Work in Progress on a Vision–Based Robotic System: Visual–Haptic Attention and Accumulation of Object Representations

Norbert Krüger, Daniel Wendorff, Gerald Sommer
Lehrstuhl für kognitive Systeme, Institut für Informatik,
Christian–Albrechts–Universität zu Kiel
Preusserstrasse 1-9, 24105 Kiel, Germany
nkr{dw,gs}@ks.informatik.uni-kiel.de

Abstract

In this paper, two modules of a behavior based robotic–vision system are described: A visual and haptic attention mechanism, and an accumulation algorithm to extract stable object representations within a perception–action cycle.

1 Introduction

The aim of our research is the design and implementation of an active vision system coupled with a robot arm (see figure 1a) which is able to recognise and grasp objects with autonomously learned representations. The system shall gain robot control over new objects (i.e., grasp a new object in a scene) by an instinctive and rudimentary behavior pattern and use the control over the object to accumulate a representation of the object and finally apply these representations to robustly track, grasp and recognise the object in a complex scene. Here we describe two modules of such a system and we give an overview about current and future research.

The design of our system is guided by a behavior based paradigm (see [2, 11]) in a dual sense. Firstly, to perform a certain action we may only need to extract a minimum of information (e.g., to fixate and zoom we do not need any shape information in our system). This is *perception for action*. Secondly, by active intervention we can make tasks easier for perception (e.g., in our system fixating and zooming potentially facilitates grasping or, as another example, robot control over the object helps for the extraction of object representations). This is *action for perception*. Usually both aspects — perception for action and action for perception — occur together in a so called *perception–action cycle* (PAC) [5, 12], i.e., perception and action support each other and depend on each other permanently.

We think that a complex vision–based system can not start to learn without some kind of prior knowledge [3, 6]. It can neither be a fully predetermined system, because the world within it operates is too complex that algorithms which solve difficult tasks could be formalised explicitly. Nor can it be a fully undetermined structure because the space of possible algorithms to be explored is much too large. Therefore, a certain amount of a priori knowledge has to be built in a complex vision system to guide learning. We think that an important part of this knowledge are basic competences (as introduced here), necessary to start a bootstrapping process in which more complex competences can be established.

In this paper, two modules of such a system are described: A visual and haptic attention mechanism, and an accumulation algorithm to extract stable object representations within a PAC. In the first module (described in section 2) the system directs its attention to new objects and manipulates the active components (i.e., cameras and grasper) such that a situation is achieved in which grasping becomes easier: grasper and object appear in the centre of a zoomed

