that lie in the fridge cluster this predicate is valid. Further, the explored locations associated with the cluster can be used to parameterize the action `putdown ?x fridge ?z`. This process of course involves thresholding the probability values in Fig. 7 which might not be trivial in all cases. Nevertheless, we achieved a good prior for the symbolic binding which allows quite efficient disambiguation in further grounding processes.

V. DISCUSSION

A. Contributions

In this work we presented an approach for learning a representation of space applicable on the task planning as well as on the sensori-motor level. In order to establish a symbolic link to the continuous world, we exploited two sources of experience: experience from exploration and experience from common sense knowledge.

While this work developed the approach in a quite exemplary manner, based on the `fridge` example, the general concept is applicable to most typical places in human made

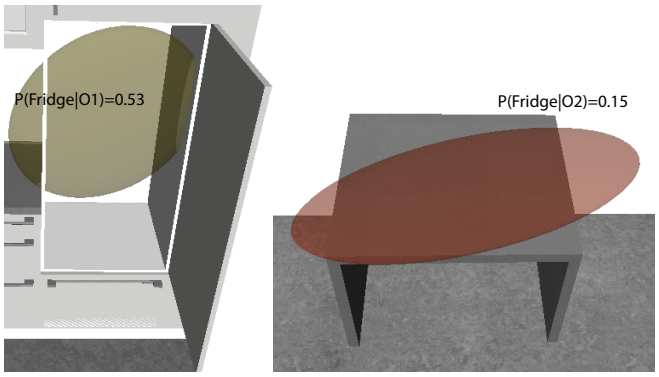


Fig. 7: Result of spatial generalization with deviation threshold $M = 400$. The cluster in the fridge $O1$ has a much higher probability of being the fridge as the cluster $O2$ on the table. This result is achieved only based on the observation of objects and the extraction of prepositional contexts.

TABLE II: Most frequent prepositions for locations

	first	second
cellar	in - 69%	from - 27%
counter	on - 81%	at - 9%
cupboard	in - 71%	from - 29%
dishwasher	in - 100%	
drawer	in - 86%	from - 14%
freezer	in - 96%	from - 4%
microwave	in - 100%	
oven	in - 81%	from - 9%
refrigerator/fridge	in - 80%	from - 20%
shelf	on - 95%	from - 5%
table	on - 73%	at - 9%

environments. The common sense knowledge provides transferable concepts for places. Considering the `table` location in Table I implies that the table is a quite versatile support surface. Based on this knowledge we would assign the symbol `table` to places which are used to support numerous different objects, independent of the current domain.

The application of common sense knowledge was demonstrated on simulated data in order to show the feasibility of the approach. The same could be done on the real data collected through exploration. However, the real kitchen data only covered the object juice in the fridge, which would result in a quite simple query to the common sense knowledge (e.g. $P(Fridge|O1) = 0.6$). Further, not all objects on the table had a significant amount of occurrences in the corpus. This stems from the fact, that often classes are used in language instead of single instances of the object (e.g. cereal vs. vitalis cereal). This problem could be addressed by means of the ontological generalization as explained in Section III-A.

B. Outlook

So far, we used the predicate `at` in order to express that an object is at a specific location. However, this preposition is not quite common and would probably not be applied in order to describe a location in the fridge. Rather, we would use prepositions which also encode the function of the location such as `in`, `on`, or `from`. In the proposed approach we used all prepositions to query for locations in order to get a significant amount of occurrences of a location independent of its function. In task planning however, there can be a huge difference between putting something on a place or in a place (e.g. open the door before putting in). This problem can be addressed by making use of the corpus again. As can be seen in Table II the query yields the correct prepositions for fridge and table and thus also allow to infer this functional aspect from common sense knowledge.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement Nº 270273 (Xperience).

REFERENCES

[1] N. Krüger, C. Geib, J. Piater, R. Petrick, M. Steedman, F. Wörgötter, A. Ude, T. Asfour, D. Kraft, D. Omrčen, A. Agostini, and R. Dillmann, "Object-action complexes: Grounded abstractions of sensorimotor processes," *Robotics and Autonomous Systems*, vol. 59, pp. 740–757, 2011.

