

Qualitative Spatial Arrangements and Natural Object Categories as a Link between 3D-Perception and Speech

Reinhard Moratz, Michael Wünnel, Robert Ross

Transregional Collaborative Research Center "Spatial Cognition"

Universität Bremen, FB 03

Postfach 330 440, D-28334 Bremen, Germany

{moratz, wuenstel, robertr}@informatik.uni-bremen.de

Overview

The visionary goal of an easy to use service robot implies intuitive styles of interaction between humans and robots. Such natural interaction can only be achieved if means are found to bridge the gap between the forms of object perception and spatial knowledge maintained by such robots, and the forms of language, used by humans, to communicate such knowledge. Part of bridging this gap, is allowing user and robot to establish joint reference on objects in the environment - without forcing the user to use unnatural means for object reference.

Our approach to establishing joint object reference makes use of natural object classification, and a computational model of basic intrinsic and relative reference systems. The system, utilizing $2\frac{1}{2}$ D laser range data, assigns natural category (e.g. "door", "chair", "table") to new objects based on their functional design. The recognizer - based on the concepts of affordances, form and function - identifies certain geometries that lead to certain functions, and therefore allow their identification [2]. With basic objects within the environment classified, we can then make use of a computational reference model, to process natural projective relations (e.g. "the briefcase to the left of the chair"), allowing users to reference objects which cannot be classified reliably by the recognition system alone.

In the current version, we mainly focus on the concept of the *supporting plane*. When the function of an object part is to support a potential other object, this part has to be parallel to the ground. A full three-dimensional segmentation based approach is not necessary when additional clues like object arrangement information is given by the user. In the future, we will augment the system with more refined 3D reconstruction abilities. The approach performs best for objects having strong functional constraints at the system's current perceptual granularity (e.g. desks, tables, chairs). However, smaller objects on the ground (e.g. waste paper baskets, briefcases etc.) can be detected but not classified reliably by our current system. These objects can however be referred to by a human and furthermore they can be referred to with reference to other objects in the environment (e.g. "the bin behind the table").

A projection of the recognized 3D objects onto the plane produces a 2D map, defined in terms of object location for directed and undirected objects, object categorization (if available), and camera position and angle. This map, is used as input for our

reference processing module. Our model of projective relations (e.g. "left", "right", "in front of", "behind") uses a reference axis which is a directed line through the center of the object used as relatum (e.g. the robot itself, the group of objects, or other salient objects) [1]. If the robot itself is the relatum then the reference direction is given by its view direction (which normally corresponds to the symmetry axis of the robot). Otherwise the directed line from robot to the center of the relatum serves as reference axis. The partitioning into sectors of equal size is a sensible model for the directions "left", "right", "front" and "back" relative to the relatum. However, this representation only applies if the robot serves as both relatum and origin. If a salient object or the group is employed as the relatum, front and back are exchanged, relative to the reference direction. The result is a qualitative distinction, as formally specified in [1].

As mentioned, this model was developed with a modest visual recognition system. However, since our new, 3D, object recognition system, is capable of detecting objects like chairs which have an intrinsic reference frame, we wish to account for intrinsic reference cases within our model. For example, "In front of the chair" is the direction into which a human would look if he sat on this chair. For such a case we can take the intrinsic reference model which we used for the robot itself. The difference is that a chair seen from a different point of view induces a "front" and a "back" acceptance area but typically no "left" or "right" area. However, we did not systematically test this intuition with human test subjects yet.

In our initial system demonstrator, users interact with the system by verbally issuing simple requests to the system. These requests - to identify items in the system's perceptual range - are detected with a Nuance Speech Recognizer¹, before being fed to a semantic analysis component. This analysis attempts to identify the category of object to be identified, the referent object, and the relationship used by the user to relate the referent object to the target object. The reference processing module then attempts to identify the target object in the 2D map using the projective relations defined. The most probable target object once computed, is then highlighted. For images of the perceived scenes and the corresponding results of the system see <http://www.sfbtr8.uni-bremen.de/A2>

In future work, our vision system is to be augmented with a light camera to combine the two- and three-dimensional recognition methods, thus allowing for a wider range of objects which can be perceived. With this new resolution capability, we will also be expanding our qualitative reference model, examining - amongst other things - differences between spatial models appropriate for English, as well as German speakers.

References

1. R. Moratz, K. Fischer, and T. Tenbrink. Cognitive Modeling of Spatial Reference for Human-Robot Interaction. *International Journal on Artificial Intelligence Tools*, 10(4):589–611, 2001.
2. Michael Wünnstel and Reinhard Moratz. Automatic object recognition within an office environment. In *Canadian Conference on Computer and Robot Vision (CRV2004)*, 2004.

¹ We gratefully thank Nuance Communications Inc. (www.nuance.com) for the use of their systems.