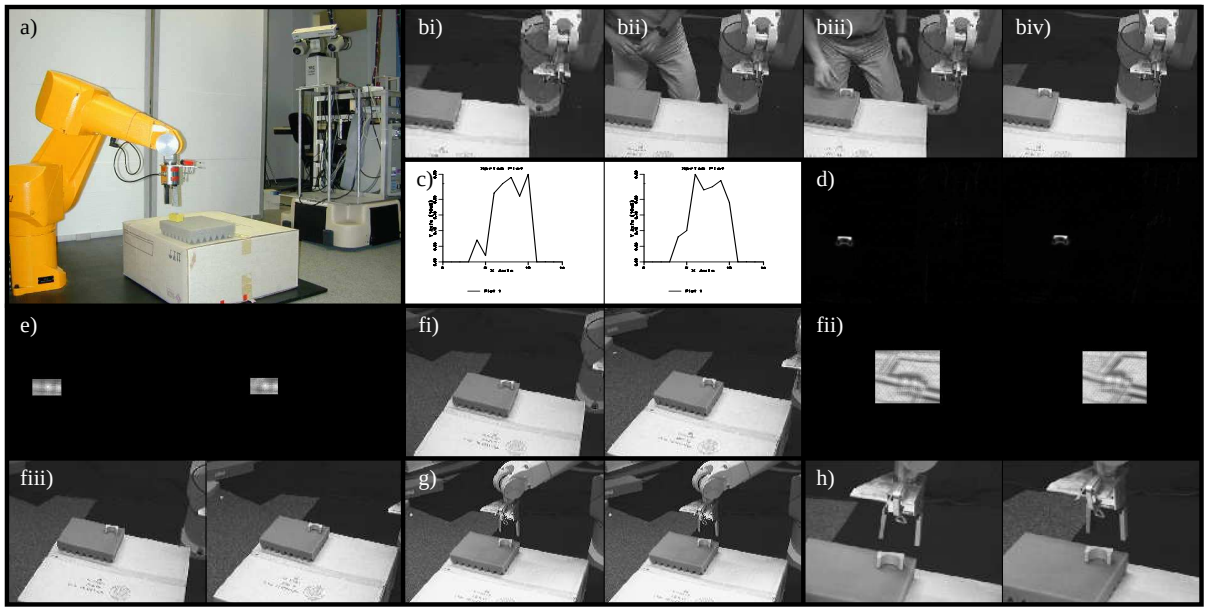

Figure 1: a) Active binocular head with robot arm. bi-biv) Images of a person entering the scene, putting an object into the scene and leaving the scene. c) Graph indicating a dynamic period by the magnitude of differences between images. d-h) Stereo images: d) Difference image before and after the dynamic period e) Similarities of a Gabor jet extracted from the centre of gravity in the left image to the jets extracted from other pixel positions of the difference area. Maxima are defined as corresponding points. fi) Fixation of the new object. fii) Similarities of the Gabor jets for fine tuning of fixation. fiii) Fixation after a second camera action. g) Movement of the robot arm to a position near the object. h) Zoom.

stereo image pair (see figure 1h). In this situation grasping of the object can be performed using only relative positions between grasper and object. The high resolution allows to accurately extract 3D–Information about the relative position and orientation of grasper and object by stereo. Our attention module is combined by a number of more primitive competences such as detection of a new object, fixation, recognition of the object under controlled conditions, movement of the grasper, zoom etc. Note that our attention mechanism is planned not to be only vision–based. We are currently redeveloping a haptic sensor [10] which allows to explore an object haptically. Therefore, our attention mechanism potentially focuses visual *and* haptic attention to the new object. The visual–haptic attention mechanism is to a wide degree predetermined but also contains adaptable components: The grasper is permantly tracked by the system. The information of motor commands and tracking results allows a self–calibration during the perception–action cycle.

The second module (roughly sketched in section 3) uses control over the object to extract a stable representation. We account for the vagueness of semantic information extracted from single images by assigning confidences to this information and accumulating this information over an image sequence of a controlled moving object. Although the information extracted from single images contains errors (see the representations on the left hand side of figure 2) a more stable representation can be achieved by combining information from different images (see right hand side of figure 2). Because the object can change its position and orientation — and this change might be wanted because another view of the object gives new information which might

not be extractable from former ones — we face the correspondence problem: Correspondences between entities describing the object in different images (or 3D interpretations extracted from stereo images) are not known. However, the parameters of motion are known since the robot manipulates the object and the transformations of entities can be compensated for each frame of the sequence. Knowing the correspondences, an algorithm can be applied to update and improve the object representation iteratively within a PAC.

In its current state, only these two modules are fully implemented. However, in section 4 we give a short overview about our current and future research aiming at a complete system. One important aspect of the design of a complex behavior based vision system is the interaction of modules developed by different people within one software package to derive complex competences from the combination of more primitive competences. We are currently developing a C++–library (KiViGraP, **Ki**eler **Vi**sion and **Gras**ping **P**roject) in which this interaction is going to occur (for details see [9]).

2 Achieving Tactile Contact by Vision–based Perception–Action Cycles

Our basic behavior to gain tactile contact with a new object can be divided into a number of more simple competences (described below). The behavior pattern can be understood to a wide degree as a reflex action: The system "aims" to get in contact to new objects to explore them visually and haptically. Going even further, it "aims" to grasp the object using a rudimentary representation to learn a more sophisticated and efficient representation (see section 3). During robot actions a permanent tracking of the grasper allows to permanently recalibrate the system.

The module described in this section is going to initiate a situation in which grasping and tactile exploration is facilitated. Since for the accumulation scheme (section 3) it is essential that the system has physical control over the object, the module described in this section can be understood as part of a bootstrapping process, that (once the system's experience has been grown) can be substituted by or transformed into a more goal–oriented behavior pattern. However, the bridge between achieving tactile contact and grasping has not yet been built and is part of current research.

In the following we describe some submodules used to achieve tactile contact. The modules described here are not understood to be performed in a sequential process but as competences which interact with each other (e.g., tracking and self–calibration) and which can be applied and mixed in a goal oriented manner. It is likely that at the very beginning of the bootstrapping process the structure and relations of the competencies are more predetermined than after a period of adaptation.

- **Detection of a new object and detection of a suitable time interval for robot action:** A new object is detected by the difference in each of the two stereo images before and after a dynamic period, i.e., a period in which people or other objects enter the scene (see figure 1bi–iv). For reasons of grasping success and maintaining safety for people interacting with the robot, it is necessary not to intervene in a dynamic situation. The system searches for a new object when a dynamic period occured — a person puts a new object into the scene — followed by a stable period — the person leaves the scene (see figure 1bi–iv). Figure 1c shows a graph indicating the dynamic in a scene. During a period in which the graph shows high values the robot is not allowed to intervene. The behavior pattern, responsible for robot and people safety can be understood as a permanent (self)protection expert which restricts all other robot processes.

