

Dialog-Based 3D-Image Recognition Using a Domain Ontology

Joana Hois¹, Michael Wünnst², John A. Bateman¹, and Thomas Röfer³

¹ I1-[OntoSpace], SFB/TR8 Spatial Cognition, Universität Bremen,
Postfach 330 440, 28334 Bremen, Germany
joana@informatik.uni-bremen.de, bateman@uni-bremen.de
<http://www.ontospace.uni-bremen.de>

² A2-[ThreeDSpace], SFB/TR8 Spatial Cognition, Universität Bremen,
Postfach 330 440, 28334 Bremen, Germany
wuenstel@informatik.uni-bremen.de
<http://feynman.informatik.uni-bremen.de>

³ DFKI Lab Bremen, Safe and Secure Cognitive Systems,
Robert-Hooke-Straße 5, 28359 Bremen, Germany
thomas.roefer@dfki.de
<http://www.dfki.de/>

Abstract. The combination of vision and speech, together with the resulting necessity for formal representations, builds a central component of an autonomous system. A robot that is supposed to navigate autonomously through space must be able to perceive its environment as automatically as possible. But each recognition system has its own inherent limits. Especially a robot whose task is to navigate through unknown terrain has to deal with unidentified or even unknown objects, thus compounding the recognition problem still further. The system described in this paper takes this into account by trying to identify objects based on their functionality where possible. To handle cases where recognition is insufficient, we examine here two further strategies: on the one hand, the linguistic reference and labeling of the unidentified objects and, on the other hand, ontological deduction. This approach then connects the probabilistic area of object recognition with the logical area of formal reasoning. In order to support formal reasoning, additional relational scene information has to be supplied by the recognition system. Moreover, for a sound ontological basis for these reasoning tasks, it is necessary to define a domain ontology that provides for the representation of real-world objects and their corresponding spatial relations in linguistic and physical respects. Physical spatial relations and objects are measured by the visual system, whereas linguistic spatial relations and objects are required for interactions with a user.

1 Introduction

Dialogs between humans and robots about natural objects in a visual scene have always required interdisciplinary research. Different components from linguistics

and psychology for generating and comprehending natural dialogs, on the one side, have to be combined with image/vision processing and knowledge representation for a formal representation of the respective objects, on the other.

In this paper, we introduce a system that analyzes a partial 3D-view of an office, in particular with a view toward desktops and the configurations of objects that are found. These 3D-images of real world scenarios are taken by a laser scanner. The images are processed by a visual component that segments and classifies the occurring objects (object types). Based on this analysis, the user is invited to correct those object classifications, so that it is possible to ask about the objects with regards to their spatial relations afterwards. The language used in the dialog and for object representation is English. This linguistic component is mainly divided into two phases, which we call the *training phase* and the *action phase*. During the training phase, the system enlists the assistance of the user for determining the types of those objects that could not be classified unambiguously by the visual component. The resulting object types from this dialog are transferred into a domain ontology, which offers a knowledge structure and additional information about the domain. During the action phase the user can ask about objects by referencing their spatial relations. Questions about objects that have a certain spatial relation to other objects and questions about object types are currently supported.

This capability enables our system to provide a basis for applications in diverse contexts. For example, it can be extended for assigning a robot instructions about activities it has to perform, while referring to objects via their spatial positions and relations. Another possibility is to generate a scene description of the 3D-view, possibly for blind people. Our main focus at present, however, is to develop this system for examining and improving the following three mutually supportive aspects:

- **Cognitive Vision.** The field of computer vision is very domain specific in general. The area of application ranges from medical to technical inspection tasks. The objective of cognitive vision is to expand the vision system’s abilities by employing cognitive properties, which are seen as being focused on these individual domains. For the perception system developed, this means that the system is able to deal with uncertainties and can learn. It may offer more than one possible solution and this can be interpreted either within the subsequent user dialog or by ontology-based deduction.
- **Linguistic User Interaction.** A human-computer dialog can improve the performance of a vision module by involving the user in the object classification and by letting the user ultimately determine an object’s type. If there are objects that could not be classified after 3D-object segmentation and perception, the user can specify the type. Moreover, the dialog component also provides the possibility to ask questions about objects in the scene with respect to their spatial relations. In this case, the system has to clarify which physical spatial relations and objects, measured by the vision system, agree with the spatial relations and objects linguistically expressed by the user.

- **Domain Ontology.** In order to combine a natural language component with an image recognition system, we need to consider the formal definition of the recognized objects that are to be referenced linguistically. In addition to those objects, the spatial relations that hold between them also need to be represented in a formal structure. This structure should not only provide a distinct representation of the concepts that are used by the user but also support reasoning strategies in order to deal with user requests and to improve the object classification. We will introduce a domain ontology guided by an ontological library for linguistic and cognitive engineering, which seems to be particularly appropriate to our problems while also offering a solid basis for further development.

There are few approaches that try to combine all of these issues. For example, in [16] a system is introduced that generates linguistic descriptions of images with references to spatial relations. Their emphasis is on relating linguistic expressions of spatial relations to the physical alignment of objects using fuzzy sets. In contrast to this work, we also want to support a reverse dialog, in which the user can use expressions of spatial relations and the system has to detect them. In addition, the use of a formal specification of concepts by means of a domain ontology seems to be more promising, as it directly supports reasoning strategies and provides a fundamental structure for linguistically expressed spatial relations as well as for the representation of spatial reference systems. Furthermore, no aspects of cognitive vision in their system are considered.

