From: L.C. Jain (ed.), Knowledge-based Intelligent Information Engineering Systems, Proceedings of KES'99 at Adelaide, Australia. 1999, pp. 423-426. ©IEEE.

An Integrated Object Representation for Recognition and Grasping

Efthimia Kefalea, Eric Maël, and Rolf P. Würtz

Institut für Neuroinformatik, Ruhr-Universität Bochum,
D–44780 Bochum, Germany
http://www.neuroinformatik.ruhr-uni-bochum.de/ini/VDM/

Abstract

As a step towards systems that can acquire knowledge automatically we have designed a system that can learn new objects with a minimum of user interaction and implemented it on our robot platform GripSee *[1]. A novel object is placed into the robot's gripper in order to define a default orientation and a default grip. The robot then places the object on a turning table and builds up a visual representation that consists of a collection of graphs, labeled with multiscale edges. A user interface that can correct errors in the representation is also part of the system. The visual representation is complemented by a grip library, which contains possible ways of grasping and manipulating the object in a robust manner. We regard this procedure as an example of* Human Assisted Learning.

1 Introduction

One of the major problems in knowledge representation is the lack of systems that can actively acquire information from the environment, in order to assess the results of reasoning processes and extend the knowledge base. Sometimes this is called the *symbol grounding problem.* For a system to be of practical use, a certain degree of control must be left to the user. What is required may be called *semi-autonomy:* a system (typically a robot) that can interact with the environment must dispose of a repertoire of skills that are carried out autonomously, but the actual *control* of behavior must be left to a human operator [1].

The knowledge dealt with here is not sophisticated high-level knowledge about complicate interactions between things in the real world, because there is currently very little chance to acquire such knowledge without extensive programming. Instead we concentrate on modeling simple knowledge acquisition tasks like learning to recognize a formerly unknown object and to grasp and manipulate it. More concretely, we describe an object representation, which is suited to support the behavior of manipulating objects and learning new objects with a minimum of user interaction.

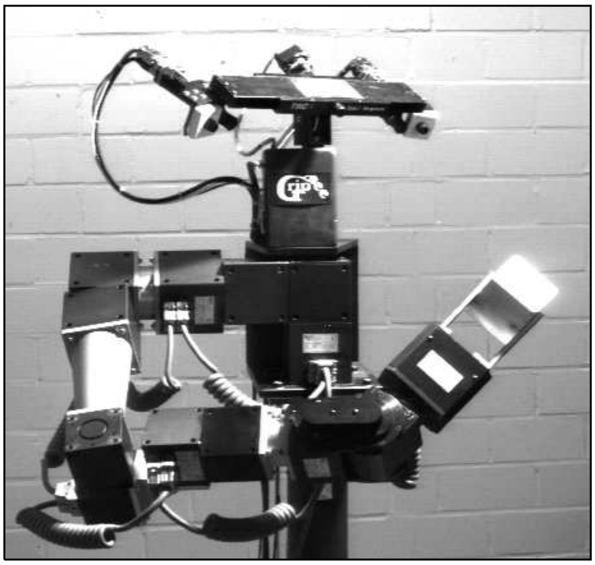

Figure 1: GripSee's 3 DoF stereo camera head and 7 DoF manipulator.

We are aware of the fact, that the use of "representations" in artificial intelligence and robotics has been heavily criticized for good reasons [2]. Indeed, highly detailed world models are difficult to obtain, and it is not sure if they really would make concrete vision tasks easier. However, enough information must be stored in the robot itself for it to be able to remember objects, situations, and suitable actions. It is this sort of memory, which we call a representation here, and we require very strongly the *learnability* of a representation, i.e. the possibility for the robot to construct it from raw data with as little human interference as possible.

