Aligned Scene Modeling of a Robot’s Vista Space – An Evaluation

Agnes Swadzba and Sven Wachsmuth
Applied Informatics, Faculty of Technology, Bielefeld University
Universitätsstraße 25, 33613 Bielefeld, Germany
{aswandzba, swachsmu}@techfak.uni-bielefeld.de

Abstract
One kind of meaningful structures in indoor rooms are supporting structures like tables and cupboards. A robot will need to know these structures for a natural interaction with the human and the environment. As bottom-up detection of such structures is a challenging problem, we propose to estimate potential supporting structures from a spatial description like “a bowl on the table”. As language and cognition schematize the space in the same way it is possible to estimate the representation of the space underlying a scene description. To do so, we introduce the aligned modeling approach which consists of rules transforming a sequence of object relations into a set of trees and a methodology to ground the abstract representation of the scene layout in the current perception using detectors for small movable objects and an extraction of planar surfaces. An analysis of 30 descriptions shows the robustness of our approach to a variety of description strategies and object detection errors.

Introduction
Spatial awareness and the ability to communicate about the environment are key capabilities enabling an agent to perform day-to-day navigation tasks. In general, space can be partitioned along the actions that are required to perceive it (Kuipers 2000; Montello 1993). The perception of the large-scale space involves locomotion of an agent while the vista space can be explored quickly from a single view point by eventually moving the gaze. Applying this definition to a domestic environment, a complete apartment belongs to the large-scale space and single rooms or room parts to the vista space. Since the ability to navigate and to localize itself is essential for a mobile robot, much research has concentrated on modeling the robot’s large-scale space providing methods for Simultaneously Localization and Mapping (SLAM) (Thrun, Burgard, and Fox 2000) and motion planning (Philippsen and Siegwart 2003). Analyzing percepts of the immediate surroundings becomes relevant for a robot if obstacles must be avoided, e.g., (Yuan et al. 2009), or an unknown environment has to be explored, e.g., (Topp and Christensen 2008). But a vista space analysis is even more important in situations with agent-agent interaction as it is the space where the interaction takes place. A collaborative interaction of the partners in and with the environment requires that both partners know where they are, what spatial structures they are talking about, and what scene elements they are going to manipulate. Our research focuses on the new area of 3D spatial modeling for human-robot interaction. A commonly used scenario where a robot is expected to acquire interactively spatial knowledge is the so-called “home tour” scenario (COGNIRON 2004). There a robot is shown around in an apartment and taught relevant scene information. In this paper we focus on situations where a room is introduced in details through a description. We address here the question of how relevant supporting structures can be inferred from a depiction and grounded visually, so that the resulting model is aligned to the describer’s representation. From psychological research we know that space is encoded in the so-called situation model (Zwaan 1999). Therefore, spatial descriptions play an important role in exchanging concepts about the surrounding. As linguistic experiments propose that cognition and language schematize the spatial world in the same way (Tversky and Lee 1998; Freksa 2008) language can be seen as a systematic framework for conveying space by selecting or neglecting certain aspects (Talmy 1983). This means that a hearer builds a model from a description which is similar to the model the speaker has built from perception (Waltz 1980). Further, models of communication partners will be aligned if information about the surrounding is exchanged verbally (Pickering and Garrod 2004). This allows a smooth communication as similar concepts are represented (Vasudevan et al. 2007). The problem with spatial descriptions is that they are underspecified (Waltz 1980) which means that an agent needs also a sensory input to resolve ambiguities in the descriptions. Based on this requirement and the findings about the nature of spatial descriptions we have proposed a computational model, called aligned scene model, which enables a robot to infer supporting scene structures from a description and to ground them in its visual perception of the scene (Swadzba et al. 2009). This paper gives an overview of the aligned scene modeling approach for a robots’ vista spaces and gives a detailed analysis of the models generated from a significant large collection of different descriptions with respect to description strategies and object detection errors.
**Related work**