In [20] a cognitive vision system is presented that generates natural language descriptions from traffic scenes. This framework uses also a fuzzy logic formalism to present schematic and conceptual knowledge, and it especially aims to explore the destination of objects moving in the scene. Apart from the different domain that is used in this approach compared to ours, the system in [20] scans the input images to detect vehicles and derive further information about the scene, for example the position of lanes. Instead of searching for specific objects in the scene, our system segments all objects independently of their types. Therefore, we have to deal with multiple object types and even with objects that are not classifiable, which is also a reason why we access a domain ontology. But generally, the system described in [20] is more concerned with the tracking of objects and their motion than with their classification and their relative spatial positions.

The system *DESCRIBER*⁴ is also a system that combines spoken language and object recognition and is able to generate phrases about objects in a scene. For this purpose, the system is trained by a data record of synthetic scenes paired with spoken descriptions about the objects in the scene, their colors, size, shape, and relative spatial relations. In contrast to [16] and [20], this system is also combined with a robotic system that supports the detection of objects corresponding to the description in novel spoken phrases [22]. The vision component of the robot uses two color video cameras for the recognition of scenes.

⁴ <http://www.media.mit.edu/cogmac/projects/describer.html>

Although the application of this system is similar to our system, the realization is very different: We are using 3D images taken by a laser, and do not need to calculate 3D information of a scene on the basis of two 2D images. Moreover, instead of recognizing objects on the basis of their color, we have developed a vision component which perceives objects not only by physically measured values but also by using additional background knowledge and techniques of cognitive vision. And instead of learning spatial relations via respective data records, our system uses general spatial calculi for generating linguistic expressions of spatial relations.

2 Cognitive Computer Vision

The field of cognitive vision has the intention of transferring ideas from cognition into computer vision. Cognitive vision can therefore be seen as a special domain within the field of computer vision. Its goals are described in [6]:

“The term cognitive vision has been introduced in the past few years to encapsulate an attempt to achieve more robust, resilient, and adaptable computer vision systems by endowing them with a cognitive faculty: the ability to learn, adapt, weigh alternative solutions, and even the ability to develop new strategies for analysis and interpretation.”

The realization of these goals can be achieved, e.g., by a virtual commentator which transforms the visual information acquired into a textual scene description [21]. To obtain such a commentator many of the claims of a cognitive vision system have to be realized. As we want to improve our object recognition system by the above cited approach, the system is motivated by it and complies with its respective claims.

Moreover, we see textual scene description as a contribution to the range of techniques by which the abilities of a system can be tested and evaluated. These aspects are therefore also part of this work and are integrated into our object recognition system: The system is supposed to be expandable, offer alternative solutions and make them available for formal (textual) interpretation.

2.1 The Object Recognition System ORCC

The recognition system ORCC⁵ (see Fig. 1) identifies and classifies objects within an indoor environment and strives for the long term goal for a scene interpretation. The system comprises modules based on the fields of computer vision, cognitive vision and computer graphics. The central component, computer vision, involves a range of methods from density segmentation up to bounding-box calculation. The module cognitive vision contains techniques that build on the results of the computer vision part. Typical cognitive capabilities supported here are the ability of learning to classify objects preferably independent from particularly designated “special” features. Thus objects are classified according to their

⁵ Acronym for **O**bject **R**ecognition using **C**ognitive **C**omputing

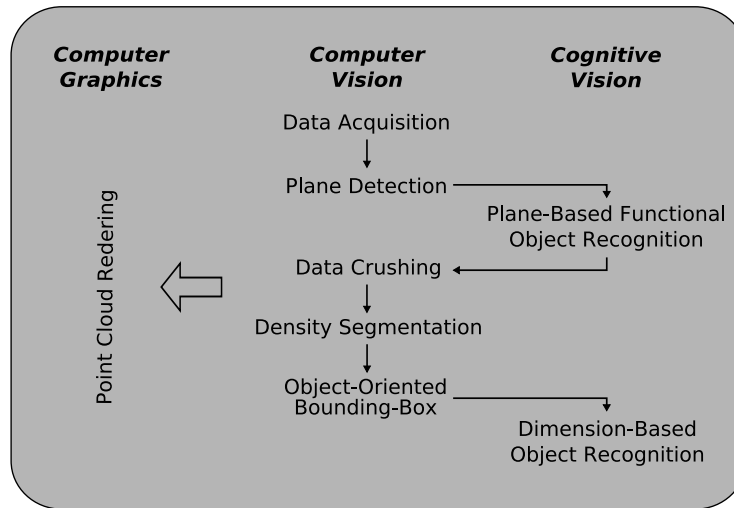


Fig. 1. ORCC System

underlying functionality (or applicability) not by some special characteristic. A table, for example, is defined by a horizontal plane within a certain height interval. The goal therefore is not to recognize a special class of objects but categories in the sense of [21]. The results are processed in a way to use them for a textual description of the scene. The third module, the computer graphics component, is used to visualize the scene and the result of the speech interaction.

2.2 Components of the ORCC System

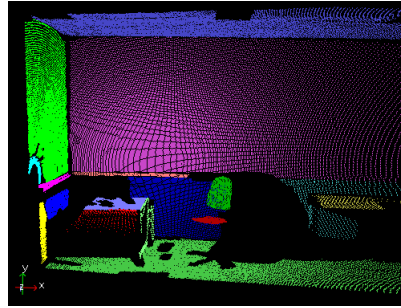
As the system uses the raw data of a real scene several features are necessary, described as follows:

Sensor The laser sensor mounted on a pan-tilt unit provides depth information from a viewpoint about 1.35 m above the ground (see Fig. 2(a)). The system has a horizontal resolution of 0.5° and the scanner is also tilted in 0.5° steps. The accuracy of the depth estimation is about 1.0 cm.

