

# Multi-Modal Scene Interpretation

Michael Wünstel, Thomas Röfer

The visionary goal of developing an easy to use service robot implies several key tasks such as speech understanding, object recognition and scene understanding. Besides the more sensor-oriented capabilities such systems need extensive meta knowledge, e.g., about mental representations of spatial relations to match the view between man and machine. Only if all parts fit together an unrestricted man machine communication can be established. Therefore a cognitive system has to address many different parts that have to be integrated, in the technical sense and especially in the cognition models [1]. Especially when connecting a perceptive component with a spatial reasoning component using a speech recognition and synthesis component, the probabilistic area of object recognition has to be coupled with the logical area of formal reasoning. The cognitive vision system *ORCC* presented here combines diverse recognition strategies that afford an extensive description of an unreserved scene: In a first step the room demarcations and structurally simple objects such as tables are extracted using as well functional as structural properties. Then further objects are segmented based on their position, followed by a structurally more complex and a more shape-oriented recognition step. Then, this spatial information is enriched with colour-based information about the objects. Afterwards, the resulting scene description can be used as an input for a speech-based man-machine dialogue about the objects within in the scene [2].

## 1 Introduction

A central requirement for future autonomous systems is the capability to verbally and interactively communicate with humans, and thus to be able to verbally and interactively learn from errors that they encounter while acting in their environment [3]. Perceptive systems in particular are not really designable robustly unless they are specialized in a specific domain. Thus especially perceptive systems are in need of a fast, i.e. verbal way of interaction.

Cognitive Vision as a field of Computer Vision also deals with the integration of the perceptive component in a holistic cognitive system. Besides the usage of cognitively motivated methods for object recognition itself the application of these methods in a cognitive system are also taken into account. This in particular implies the necessity of a gradual rating of different alternatives in the recognition process. The perceptive system *ORCC* presented here uses cognitively motivated methods for recognition that are partially independent of the kind of the sensor. The system provides a complete functional and textual scene description that also takes alternative interpretations into account for the subsequent cognitive processing within a speech-interaction module. Thus the difficult but indispensable process of man-machine interaction that itself is a prerequisite for the man-machine learning process is represented within an indoor scene. For this task methods for spatial calculi from the field of Spatial Cognition in an ontology based realization are used. These methods however will not be treated in detail in this article.

## 2 System

There is a set of different models dealing with the process of object recognition in humans. One way of distinguishing these different models is their classification into four base models of perceptual cognition [4]: The recognition by colour or simple geometric features, the recognition by the structural arrangement, recognition by comparison with a sim-

ple memory image of the object, and finally the recognition through different memorized views of the object. Concerning the underlying cognition approaches two different methods can be distinguished [5]: On the one hand symbolically oriented systems, in which a cognitive model that is interpretable by humans serves as a basis, and on the other hand emergent or connectionist systems that evolve closely interlocked with the existing sensors (embodiment) and ultimately do not possess any direct formally describable internal representation but in which the process of cognition corresponds to a kind of unconscious inference. A recognition method that correlates with recognition by structural parts is the concept of affordances. Gibson [1] assumes that we closely associate the objects in our surroundings with their possible functional applicableness. In his interpretation Gibson does not use any internal representations. This functional aspect plays also a major role in the general bottom-up concept "the Society of minds" by Minsky [6] that incorporates "both symbolic and connectionist notions" [7].

The functional approach allows the recognition of objects even if their typical structure is no longer available or detectable. For example that can be the case with chairs. On the other hand the functional aspect can be subordinated to the design, and thus e.g. the arrangement of objects be superordinated over their functional properties. Besides this form and function based aspects the texture plays a further important role, especially in domestic environments. Textures can e.g. serve as a supplement of a three-dimensional structurally detected object or can provide 2-D objects such as textual descriptions. Thus a cognitive system has to be able to detect 3-D depth information as well as 2-D colour information and register both to each other. In addition such a system has to be capable to detect specific textures, e.g. markers or text.

The advantage of the here presented *ORCC*-system is the combination of a 2-D colour intensity camera and a 3-D laser range scanner. Thus the in a way more powerful 3-D data can be used for central segmentation and classification tasks and