This section discusses relevant work on inferring scene models from verbal descriptions, extracting intermediate scene structures from visual perception, and integrating verbal and visual scene interpretations. First systems like the Spatial Analog Model have transformed spatial prepositions like “the box is on the table” into a 3D box model satisfying gravity conditions (Boggess 1979). The Words-into-Pictures approach extends this by constructing objects from a finite set of geometric primitives and incorporating constraints on orientation and position from object relations into a potential field (Olivier, Maeda, and Tsuji 1994). In contrast to this abstract modeling, recent robotic methods provide meaningful labeling (e.g., as “wall”) of 3D points using top-down knowledge encoded in ontologies (Nüchter and Hertzberg 2008; Cantzler, Fisher, and Devy 2002; Grau 1997) or bottom-up information from point-based features (Munoz et al. 2009; Rusu et al. 2008; Triebel et al. 2007). So far, the integration of verbal descriptions and visual perception has been studied with single caption words (Jamieson et al. 2010; Wang, Hoiem, and Forsyth 2009) or simple table-top scenarios. Spatial relations and visual perceptions have been associated in a probabilistic manner using Bayesian inference processes (Wachsmuth and Sagerer 2002), stochastic layers (Mavridis and Roy 2006), or potential fields (Brenner et al. 2007).

**The computational model**

This section introduces the computational model for acquiring an aligned scene representation which reflects the tutor’s scene representation and the visual reality perceived as bottom-up extracted 3D planar patches. An abstract representation is inferred from a scene description and grounded in the visual perception utilizing object detection and surface extraction results.

**From verbal descriptions to set of trees**

Studies on the nature of spatial descriptions have revealed that spatial knowledge is organized hierarchically (Hirtle and Jonides 1985). This is visible in the partitioning of space into figures and spatial relations (Tversky and Lee 1998) and the use of horizontal and vertical planes as reference frames (Tversky, Kim, and Cohen 1999). As the spatial relations are a prominent part of spatial descriptions we have conducted an empirical study to learn more about the nature of these relations. In this study 30 free-form descriptions of two rooms presented as pictures have been collected (→ Figure 1). It can be assumed that depictions of true scenes (Henderson and Hollingworth 1999) are more realistic than descriptions of ersatz scenes (displays of arbitrary arrayed objects). This allows us to use these descriptions to test our scene modeling on 3D data acquired from the same view on the scene as the pictures have been taken. A detailed analysis of the descriptions shows that spatial room structures like pieces of furniture or other room parts serve as crystallization points for room descriptions (Vorweg and Rickheit 2009). Typical relations are “the lion is on the chair” or “a toy car is in front of the koala” (→ Figure 1(a)). Objects are put into relation with their supporting room structure or with other objects being on the same room structure. This can be seen quantitatively in Figure 1(c) which shows the usage frequencies of different relation types. The majority of specified relations are relations between objects on the same super-ordinate structure (B) or relations between object and the corresponding super-ordinate structure (A, nearly never R). Therefore, we propose the following terminology:

**Definition. Orthogonal relation,**

\[ \text{Parallel relation}, \quad \text{a relation between an object and its super-ordinate structure (e.g., “a lion on the chair”) \} } \\

**Definition. Parallel relation,**

\[ \text{a relation between objects localized on the same structure (e.g., “a car in front of the koala” both are on the table)} \]