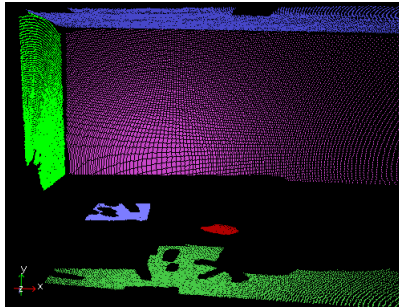
Plane Detection For plane detection a normal vector based algorithm is used [23]: The normal vector in every data point is calculated using its corresponding surroundings. Only if the local quality of the normal vector (given by the eigenvalues within a Singular Value Decomposition step) and neighbor quality (given by the comparison with neighboring normals) falls within an appropriate range of tolerance it is used. The segments themselves are obtained via a region growing algorithm (see Fig. 2(b)).



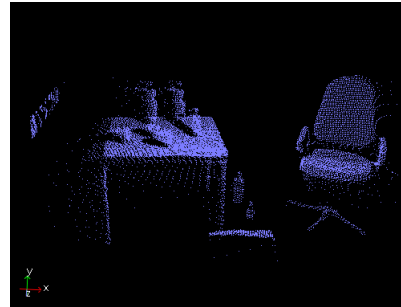
(a) Sensor



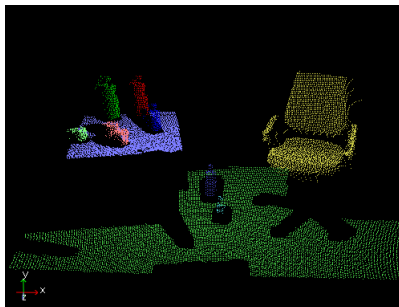
(b) Detected Planes



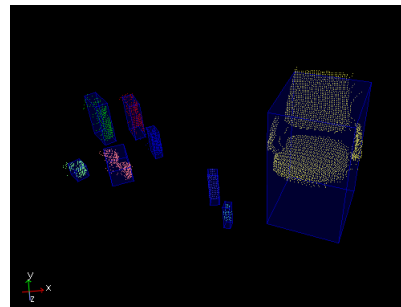
(c) Plane-based functional object recognition: floor, ceiling, walls, tabletop



(d) Data crushing



(e) Density segmentation



(f) Object-oriented bounding-box

Fig. 2. Diverse stages of the ORCC System

Plane-Based Functional Object Recognition The task of this non-final classification step is to find objects that are necessary for the segmentation of an object (see Fig. 2(c)). The object models used here hold direct attributes like *orientation* or *size* and relational attributes like *distance to* or *deviation of the orientation from the horizontal plane*. A tabletop (the plane of the tabletop), on the one hand, is located within a certain interval above the floor, has a size that ranges within a certain interval, and its orientation is horizontal. On the other hand, a wall has a certain distance (ideally zero) to the ceiling and is oriented vertically. A common approach for the modeling of such problems is to use Bayesian networks [13]. The structure of the network is given by the different features of the objects. The result for each object is a multivariate Bayesian function. The mean and the variance values are determined analytically and are stored within an XML file.

Data Crushing In the overall approach there is no search for specific objects but an estimation of segmented objects. To be able to segment the objects based on their three-dimensional point distribution, it is necessary to remove structural elements of the scene, such as walls (see Fig. 2(d)). After the removal of the space enclosing points (objects), the predominantly interesting points are obtained.

Density Segmentation The segmentation step detaches objects from their surroundings. In some approaches, special models (e.g. cars) are determined within a scene [1]. But because we also need to be able to segment unknown objects, a point density approach is chosen. As the objects are predominantly positioned on planes like tabletops or floors, this information is used for the density segmentation (see Fig. 2(e)).

Object-Oriented Bounding-Box Calculation The calculation of the object-oriented bounding box (see Fig. 2(f)) is a preliminary step of the dimension-based object recognition procedure. It is separated from the latter as it has to deal with tail points, which appear at edges; these are due to special characteristics of the scanning equipment [17].

Dimension-Based Object Recognition The dimensions of an object category are described by naïve, continuous Bayesian networks. Every object class corresponds to a single Bayesian Network that is parameterized by its mean and variance values. These values are determined analytically using the values from a training set. Given the dimensions of an object resulting from the step before, the grade of membership to each class is determined. The object is classified as belonging to the class with the highest probability value, although competing alternatives are also given [25].

3 Representation of Objects and Spatial Relations

For our purposes we need to represent the scene in terms of its occurring objects and their relations, which requires explicit taxonomical and partonomical structures. This representation is also affected, however, by the linguistic interaction intended between the user and the system. Hence, the structure has to take into account linguistic expressions that possibly correspond with the scene together with their meanings. To cover all these criteria, we decided to make use of an ontological structure that directly supports our requirements.

Ontologies are nowadays widely used in diverse applications with different granularities and with varying formal considerations. They are used to clarify the concepts being modeled and the relations holding between them. They also provide a reasoning mechanism by the ontological structure itself. However, there is still no standard conceptual modeling guideline, although there are series of principles by which ontologies can be modeled and classified [8]. Ontologies have also raised attention since the beginning of the Semantic Web development [3], which is intended to provide a navigable structure aligned by its semantic concepts. It is also a main focus to ensure re-usability and sharing of this ontologically defined knowledge.

In the past, several research efforts have been made to represent space and spatial relations using insights from ontological development (see [2] for an exhaustive overview). The work reported here applies one of these efforts in order to examine its practical usability and the necessity for further extensions. For computability reasons, we use description logic for the development of our domain ontology.