The general principle underlying our approach to knowledge representation is the *association* of actions to situations encountered before. Because no situation is ever identically encountered a second time a certain degree of generalization is required in

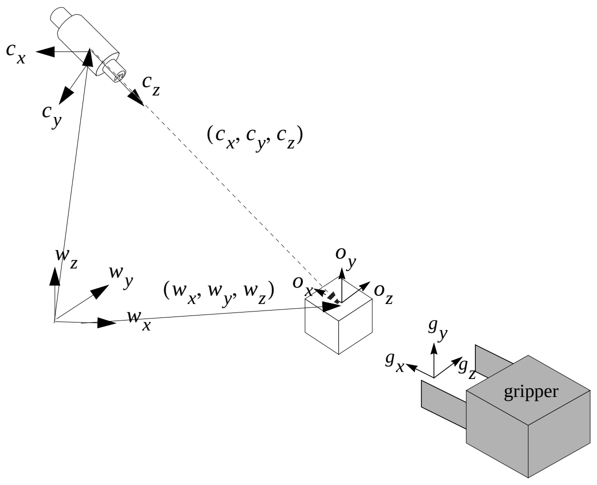

Figure 2: World, camera, object and gripper coordinate systems

the analysis of the situation (recognition process) as well as in the parameters guiding the action.

The paper is organized following the standard distinction between data structures and algorithms: First, we describe the representation of a single object, then the procedure employed to learn a new object, and finally the procedures that are used to recognize and grasp a known object.

2 The object representation

An object representation suited for visually guided grasping has to integrate 2D visual features and 3D grip information about a known object, in order to apply a known grip when the situation requires it. According to our general philosophy, autonomous learning of the representation is highly desirable, therefore complicated constructs like CAD models are not considered. Rather, we adopt the view that visual recognition and application of a grip is mainly a recollection of what has been seen or done before, with the necessary slight modifications to adapt to the situation at hand.

2.1 Hardware constraints

GripSee is an anthropomorphic robot system developed at our institute as a research platform and demonstrator for a coming generation of service robots. It is equipped with a redundant manipulator with a parallel jaw gripper featured with tactile sensors and an active stereo camera head (see figure 1). The stereo camera head is calibrated autonomously to the predefined manipulator kinematics [6]. The robot's proportions resemble a sitting human, which makes it suited for exploring and grasping objects in a table scenario using information from its visual and tactile sensors [1].

2.2 Grip representation

In contrast to the very sophisticated information required for grasping with a multifingered hand, a grip for our parallel jaw gripper can simply be represented by a homogeneous transformation matrix, which stands for the gripper's position and orientation in object coordinates (see figure 2), plus the opening width of the gripper suitable for the object (see [5] for details). In the course of the full integration of our tactile sensors, additional descriptors will be the grasping force and a desired contact distribution.

2.3 Visual representation

The visual representation is view-based, i.e., for each different orientation of the object a set of visual features is stored, which are extracted from the left and right stereo images and grouped into a *model graph*, which preserves the topological relationships between the features. These model graphs are stored in a library for different objects with different orientations and are used to recognize known objects with a graph matching process, which is invariant under translation and scale.

The situation is made more difficult by the fact that one and the same object can appear on the table not only in different locations and orientations around the vertical axis, but can also rest on different sides, which potentially changes the visual appearance completely. At the current state of our system, such different stable positions are treated as independent versions of the object and each of them is stored and learned separately.

2.4 Integrated representation

A complete object representation consists of a grip library, which contains position and orientation of suitable grips in object coordinates, and for each stable pose of the object on the table a set of model graphs, which cover a complete 360° set of rotations in a reasonable resolution (9°). Each model graph is associated with a *rotation matrix*, which relates object and camera orientations and with the *offset* of the projection of the *grasping center* (the point between the gripper tips during a successful grip) to the center of gravity of the graph nodes in the image plane.