These definitions follow the basic physical fact that due to gravity objects are placed on tables, attached to walls, or contained in cupboards (Torralba, Russell, and Yuen 2009; Olivier, Maeda, and Tsuji 1994). The explicit and implicit references to super-ordinate structures reflect the hierarchical character of the underlying situation models, so that tree-like structures are a suitable representation to maintain such models. Handling relations only based on their assignment to the orthogonal or parallel type allows a scene modeling that is independent from the exact view on the scene and the used reference frames. The challenges faced when transforming a description into a tree representation \( \mathcal{T} \) are incompleteness, ambiguity, and disordered arrangement. To cope with these problems, we assume that descriptions are sequences of object relations. An existing tree set \( \mathcal{T}_{n-1} \) is updated by the current relation to \( \mathcal{T}_n \) according to certain rules. Figure 1(d) introduces our compact illustration of tree sets. The nodes correspond to scene items while the edges indicate the relations between them. Each parent node is the supporting structure for all its children. A tree set is updated by \( o = \text{obj}(“o”) \rightarrow n_o \), which inserts for an object \( o \) with the label “o” a node \( n_o \) into \( \mathcal{T}_n \). delete \( (n) \) deletes the node \( n \) from tree set \( \mathcal{T} \) and child \( (n_o, n_p) \) inserts a directed edge from the parent node \( n_p \) to the child node \( n_o \) expressing the supporting characteristic of the parent structure to the child object.

A **parallel** relation, \( \text{rel}_p(o_1, o_2) \), between object \( o_1 \) and \( o_2 \) is handled by the basic rule shown in Figure 2(a). Their supporting structure is inserted as new node \( n_p \) where \( n_{o1} \) and \( n_{o2} \) become its child nodes. If one of the related objects has already a parent the exception rule is applied Figure 2(b).

An **orthogonal** relation, \( \text{rel}_o(o_1, o_2) \), is handled as shown in Figure 3(a). A directed edge between the nodes \( n_{o1} \) and \( n_{o2} \) implements the given relation. In the exceptional case both trees or subtrees can be fused to one tree with \( n_{o2} \) as the root node if the existing parent node has an empty label or “o2” as label (see Figure 3(b)).

Figure 4 shows an example description and its tree set representation resulting from the sequence-wise processing of the description with the rules defined above.
Grounding the scene model in the 3D perception

This section presents how the set of trees resulting from the computation presented in the previous section can be mapped on a 3D perception (e.g., a SwissRanger scan) of the described scene. The challenges are ambiguous labels like object categories (e.g., “soft toys”) and parent nodes with an empty label, the so-called “virtual patches”. As man-made environments mainly consist of planar surfaces it is reasonable to estimate supporting structures as planar patches. We propose to compute the patch parameters using the 3D locations of small movable objects attached to them. We assume that detectors for small compact objects, which are not part of the spatial layout of a room like “lion”, “koala”, . . . , are easier to train and detect more robust than detectors for spatial structures like “chair” or “table”. As the development of a robust object detection system is out of the scope of this work, we have labeled manually objects in the SwissRanger amplitude image simulating an output of an object detector (→ Figure 5(a)). The following algorithm consists of four steps. First, the confirmed objects of a parent node (distinct objects in the scene) are identified. Second, these objects and the corresponding relation type are used to estimate the patch parameters of the parent nodes. Third, the estimated patch is used to resolve existing category labels. And fourth, ambiguities and errors are handled.

(I) \( O \), the set of objects known to the robot (→ Figure 5(a)) is divided for a certain description (→ Figure 4) into a set of confirmed objects \( O_{\text{con}} \) which are those objects that have been mentioned in the description and a set of potential objects \( O_{\text{pot}} \) holding the remaining objects.

(II) For each parent node \( n_p \in T \) its potential planar patch \( \mathcal{P}_p \) is computed using its distinct objects \( O_{\text{con}} \subseteq O_{\text{con}} \). A planar patch is described by an orientation and a position in the global coordinate system:

\[
\mathcal{P} : \mathbf{n} \cdot \mathbf{x} - d = 0.
\]

Knowledge about the type of relation used to assign objects supports the computation of the patch orientation. A patch where objects are located on is estimated as horizontal patch:

\[
\mathbf{n}_p = (0, 1, 0)^T.
\]