3.1 Domain Ontology Modeling

We have created a domain ontology that consists not only of objects that can be recognized by the ORCC system or that are known to our image recognition tools, but also of general objects that can be present in an office scene and that we have already explored in earlier office domain ontologies [12]. Its structure is guided by an ontology framework with an emphasis on linguistic and cognitive aspects, namely DOLCE [15].

Originally developed as a part of the WonderWeb project⁶, DOLCE was designed as a foundational ontology. As it is strongly influenced by natural language and human commonsense, it seems a promising basis for our current work. Its main concepts are divided into four categories, to wit *Perdurants*, *Endurants*, *Qualities*, and *Abstracts*. Perdurants describe entities that unfold in time (e.g., an event of some kind), whereas Endurants describe entities that are wholly present at each point in time (e.g., a cup, a glass, a desk, etc.). Qualities inhere in entities, so every entity can have certain qualities, even Qualities themselves. The value of these qualities are expressed by entities of the *Abstract* concept,

⁶ <http://wonderweb.semanticweb.org>

called *quale*. A quale describes a position of an individual quality in a *quality space*, which corresponds to *conceptual spaces* described in [9].

Objects in ordinary office scenes are types of the concept Endurant, more precisely of its subconcept *Physical Object*. Such concepts are defined by having direct spatial qualities and as being wholly present in time. The object types we have modeled in our domain ontology are hierarchically structured by means of their functions, as emphasized in [5]. This functional bias is of particular importance: In the long term, we are planning to connect object recognition with possible robot actions that depend on ‘functional roles’ an object can have. For instance, additional information about a mug, as some kind of a drinking vessel, is that it can be filled with liquid, an instance of *Amount Of Matter*. Therefore mugs, cups, coffeepots, bottles are subsumed under drinking vessel; staplers, hole punchers, scissors are subsumed under office-supply, etc. This information can then be derived and used for performing possible actions, such as drinking. In addition, a human-computer interaction can linguistically refer to such actions.

In our domain ontology, we provide a fundamental data set for object types that are common in an office. If the system detects an object type that cannot be classified and the user suggests a type that is not known to the ontology, the system will handle this by classifying this concept simply as a physical object. If the user refers to this object later by using a specific type that is already known from the domain ontology, the system will refine the respective object type and for this scene the user can refer to this object with both types.

3.2 Spatial Relations

Spatial relations between objects in a scene are used for the linguistic user dialog during the *action phase* (see section 4), in which the user may use relations like “Show me the object that is behind the book” and the system has to infer what “behind something” means and which book the user is referring to, regarding the current object configuration of the scene.

To describe spatial relations in linguistic terms, we have to consider the different linguistic frames of reference that lead to these terms and that are employed by the user. An essential result from linguistic and psychological research concerning linguistic expressions of spatial relations is the general difference between three main frames of reference, named *intrinsic*, *relative*, and *absolute* [14].

In the first case, spatial relations between two objects are described by the position of one object (the *referent*) relative to an intrinsic orientation of another object (the *relatum*). In relative reference systems, the relative position of a referent to its relatum is described from the viewpoint of a third position (the *origin*). Finally, if the description of spatial relations between two objects is made with respect to some externally fixed direction, we speak of an absolute reference system. For a detailed discussion of reference systems and their literature, as well as extensive examples of linguistic terms involving spatial relations, see [24].

With the current status of our system, we consider only projective linguistic expressions of spatial relations corresponding to a relative reference system. As

absolute reference systems do not seem to be employed for indoor office environments in our western culture [19], this restriction is well motivated and therefore we do not take into account absolute relations at present. The integration of intrinsic references is also left to future work, although this is important because intrinsic references are very common in natural language [11].

As indicated above, a spatial relation comprises physical objects each being origin, relatum, or referent. In DOLCE, a Physical Object has an individual spatial (and temporal) quality that defines the physical position of the object in space (and time) [15]. The quale of this spatial quality is a spatial region in geometric space. Although the properties of the geometric space adopted is not further defined in DOLCE, they are intended to represent the absolute value of the spatial position of one physical object. They are not supposed to express spatial relations that belong to linguistic terms and describe assignments of abstract concepts to certain physical conditions between physical objects. Hence, spatial relations addressed in this paper are not represented in the existing ontological framework yet.⁷

In order to deal with this central area of spatial conceptualization, we are exploring a novel definition of spatial relations as instances of the DOLCE category *Non-agentive Social Object* that relate the objects to their relation. This is because Non-agentive Social Objects in DOLCE are defined as entities that are generically dependent on a community of agents. In this sense linguistic terms of spatial relations are in this sense socially constructed and maintained. We call this entity *Spatial Relation*. The values, it can take, are subsumed under the abstract concept *Spatial Relation Region*. Instances of Spatial Relation take as arguments three Physical Endurants distinguished by being origin, relatum, or referent, and one Spatial Relation Region, the value of the spatial relation. This is illustrated in Fig. 3. Even though we are currently only considering relations in terms of a relative reference system, this representation also covers intrinsic and absolute reference systems. In this case origin and relatum are equal.

In our application, a distinguished viewpoint of the user (the origin), on which the spatial relations of the objects depend, has to be declared. As the user’s viewpoint is equal to the position of the laser with respect to the scene, we can use an initial instantiation for the laser in our domain ontology as origin. Although it is possible for the user to have a position different from the laser, we are currently constrained to the laser’s viewpoint as this is where the scene is scanned from and displayed to the user. This is also done with respect to the scope of our application: A robot getting instructions from users that share the same viewpoint toward the scene from the position of the robot, when users are not in the same room as the robot or when they are located at the same position as the robot (for example, in case of a wheelchair).