3 Learning the representation

Learning of a new object (or more precisely, one stable pose of a new object) is initiated by putting the object onto the table and having the robot create the various views by moving the object around. This procedure has two serious difficulties. First, a good grip must already be known for the robot to manipulate the object in a predictable manner. Second, the actual orientation of the object should

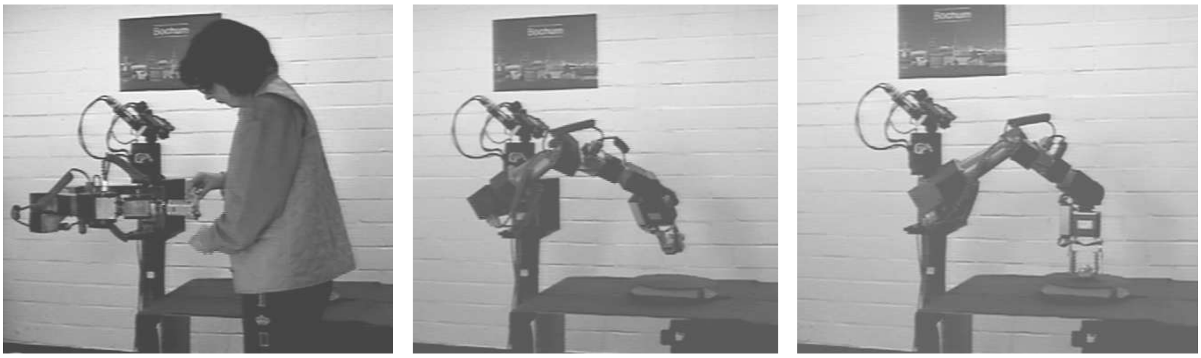

Figure 3: Initialization of the learning process. A human operator shows a simple grip, which is then used by the robot to place the object onto a turntable and collect the views required for recognition.

be known with good precision, because the error is likely to accumulate over the various views. The problem of learning new grips from scratch can only be solved by relying on tactile information. We have currently constructed and implemented tactile sensors on our gripper, but the extraction of detailed object information is subject of future research.

In this situation, we have decided to solve both problems by what we call *human-assisted learning*. The general idea is that the acquisition of knowledge is as autonomous as possible, but a human operator still makes decisions about what is important and thus guides the process. Concretely, in the current case our learning procedure is as follows. The operator presents the novel object by putting it into the gripper (which has a defined position and orientation at that moment), in a position and orientation that are ideal for grasping. They thus define both a default grip and the object coordinate system. The robot closes the gripper, puts the object onto the center of a turning table, *fixates* on the grasping center, and takes a stereo image pair of the first object view. Then, the turning table is rotated by a specified increment (9°) and a second view is taken. This is repeated until a full circle of object views is acquired. After acquisition, all images are rotated around their center to compensate for the rotation associated with a combination of tilt and vergence.

After the images of the views are taken, they are converted into a collection of *labeled graphs*. In this learning step, it is assumed that the background is uniform (which is the case for the surface of the turning table) in order to avoid the necessity of complicated segmentation methods and to assure as clean graphs as possible in the representation (see [3] for details).

As we are mainly interested in simple objects, the first step is *contour extraction*. Input images are preprocessed in two different steps, namely the calculation of *Mallat multiscale edges*, and a subsequent contour following step, which assigns confidence values to edges, starting from the modulus values of the Mallat transform and modifying them according to coherent edge information in the neighborhood.

A graph is then constructed beginning with *homogeneous sampling* of the image. A square lattice of points with a spacing of several pixels is generated. Graph nodes are positioned on these image points. Each node is connected to all its neighbors resulting in a maximum of eight neighbors for each node of the graph. A *thinning step*, which discards all nodes whose Mallat responses are below a threshold, cuts the graph down into regions which actually contain contours. The resulting graph is, in general, no longer connected, because there may be significant contours in the background. Therefore, all graphs except the one with the maximal number of nodes are eliminated in a *separation* step.

After the preceding steps all remaining nodes are still located on the original square lattice and lie on or directly neighboring to lines of local modulus maxima. Now, by *local adaptation* each node moves to the position of the closest modulus maximum. This leads to a contour-adapted graph with nodes positioned on edges only. This local shift can result in neighboring nodes lying on the same image position. Those multiple nodes are eliminated by a final *simplification step*.