A supporting structure containing objects (indicated by an in-relation) has a vertical orientation:

\[
\mathbf{n}_p = \mathbf{a}_p \times \mathbf{b}_p.
\]
(b) a vertical plane

Figure 6: The figures show two potential patches. A patch is represented by an ellipse, its normal vector \( \vec{n}_p \), and its centroid \( \vec{c}_p \). Each object is represented by its convex hull and its 3D location \( \vec{l}_i \), \( \vec{r}_{obt} \), \( \vec{r}_{vrai} \), \( \vec{r}_{red} \), \( \vec{r}_{pokemon} \).

where \( \vec{a}_p = (0, 1, 0)^T \) and \( \vec{b}_p \) the orientation of a line fitted robustly via RANSAC through the object points projected in the \( xy \)-plane (ground plane). If the relation between child and parent node is not known, the orientation of the 3D arrangement of the confirmed objects with respect to the ground plane determines the orientation of the potential patch (if \( \theta < 45^\circ \): horizontal else: vertical). The position of the patch is computed from the centroid \( \vec{c}_p \) of the objects’ 3D locations: \( \vec{d}_p = \vec{n}_p \cdot \vec{c}_p \). Figure 6 visualizes a horizontal and a vertical potential patch.

(III) Finally, \( P^\text{pot} \) can be utilized to resolve possible category labels. Objects in \( O^\text{pot} \) having the specified category and a distance to the corresponding patch smaller than a given threshold are assigned to the potential patch. It is assumed that the human tutor has referred to these objects. Figure 7(a) shows the initial scene model of the example description in Figure 4. Patches for “corner”, “table”, “chair”, “cupboard2”, and “cupboard3”, and two virtual patches with an empty label “.” are estimated. The category label “soft-toy on the table” is resolved to “kangaroo”, “bear”, “frog”, and “stitch”. The category label “games in cupboard3” is resolved to “games1” and “games2”.

(Final) The initial scene model contains errors (cupboard3 is too big) and virtual patches (patches with an empty label “.”) due to errors in the resolution of category labels and underspecifications in the descriptions. Both problems can be solved by incorporating bottom-up extracted patches \( \{ P^\text{real}_i \}_{i=1}^{m} \). A real patch \( P^\text{real} \) assigned to a potential patch \( P^\text{pot} \) if they share a similar orientation and position in space. Several real patches can be assigned to one potential patch and one real patch can be assigned to multiple potential patches. The matching type indicates the status of an initial scene model. If it contains only injective mappings from real patches to potential patches the model is correct while a constellation where one real patch \( P^\text{real} \) is assigned to several potential patches with competing labels, e.g., “p1” and “p2”, means an object mismatch. To correct this, all objects of \( P^\text{pot}_1 \) and \( P^\text{pot}_2 \) are checked whether they are positioned in/on the real patch \( P^\text{real} \). \( P^\text{real} \) is put to that potential patch holding the biggest percentage of objects lying in/on \( P^\text{real} \).

In the case where a bottom-up patch \( P^\text{real} \) is assigned to a set of potential patches where exactly one patch has a non-empty label all parent nodes can be merged to one node labeled with “p” removing virtual patches.

Evaluation

Figure 8 shows the generated aligned scene models based on the 30 descriptions collected in the studies. The room structures are represented by ellipses and are localized on level 1 in the tree sets \( T \) since objects positioned in the leaves of the trees are used to estimate these structures. The histograms in Figure 9 show the distribution of the structure labels found in the models. A structure appearing in many models is a prominent element of the scenario. In general, no differences in the descriptions of the pilot study and the main study can be observed. Each model is aligned to the specific level of detail given in the description. Most of the descriptions are quite detailed with many relations between objects and room structures. There are only few descriptions where objects have been simply listed (compare model 1, 2, 5, and 7 in Figure 8(a) where nearly no spatial structures have been concluded). In the playroom scenario (→ Figure 9(a)) the “table” and the “chair” are the most dominant structures. Further, the shelves at the wall are also interesting structures. Half of the participants have given enough information to conclude patches for “cupboard2” and “cupboard3”. The individual reference of the shelves may be primed by the different colors of the shelves as the contrary effect is visible in the descriptions of the living room \( S_2 \) where the shelves have the same color and are fused to the structure “cupboard” (→ Figure 9(b)). The other meaningful structures in the living room are “table”, “sofa”, and “wall”. In general, the system fails to compute a potential patch for a spatial structures (indicated by “Cannot compute potential patch for ...”) when only the category of objects located in resp. on the structure is known (like “there are soft toys in the cupboard”). Without knowing at least one specific object this category label cannot be resolved. Finally, it can be said that all generated models are an appropriate representation