⁷ Note: In DOLCE, there are also *places*, e. g. “the underneath of a table”, defined as a *Feature* of objects. But this should not be confused with spatial relations, like “something *is* underneath a table”, because such places are regarded as “parasitic entities” that constantly depend on their *hosts*; they do not directly act as physical spaces for other objects.

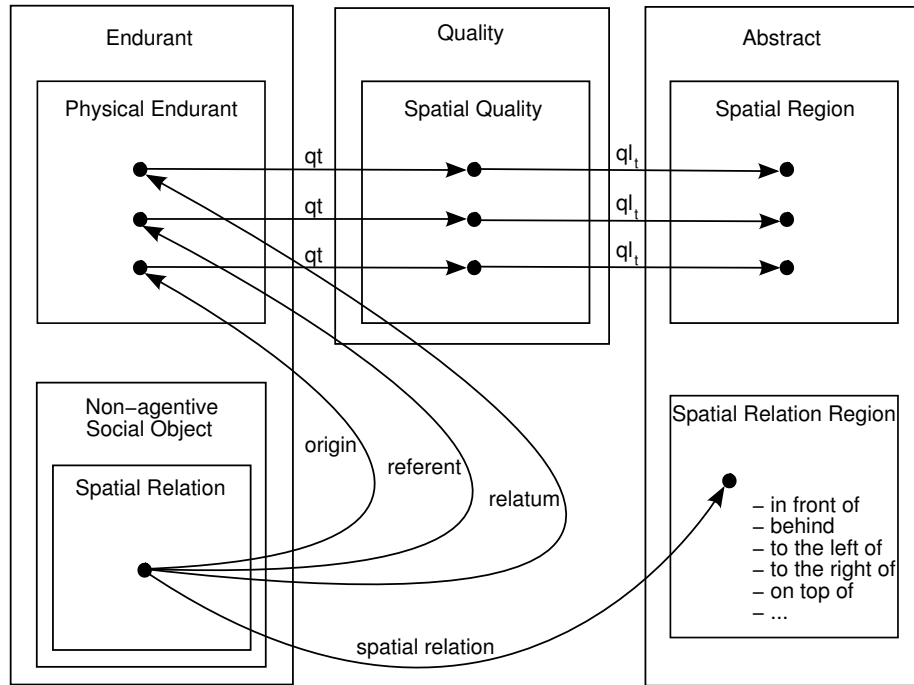


Fig. 3. Spatial relations

Starting from this origin, we calculate the relative position of objects and integrate them into the domain ontology. The calculation is based on the values measured by the ORCC system, that is the orientation and the distance according to the laser. The assignment of the linguistic terms of spatial relations is oriented by *heterogeneous non-overlapping acceptance areas*, introduced in [10]. The system assigns appropriate linguistic spatial terms by calculating the angles between two objects (referent and relatum) with respect to the laser sensor (origin). Term assignments in our application conform with the distribution of angles and their linguistic projective terms (left, right, in front of, behind) that is described in [18], regarding only relative references. This is taken at present for simplicity, as it will be for further investigations to show how we can best integrate the representation of diverse spatial calculi in DOLCE and how this can be related to linguistic expressions of spatial relations.⁸

Generally, our domain ontology has the following major advantages that contribute to the system: 1. It represents all relevant information in our system about object representation and spatial relations shared by the dialog and the

⁸ Our approach is to extend the notion of the axiomatization of simple spatial regions involving single physical objects in terms of geometric spaces. Our abstract spatial relation quality regions, for example, correspond more to axiomatizations of spatial calculi involving orientation, such as Freksa's Doublecross calculus [7].

classification component. 2. Spatial relations that are expressed in natural language can be directly accessed due to the ontological representation of spatial relations. 3. As it is developed on the basis of a fundamental ontology, it shall easily be combined with ontologies guided by the same fundamental basis. 4. This structure also eases extending spatial relations for other reference systems

Although the classification of objects is not directly influenced through this domain ontology, the quality of the dialog can be improved, as additional information about possible constellations of objects in an office scene also forms part of our domain ontology. This information is helpful for the user dialog, as described in the next section, but not vital. To infer new knowledge from the ontology, the system is using the tableaux reasoner Pellet⁹.

4 Linguistic Dialog

With the current status of the system's dialog, we are offering rather basic user interactions, as our aim is not yet to develop sophisticated natural language dialogs that provides multiple variants for similar meanings. We will demonstrate the representation of, and referencing to, physical objects and their spatial relations represented in the domain ontology on the basis of the dialogs instead. In particular, our linguistic component is part of the realization of the cognitive vision claims, mentioned above. An improvement of this component is considered below (see section 6).

Generally we divide our human-computer interaction with the system into two phases: In the first phase (the *training phase*), after the scene has just been scanned, the system generates the ontological instances of the objects as far as they are identified by the image recognition system. If an object was not identified, or was identified as several types, the system tries to identify the object by consulting the user. Thus, objects that were classified unambiguously by the ORCC system are directly instantiated into the domain ontology together with their spatial relations to surrounding objects. Objects with ambiguous classifications have to be determined by the user during the training phase.