- **Fixation, approaching and zooming:** In case of detection of a change in the images

before and after the dynamic period we fixate the new object. The internal camera parameters of our binocular camera-head are calibrated at an initialisation stage. Then the system recalibrates itself after a movement by computing the new projection parameters from the motion commands given to the camera head. This recalibration is relatively stable even after a number of movements.

The two areas which represent differences in the image (or more precisely their centre of gravity) give us two corresponding points for which we can compute a 3D–position with our calibrated system. Knowing its 3D–position we could easily fixate the object and then doing the 3D–estimation. However, since the correspondence of two objects is defined by the centre of gravity of areas (which might not be very precise), the system may additionally use information about similarities within a small area around our difference areas. We compare image patches (with a method similar to [8] based on Gabor wavelets and jets) to find more precise correspondences in the two stereo images (see figure 1e). The system can achieve a higher robustness by iterativeley computing the distance of the object and the image center after fixation. Note that these distances also can be used as a measure for the performance of the system, i.e., can also be used in a more global feedback loop to optimize the system.

Finally, the robot arm is moved to a position near the computed 3D–position of the object (see figure 1g) and the system can perform a zoom to get a higher resolution of both, the object and the grasper (see figure 1h).

- **Recognizing grasper and object and performing a second move (not fully implemented):** After achieving higher resolution, we can analyse the relative position of object and grasper (see figure 1fi–iii). In case that the system was not able to achieve tactile contact (which can now be checked visually and haptically), the system can repeat moves of the grasper to approach the object. Note that after fixating and zooming we have reduced the matching problem (finding grasper and object) significantly. Furthermore, since we can manipulate the grasper, it is not necessary to search for an arbitrary aspect of the grasper but its current 2D–aspect can be actively controlled.

  In a further step the object has to be grasped. Since our grasper allows to determine the success of grasping by measuring the width of the jaws after grasping, a repetition of grasping can be performed in case of non–success. Furthermore, this measure of success can be used as feedback on a more global learning level.

- **Tracking and self–calibration:** The system is equipped with a permanent grasper–tracking mechanism which is also based on the jet–representation in [8]. The 2D–tracking results and the motion parameters given to the robot can be compared to recalibrate the system by a simple update rule. It seems to be important that calibration does not only occur at the beginning of a process (often with an artificial calibration pattern) but is performed permanently during the normal perception–action cycle. Therefore, we have to face the tracking of the grasper in our quite uncontrolled environment. This is known as a very hard matching task which we are able to solve even with our rudimentary object representation by allowing only 'sure' matches to be used for self–calibration (for details see [14]). Here again, the system's ability to measure the success of performing competences is of significant importance.

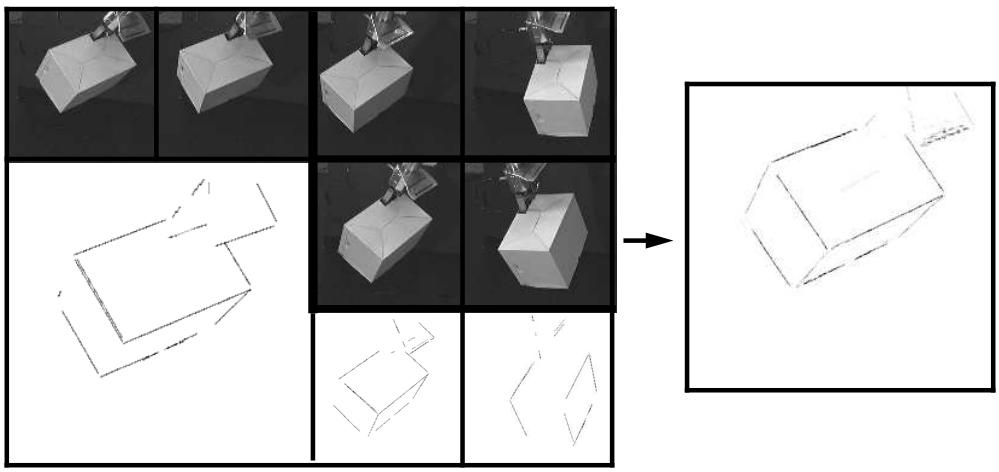

Figure 2: **left)** top: left and right image of an object. bottom: the projected 3D representation extracted from the stereo images. **middle)** Two pairs of stereo images (top: left camera image, middle: right camera image) and the the projected 3D representation (bottom). **right)** Projected 3D Representation accumulated over a set of stereo images. The system's confidence for the presence of line segments is represented as grey value (Dark values represent high confidences).