Planar patches can be extracted via region growing utilizing the conormality and coplanarity value of point pairs (Stamos and Allen 2002). Several runs of the RANdon SAmple Consensus (RANSAC) algorithm (Fischler and Bolles 1981) improve the initial regions. A detailed description of the planar patch extraction algorithm can be found in (Swadzba and Wachsmuth 2008).
of the respective scenario. The analysis of this large set of
descriptions ensures that we have covered different description
strategies. The strength of our approach is that for ev ery
description room structures can be inferred and modeled
across different scenes and appearances and that the level of
details in the model is aligned to the given description pre-
serving resources.

As assuming perfect object detection is not realistic we
have also examined the influence of typical errors on the
model formation process. Figure 10 shows an object set
where typical errors like scaling (bounding box is too small
or too big), translation (bounding box is misplaced), and
missing errors (no bounding box available) have been ap-
plicated randomly selected objects. Half of the objects in
the set of known objects is affected by the errors. The figure
presents also the scene model estimated using this erro-
neous object set. The table in Figure 10 gives the deviation of
the model from the ground truth model in Figure 7(b). Over-
all, the patches are estimated robustly. “cupboard2”, “table”,
and “corner” show only minor deviations from the original
patches. The twisting of “cupboard3”-patch by 14.53° is ac-
ceptable since the overall position of the patch is correct. It is
still possible to assign the correct real patch to “cupboard3”
because we allow due to noise a deviation of up to 30° be-
 tween the potential patch normal and the real patch normal.
Only the “chair”-patch is too big and misplaced. If 3D data
of the chair itself could be perceived step (IV) would detect
that the “robot”-object is misplaced with respect to the chair
as it handles mismatched objects arising from errors when
resolving category labels. As the computation of the 3D ob-
ject hulls incorporates a removing of outlier points it can
handle slightly misplaced object boxes. Further, the orienta-
tion of the horizontal patches is only influenced by the ori-
entation of the camera and the RANSAC based approach for
computing the vertical patches can deal with remaining out-
liers. Also, missing objects can be compensated if enough
other objects are known for the corresponding parent node.
To summarize, our method proposes a robust estimation
of the patch orientation that can handle object outliers. The
computation of the patch position is more sensitive to noise.
If enough correct objects are given a wrong object position
 can be averaged out. But if only one object is associated to
a parent node a big translation or scaling error will corrupt
the potential patch of this node. In general, the corruption
affects the position and expansion of a node patch.
Conclusion

In this paper we have analyzed the performance of the aligned scene modeling approach regarding robustness to different room description strategies by analyzing a large set of descriptions. The computational model is based on the fact that the construction of spatial descriptions is driven by gravity, which means that the descriptions mainly consist of orthogonal relations between objects and their supporting structures and parallel relations between objects located on/in the same structure. The classification of relations in a description only into two types allows the design of compact rules that can transform relations to an abstract tree representation independent from the order in which the relations are given. The evaluation shows that the strategies developed for grounding the abstract model in the perception of the scene are robust to missing data and object detection errors.

Acknowledgement. The work has been funded by the CRC673 “Alignment in Communication”.

References