To be as non-intrusive as possible, the system analyzes the ambiguously classified objects to see whether they are most likely to be of a specific type by ontological reasoning. For some concepts, we can assume that objects of the same kind happen to be close to each other, like books, bottles, or mugs. However, it is crucial to emphasize that we do not restrict the concepts in this way. A bottle does not necessarily need to be nearby other bottles. We just note that they can be usually grouped together. In case of an unclassified object that is surrounded by three objects of one type, the system asks the user if the unclassified object (which is marked in the image during this dialog) is also of that type (or, if not, what else). This technique does not bother the user with many different and unlikely suggestions, but only the most probable one. Hence, the domain ontology already facilitates a simpler, more controlled style of interaction and

⁹ <http://www.mindswap.org/2003/pellet>

it improves the convenience of the system. A similar assumption is that certain kinds of objects are also often close to each other, like a mouse is usually near to a computer, a keyboard, or a laptop, and not near other mice, which gives us a similar reasoning strategy as just described. Such modeled background knowledge is usually related to the functionality of objects. The probability value of an object type that is calculated by the ORCC system is also a hint for prioritizing the list of possible types. But we do not omit types that have very low priority values as these may still prove to be relevant later. In either case, users are always free to define object types as they like.

In addition to each request of the system about possible object types, the referring object is marked in the image. Hence, the user can determine the respective type directly looking at the image. The system will consult the user for clarification only in case an object was classified ambiguously. After the type has been determined, an instance of the object is integrated into the domain ontology, including spatial relations to surrounding objects, as described in the previous section. In Fig. 5 an example of a training phase dialog is shown.

In the second phase (the *action phase*), the user can either request specific types of objects by referencing spatial relations to other objects, or can request one or several objects, that have a certain relation to a specifically described object. Although the latter request is just a “show me”-operation, this can be easily extended to other, more application-relevant instructions, such as “give me”, “take”, or “do something with”.

Our system can handle only a few linguistic queries at present, such as “Show me the object that is to the right of the mug!” or “What is the object behind the mug?”. After syntactically processing the user’s request, the ontological knowledge about relative positions between different objects takes effect. Depending on the spatial relations that the user expressed in the request, the system tries to derive the correct event. Currently linguistic expressions, such as “to the right/left of”, “on (top of)”, “in front of”, and “behind”, are supported. If the user, for example, asks for “all objects” that have a certain position relative to a specific object (the *relatum*), all possibilities are looked up transitively in the domain ontology. In case the request is about a specific object with an indication of its relative position, the ontological instantiation of this type and its spatial relations are explored. An example of an action phase dialog is shown in Fig. 6.

As we have already mentioned, this human-user interaction should be seen primarily as a first approach toward one possible application of the combination between image recognition, linguistic references and ontological modeling of spatial relations. There are certainly other fields of application which include functional aspects of objects, like assigning a task to a robot to perform certain actions depending on the objects’ functionalities, or different interaction scenarios, such as generating a description of the scene. This will be investigated in future developments. The results of our current experiment serve mainly as a basis for investigating the representation of spatial relations, formally as well as linguistically, which are evaluated in the next section.

5 First Results

To show first results of our system’s processing sequence, an example scene is analyzed in this section.

Table 1. Result of the ORCC analysis of the suggested object types for each object (the bold types are the correct object types)

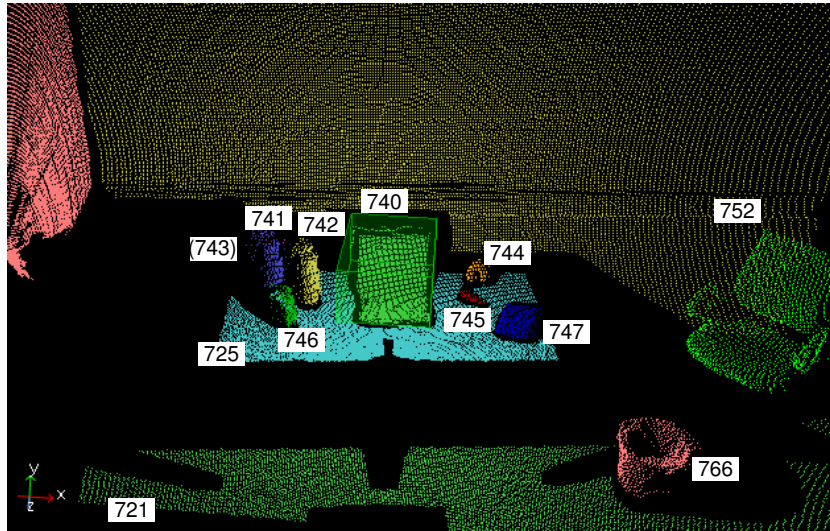
ID	Classification & Probability		
721	floor (0.99)		
725	tabletop (0.96)		
740	laptop (0.58)	bin (0.15)	briefcase (<0.01)
741	bottle (<0.01)		
742	bottle (0.06)		
743	coffeepot (0.61)	bottle (<0.01)	<i>Segmentation failure</i>
744	mug (0.71)	holepuncher (0.50)	stapler (0.09)
745	mug (0.33)	stapler (0.16)	holepuncher (0.09)
746	bottle (0.23)	coffeepot (0.02)	
747	book (0.61)		
752	chair (0.15)	bin (<0.01)	stapler (<0.01)
766	laptop (0.25)	bin (0.18)	briefcase (<0.01)

Fig. 4 shows a 3D scene (4(a)) in which a part of an office with its typical objects is illustrated¹⁰, along with the corresponding 2D pictures (4(b) and 4(c)). The position of the laser indicates the viewpoint toward the scene. Initially the system has detected six objects unambiguously and correctly: The floor (ID: 721), a table (ID: 725), two bottles (ID: 741, 742), a book (ID: 747), and a chair (ID: 752). Hence, the respective objects and their corresponding spatial relations are integrated into the domain ontology relating to this scene. Six remaining objects have been detected by the ORCC system, which finally have to be determined by the user. The resulting dialog is shown in Fig. 5.