[2] C. Geib, K. Mourao, R. Petrick, N. Pugeault, M. Steedman, N. Krueger, and F. Wörrgötter, "Object action complexes as an interface for planning and robot control," in *IEEE International Conference on Humanoid Robots (Humanoids)*, 2006.

[3] R. Petrick, D. Kraft, K. Mourão, N. Pugeault, N. Krüger, and M. Steedman, "Representation and Integration: Combining Robot Control, High-Level Planning, and Action Learning," in *International Cognitive Robotics Workshop (CogRob 2008) at ECAI 2008*, 2008, pp. 32–41.

[4] D. McDermott, M. Ghallab, A. Howe, C. Knoblock, A. Ram, M. Veloso, D. Weld, and D. Wilkins, "Pddl—the planning domain definition language," New Haven, CT: Yale Center for Computational Vision and Control, Tech. Rep., 1998.

[5] S. Cambon, R. Alami, and F. Gravot, "A hybrid approach to intricate motion, manipulation and task planning," *The International Journal of Robotics Research*, vol. 28, no. 1, pp. 104–126, 2009.

[6] J. Wolfe, B. Marthi, and S. Russell, "Combined task and motion planning for mobile manipulation," in *International Conference on Automated Planning and Scheduling*, Toronto, Canada, 05/2010 2010.

[7] L. Kaelbling and T. Lozano-Perez, "Hierarchical task and motion planning in the now," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2011, pp. 1470–1477.

[8] E. Plaku, "Planning in discrete and continuous spaces: From ltl tasks to robot motions," in *Advances in Autonomous Robotics*, ser. Lecture Notes in Computer Science, G. Herrmann, M. Studley, M. Pearson, A. Conn, C. Melhuish, M. Witkowski, J.-H. Kim, and P. Vadakkepat, Eds. Springer Berlin Heidelberg, 2012, vol. 7429, pp. 331–342.

[9] B. Kuipers, "The spatial semantic hierarchy," *Artificial Intelligence*, vol. 119, pp. 191–233, 2000.

[10] A. Pronobis, O. M. Mozos, B. Caputo, and P. Jensfelt, "Multi-modal semantic place classification," *The International Journal of Robotics Research, Special Issue on Robotic Vision*, vol. 29, no. 2-3, pp. 298–320, Feb. 2010.

[11] A. Nüchter, H. Surmann, K. Lingemann, and J. Hertzberg, "Semantic scene analysis of scanned 3d indoor environments," in *Eighth International Fall Workshop on Vision, Modeling, and Visualization*, 2003.

[12] O. M. Mozos, P. Jensfelt, H. Zender, G.-J. M. Kruijff, and W. Burgard, "From labels to semantics: An integrated system for conceptual spatial representations of indoor environments for mobile robots," in *IEEE International Conference on Robotics and Automation (ICRA) Workshop: Semantic information in robotics*, 2007, pp. 33–40.

[13] C. Galindo, J.-A. Fernández-Madrigal, J. González, and A. Saffiotti, "Robot task planning using semantic maps," *Robot. Auton. Syst.*, vol. 56, no. 11, pp. 955–966, Nov. 2008.

[14] P. Viswanathan, D. Meger, T. Southey, J. Little, and A. Mackworth, "Automated spatial-semantic modeling with applications to place labeling and informed search," in *Canadian Conference on Computer and Robot Vision*, 2009, pp. 284–291.

[15] S. Tellex, T. Kollar, S. Dickerson, M. Walter, A. Banerjee, S. Teller, and N. Roy, "Understanding natural language commands for robotic navigation and mobile manipulation," in *Proceedings of the 25th National Conference on Artificial Intelligence*. AAAI, 2011, pp. 1507–1514.

[16] T. Kollar, S. Tellex, D. Roy, and N. Roy, "Toward understanding natural language directions," in *Proceedings of the 5th International Conference on Human-Robot Interaction (HRI)*. IEEE, 2010, pp. 259–266.

[17] D. Chen and R. Mooney, "Learning to interpret natural language navigation instructions from observations," in *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI-2011)*, 2011, pp. 859–865.