Two alternatives to using the turntable may be considered, namely having the robot place the object onto the table in all necessary orientations or holding it in the gripper while storing the visual information. Both have not been pursued so far in order to minimize the mechanical strain on the hardware. The second possibility, poses the additional, problem of segmenting and subtracting the gripper itself from the image of the object. Obviously, the quality of the object recognition depends critically on the quality of the model graphs. Our system therefore contains the possibility to modify the node positions by hand, but under good illumination the automatically created graphs are usually good enough.

4 Using the representation

4.1 Object recognition

The recognition step yields the identity, position and orientation of an object on the table. In the presence of multiple objects, one of them must be selected by either operator interaction [1] or some attention control mechanism. After the region of interest is defined the camera head fixates on the center of this region (see above).

After that step the recognition proper proceeds by *Elastic Graph Matching* [4]. Its main goal is to establish correspondences between model nodes and points in the actual image, so that feature similarities are only evaluated at corresponding points. To this end, Mallat feature vectors are calculated in the pair of stereo images to be analyzed. Then a distorted copy of a model graph in the stereo image is optimized to fulfill the following (possibly conflicting) constraints:

- The local image information stored at each node position of the model graph must match the image region around the position where the node is positioned in the neighboring view.
- The distances between the matched node positions should not differ too much from the original distances.
- The matches of one model graph in both stereo images must approximately adhere to the *epipolar constraint.*

This matching procedure is carried out for all available model graphs and the best matching graph is assumed to represent the correct object in the correct orientation. The position can then be retrieved from the location of the matching nodes.

4.2 Grasping

After identity, position, and orientation of the object are known, a suitable grip is selected from the library and transformed from object coordinates to world coordinates using the rotation matrix and offset belonging to the best matching view. The selection is either based on a command from the operator (e.g., in the form of a hand gesture) or on a kinematical optimization criterion. Then an arm trajectory is planned and executed and the object is picked up for further manipulation.

5 Conclusions

The representation scheme we have introduced is well suited for *rigid* everyday objects, but not for objects of flexible shape such as wires or cables. Furthermore, the gripper can grasp objects which do not exceed certain dimensions, which is a constraint applying also to humans.

The major problem for the system as described here is the lack of feedback during the grasping process. Therefore, the gripper has now been equipped with tactile sensors. As a part of ongoing research we work on refining the estimates of the object's position and orientation produced by the recognition module using tactile information during the application of the grip. Another line of research consists of learning new grips by *imitating* the trajectories presented by a human operator. Finally, the representation will be extended by a detailed description of the tactile impressions encountered during grasping, which may allow the measurement and subsequent use of quite detailed 3D-information.

Acknowledgments

This work has been funded by grants from the DFG (Graduiertenkolleg KOGNET) and the German Federal Ministry for Education and Research (01 IN 504 E9). Programs were developed in the FLAVOR environment. We wish to thank Abdelkader Mechrouki, Mark Becker, Michael Neef, Peer Schmidt, and Jochen Triesch for help with many of the problems encountered during the development.

References

[1] M. Becker, E. Kefalea, E. Maël, C. von der Malsburg, M. Pagel, J. Triesch, J. C. Vorbrüggen, R. P. Würtz, and S. Zadel. GripSee: A gesture-controlled robot for object perception and manipulation. *Autonomous Robots*, 6(2), 1999. In press.

[2] R. A. Brooks. Intelligence without representation. *Artificial Intelligence Journal*, 47:139–160, 1991.

[3] E. Kefalea. *Flexible Object Recognition for a Grasping Robot.* PhD thesis, University of Bonn, Germany, 1999. In preparation.

[4] M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Würtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3):300–311, 1993.

[5] E. Maël. *Adaptive and Flexible Robotics for Visual and Tactile Grasping.* PhD thesis, Ruhr-Universität Bochum, 1999. In preparation.

[6] M. Pagel, E. Maël, and C. von der Malsburg. Self calibration of the fixation movement of a stereo camera head. *Machine Learning*, 31:169–186, 1998.