5.1 Example of a Training Phase Dialog

In this dialog, the system asks the user about the remaining objects. It suggests the different types that the visual component has classified for each object. The results of this classification are shown in Table 1. Additionally to each question the respective object is marked in the scene. Beginning with the first question (“Is the marked object (id:740) a/an laptop, or briefcase, or something else?”, the marking of the object is shown in Fig. 4(a)), the object with ID 740 is ambiguously classified as a laptop, a bin, or a briefcase (see Table 1). One of the additional kinds of information about possible spatial constellations of objects, mentioned above, is that laptops are usually on a table and bins on a floor. This

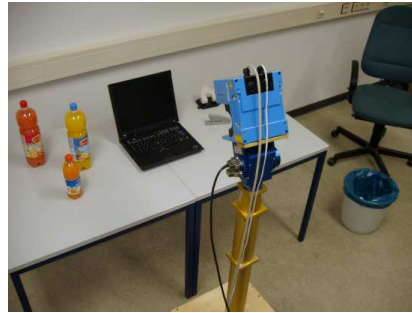
¹⁰ An ID of each object is inserted for a comprehensible reference in the text.



(a) 3D



(b) 2D



(c) 2D

Fig. 4. Example of a scanned scene (3D and 2D)

relation is represented in our domain ontology as an optional attribute: Objects of type laptop are related to an object of type table by the spatial relation *on (top of)* while objects of type bin are related to an object of type floor. As we mentioned above, these are not strict conditions for being a laptop or a bin, but optional ones, and they are considered during the training phase.

For this reason, the system seeks for and detects the Spatial Relation with ID 740 as referent that has the relation *on (top of)* and analyzes the object type of the corresponding relatum. Such “on (top of)” relations are based on the results from the ORCC system, which calculates the vertical alignment of objects on the basis of density segmentation (see Section 2.1). If the relatum is a table, the system will conclude that the object is more likely a laptop and if the relatum is a floor, it will conclude that the object is a bin. As a result of this analysis,

The answer to the question about the objects behind the stapler lists the hole puncher, which also matches with the image in Fig. 4(a). Depending on the frame of reference and the acceptance areas that we introduced in section 3, the system specifies that the bin is the only object in front of the chair and the table is to the left of the chair from the user’s viewpoint, which also seems intuitive in combination with Fig. 4(a). In the case that the system has to list all objects to the left of the laptop, it only refers to the objects that were instantiated during the training phase. Thus, it lists three bottles that actually exist in the image (Fig. 4(b)) without the additional object (Fig. 4(a), ID 743) that were segmented by the visual component. Again according to the 2D image in Fig. 4(b), the system lists the hole puncher, the book, and the stapler as the objects that are to the right of the laptop.

Another result from the use of acceptance areas and the relative frame of reference is that there are no objects behind the book and the stapler is the object to the left of the book according to the system. It is important in this case to remember the viewpoint of the user, from which the relative positions are calculated. It may also be usual for the user to refer to the stapler or the hole puncher as being behind the book in respect of their alignments on the table. The planned integration of additional frames of reference, such as the intrinsic reference required here, will incorporate such possibilities.

Finally the question about all objects that are on the floor gets the reply from the system containing the bin, the table, and the chair, which again matches correctly with Fig. 4(b).

6 Conclusions and Future Work

In this paper, we have argued that visual systems can be supported by linguistic user dialogs as well as a domain ontology. We have presented a cognitive visual system that analyzes a real world 3D image. A first approach toward the representation of spatial relations between physical objects in an ontological structure was introduced and we have discussed the resulting human-user interaction dialogs.

Further research aspects include the incorporation of representations of spatial calculi, for instance, RCC-8 [4] or Doublecross [7]. Also the computation of responding to user requests during the action phase that ask only for *one* object has to be refined as it currently only considers distance and relation but not proximity to spatial relation axes. The differences in the linguistic expressions of the resulting spatial relations will be examined and evaluated empirically for their suitability for differing tasks. The representation of spatial relations in ontologies still remains a crucial concern. Further considerations have to be made with respect to current efforts in ontological engineering. The interrelation between representing spatial relations depending on spatial calculi in ontologies and spatial relations depending on linguistic expressions must also be made clearer. To this end, we believe that the combination described here of functionally interpreted perceptual visual input with a formal domain ontology offers an

excellent foundation for subsequent research on the perception, grounding, and use of spatial descriptions in interactions between artificial agents and humans.

One of our next foci is on including further linguistic frames of reference. Linguistic terms according to intrinsic references seem to be very promising and will be integrated into the system, especially in connection with a detection of objects' intrinsic orientation within the ORCC system. We will also consider *internal relations*, such as “the object in the back corner of the room” — in contrast to *external relations* described in intrinsic, relative, and absolute reference systems. Such extensions are also essential to clarify user request ambiguities. The possibility of different viewpoints from the laser will also be important. We plan to extend this in the long term to multi-robot scenarios.

An extensive evaluation of our system is also to be done. This depends mostly on the configuration of larger data records that we have to prepare. Moreover, we are planning to carry out experiments with our system and humans to investigate the system's performance and quality. The possibilities of the user interaction will also be extended in order to support this evaluation as required.

Acknowledgements

The Collaborative Research Center for Spatial Cognition (Sonderforschungsbereich Transregio SFB/TR8) of the Universität Bremen and the Universität Freiburg is funded by the Deutsche Forschungsgemeinschaft (DFG), whose support we gratefully acknowledge.