[18] P. Singh, T. Lin, E. T. Mueller, G. Lim, T. Perkins, and W. L. Zhu, "Open mind common sense: Knowledge acquisition from the general public," in *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*. Springer, 2002, pp. 1223–1237.

[19] C. Teo, Y. Yang, H. Daumé III, C. Fermüller, and Y. Aloimonos, "A corpus-guided framework for robotic visual perception," in *Workshop on Language-Action Tools for Cognitive Artificial Agents, held at the 25th National Conference on Artificial Intelligence*. San Francisco: AAAI, 2011, pp. 36–42.

[20] C. Teo, Y. Yang, H. Daumé III, C. Fermüller, and Y. Aloimonos, "Toward a Watson that sees: Language-guided action recognition for robots," in *IEEE International Conference on Robotics and Automation*. St. Paul, MN: IEEE, 2012, pp. 374–381.

[21] A. Kasper, Z. Xue, and R. Dillmann, "The kit object models database: An object model database for object recognition, localization and manipulation in service robotics," *The International Journal of Robotics Research*, vol. 31, no. 8, pp. 927–934, 2012.

[22] K. Mourão, L. S. Zettlemoyer, R. P. A. Petrick, and M. Steedman, "Learning strips operators from noisy and incomplete observations," in *Uncertainty in Artificial Intelligence (UAI)*, 2012.

[23] R. P. A. Petrick and F. Bacchus, "A knowledge-based approach to planning with incomplete information and sensing," in *International Conference on Artificial Intelligence Planning and Scheduling (AIPS-2002)*, M. Ghallab, J. Hertzberg, and P. Traverso, Eds. Menlo Park, CA: AAAI Press, Apr. 2002, pp. 212–221.

[24] ——, "Extending the knowledge-based approach to planning with incomplete information and sensing," in *International Conference on Principles of Knowledge Representation and Reasoning (KR-2004)*, D. Dubois, C. Welty, and M.-A. Williams, Eds. Menlo Park, CA: AAAI Press, June 2004, pp. 613–622.

[25] F. Stulp, A. Fedrizzi, L. Msenlechner, and M. Beetz, "Learning and Reasoning with Action-Related Places for Robust Mobile Manipulation," *Journal of Artificial Intelligence Research (JAIR)*, vol. 43, pp. 1–42, 2012.

[26] V. Maz'ya and G. Schmidt, "On approximate approximations using gaussian kernels," *IMA Journal of Numerical Analysis*, vol. 16, pp. 13–29, 1996.

[27] K. Welke, "Memory-based active visual search for humanoid robots," Ph.D. dissertation, Karlsruhe Institute of Technology (KIT), Computer Science Faculty, Institute for Anthropomatics (IFA), 2011.

[28] J. Williams and P. Maybeck, "Cost-function-based gaussian mixture reduction for target tracking," in *Sixth International Conference of Information Fusion*, vol. 2, 2003, pp. 1047–1054.

[29] M. West, "Approximating posterior distributions by mixtures," *Journal of the Royal Statistical Society (Ser. B)*, 1993.

[30] A. Runnalls, "Kullback-leibler approach to gaussian mixture reduction," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 43, no. 3, pp. 989–999, 2007.

[31] T. Asfour, K. Regenstein, P. Azad, J. Schröder, N. Vahrenkamp, and R. Dillmann, "ARMAR-III: An integrated humanoid platform for sensory-motor control," in *IEEE International Conference on Humanoid Robots (Humanoids)*, 2006, pp. 169–175.

[32] T. Asfour, P. Azad, N. Vahrenkamp, K. Regenstein, A. Bierbaum, K. Welke, J. Schröder, and R. Dillmann, "Toward humanoid manipulation in human-centred environments," *Robotics and Autonomous Systems*, vol. 56, no. 1, pp. 54–65, 2008.

[33] Y. Goldberg and J. Orwant, "A dataset of syntactic-ngrams over time from a very large corpus of english books," in *Second Joint Conference on Lexical and Computational Semantics*, 2013, pp. 241–247.

[34] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a Large Annotated Corpus of English: The Penn Treebank," *Computational linguistics*, vol. 19, pp. 313–330, 1993.

[35] M.-C. de Marneffe and C. D. Manning, "Stanford typed dependencies manual," 2008.