3 Accumulation of Inaccurate Information to a Robust Object Representation

After grasping the object, an accumulation scheme can be applied to extract a representation of the object. Our accumulation algorithm can be defined independently of the entities used to represent objects. The algorithm also is independent of the concrete equivalence relation or transformation used to define correspondences. It only requires an object representation by certain entities for which a metric is defined and to which certain transformations or equivalence relations (here a rigid body motion) can be applied. The basic idea of the scheme is that for each entity a confidence is updated. This confidence increases when correspondences under the known transformation can be found in many frames. This accumulation algorithm is an extension of an algorithm introduced in [7] which has only dealt with 2D representation and translational motion. Figure 2 shows the application of this scheme to representations consisting of 3D line–segments extracted from stereo images. For these entities the change of the transformation and a metric can be computed explicitly (for details see [1]).

In this scheme, an entity (here, a 3D line segment) is regarded to be existent only if it has accumulated confidence over time, or more precise, it is understood as an invariant entity in the time–space continuum under the equivalence relation 'rigid body motion'. Therefore an interpretation (as a 3D–line segment) is grounded in its change under the controlled movement of the object: the entity '3D–line segment' establishes itself only if it has been reconfirmed within the perception–action cycle. Our ansatz is therefore related to the so called symbol grounding problem (see [4]), i.e., to the problem to assign meaning to abstract entities. Here 'meaning' can be interpreted as an observable and foreseeable change under a self–performed motion.

4 Ongoing and future research

We have introduced two basic competences of an object recognition and manipulation system. In both modules perception and action are tightly intertwined within perception–action cycles

[5, 12].

Important components of such a system are still missing, such as performing grasping of the object after the visual and haptic attention mechanism. However, for such a grasp the attention mechanism gives a good starting point, because we have only to operate with relative positions and since we gained high resolution of the important aspects of the scene by active control of the camera. Furthermore, we intend to also use haptic information for performing the grasp. A further important problem is the application of our extracted representations to recognition and grasping tasks. As a first step, we could successfully apply an accumulated representation to the tracking problem using the pose estimation algorithm in [13].

We argue that gaining control over the object by grasping is a helpful prerequisite for the extraction of object representations. Correspondences, necessary for accumulation, can be computed since the system has control over the object. Furthermore, the system can decide by itself when the accumulation process shall stop, i.e. when a satisfactorial representation has been achieved. Furthermore, it can move the object in a position in which accumulation becomes easier (for example with homogeneous background).

We find the design of a vision–based robot system in which basic competences (such as introduced here) interact with each other to derive more complex behavior patterns is a challenging and demanding perspective. It desires the integration of different disciplines such as robotics, computer vision, signal processing and statistical learning as well as the integration of software developed by different people. Finally, the success of such a system should be measured empirically.

References

[1] M. Ackermann. Akkumulieren von Objektrepräsentationen im Wahrnehmungs–Handlungs Zyklus. *Christian–Albrechts Universität zu Kiel, Institut für Informatik und praktische Mathematik (Diplomarbeit)*, 2000.

[2] R.A. Brooks. Intelligence without reason. *International Joint Conference on Artificial Intelligence*, pages 569–595, 1991.

[3] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58, 1995.

[4] S. Harnad. The symbol grounding problem. *Physica*, D(42):335–346, 1990.

[5] J.J Koenderink. Wechsler's vision: An essay review of computational vision by Harry Wechsler. *Ecological Psychology*, 4:121—128, 1992.

[6] N. Krüger. *Visual Learning with a priori Constraints (Phd Thesis)*. Shaker Verlag, Germany, 1998.

[7] N. Krüger and G. Peters. Orassyll: Object recognition with autonomously learned and sparse symbolic representations based on metrically organized local line detectors. *Computer Vision and Image Understanding*, 77, 2000.

[8] 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 dynamik link architecture. *IEEE Transactions on Computers*, 42(3):300–311, 1993.

[9] KiViGraP (Homepage of the Kieler Vision and Grasping Project). http://www.ks.informatik.uni-kiel.de/~kivi/kivi.html.

[10] Peer Schmidt. Entwicklung und Aufbau von taktiler Sensorik für eine Roboterhand. *Institut für Neuroinformatik Bochum (Internal Report)*, 2000.

[11] G. Sommer. Verhaltensbasierter Entwurf technischer visueller Systeme. *Künstliche Intelligenz*, 3:42–45, 1995.

[12] G. Sommer. Algebraic aspects of designing behaviour based systems. In G. Sommer and J.J. Koenderink, editors, *Algebraic Frames for the Perception and Action Cycle*, pages 1–28. Springer Verlag, 1997.

[13] G. Sommer, B. Rosenhahn, and Y. Zang. Pose estimation using geometric constraints. Technischer Bericht (Institut für Informatik und Praktische Mathematik, Christian–Albrechts–Universität zu Kiel), 2000.

[14] D. Wendorff. *Diplomarbeit (in progress)*.