This work has been especially supported by the SFB/TR8 subprojects A2-[ThreeDSpace] and I1-[OntoSpace].

References

1. Drago Anguelov, Ben Taskar, Vasco Chatalbashev, Daphne Koller, Dinkar Gupta, Jeremy Heitz, and Andrew Y. Ng. Discriminative learning of markov random fields for segmentation of 3D range data. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego, California, June 2005.
2. John A. Bateman and Scott Farrar. Spatial Ontology Baseline. SFB/TR8 internal report I1-[OntoSpace] D2, Collaborative Research Center for Spatial Cognition, University of Bremen, University of Freiburg, Germany, 2004.
3. Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, May 2001.
4. Anthony G. Cohn, Brandon Bennett, John Gooday, and Nicholas Mark Gotts. Qualitative spatial representation and reasoning with the region connection calculus. *GeoInformatica*, 1(3):275–316, 1997.
5. Kenny R. Coventry, Richard Carmichael, and Simon C. Garrod. Spatial prepositions, object-specific function and task requirements. *Journal of Semantics*, 11:289–309, 1994.
6. Peter Auer et al. A Research Roadmap of Cognitive Vision, ECVision: European Network for Research in Cognitive Vision Systems. http://www.eucognition.org/ecvision/research_planning/ECVisionRoadmapv5.0.pdf, (19.04.2005).

7. Christian Freksa. Using orientation information for qualitative spatial reasoning. In Andrew U. Frank, Irene Campari, and Ubaldo Formentini, editors, *Spatio-Temporal Reasoning*, pages 162–178, 1992.
8. Asunción Gómez-Pérez, Mariano Fernández-López, and Oscar Corcho. *Ontological Engineering with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer-Verlag, London, 2004.
9. Peter Gärdenfors. *Conceptual Spaces: The Geometry of Thought*. A Bradford Book. MIT Press, 2000.
10. Daniel Hernández. *Qualitative Representation of Spatial Knowledge*, volume 804 of *Lecture Notes in Computer Science*. Springer Verlag, Berlin, 1994.
11. Annette Herskovits. *Language and Spatial Cognition: an interdisciplinary study of the prepositions in English*. Studies in Natural Language Processing. Cambridge University Press, London, 1986.
12. Joana Hois, Kerstin Schill, and John A. Bateman. Integrating Uncertain Knowledge in a Domain Ontology for Room Concept Classifications. In Max Bramer, Frans Coenen, and Andrew Tuson, editors, *The Twenty-sixth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, Research and Development in Intelligent Systems, Cambridge, UK, December 2006. Springer-Verlag. to appear.
13. Björn Krebs, M. Burkhardt, and Friedrich M. Wahl. Integration of Multiple Feature Detection by a Bayesian Net for 3D Object Recognition. In *Mustererkennung*, pages 143–150, 1998.
14. Stephen C. Levinson. *Space in Language and Cognition*. Cambridge University Press, Cambridge, 2003.
15. Claudio Masolo, Stefano Borgo, Aldo Gangemi, Nicola Guarino, and Alessandro Oltramari. Ontologies library (final). WonderWeb Deliverable D18, ISTC-CNR, Padova, Italy, December 2003.
16. Pascal Matsakis, Jim Keller, Laurent Wendling, Jonathan Marjamaa, and Ozy Sjahputera. Linguistic description of relative positions in images. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, 4(32):573–588, August 2001.
17. Andreas Meyer. Merkmals- und formbasierte 3D-Objekterkennung für Büroszenen. Diplomarbeit, Universität Bremen, 2005.
18. Reinhard Moratz and Thora Tenbrink. Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations. *Spatial Cognition and Computation*, 6(1):63–106, 2006.
19. Reinhard Moratz, Thora Tenbrink, John A. Bateman, and Kerstin Fischer. Spatial knowledge representation for human-robot interaction. In Christian Freksa, Wilfried Brauer, Christopher Habel, and Karl Friedrich Wender, editors, *Spatial Cognition III: Routes and Navigation, Human Memory and Learning, Spatial Representation and Spatial Reasoning*, volume 2685 of *Lecture Notes in Artificial Intelligence*, pages 263–286, Berlin, 2003. Springer.
20. Hans-Hellmut Nagel. Steps toward a Cognitive Vision System. *AI Magazine*, 25(2):31–50, 2004.
21. Hans-Hellmut Nagel. Cognitive Vision Systems (CogViSys). http://cogvisys.iaks.uni-karlsruhe.de/homepage_CogViSys_V3B.html, (31.08.2001).
22. Deb Roy, Peter Gorniak, Niloy Mukherjee, and Josh Juster. A Trainable Spoken Language Understanding System For Visual Object Selection. In *International Conference of Spoken Language Processing*, 2002.
23. Ioannis Stamos and Peter K. Allen. 3-D Model Construction using Range and Image Data. In *Proceedings of the IEEE Conference on Computer Vision and*

Pattern Recognition (CVPR-00), pages 531–536, Los Alamitos, June 13–15 2000. IEEE.

24. Thora Tenbrink. Semantics and Application of Spatial Dimensional Terms in English and German. SFB/TR8 internal report I1-[OntoSpace], Collaborative Research Center for Spatial Cognition, University of Bremen, University of Freiburg, Germany, 2005.
25. Michael Wüstel and Thomas Röfer. A Probabilistic Approach for Object Recognition in a Real 3-D Office Environment. In *WSCG'2006 Posters Proceedings*, 2006.