Two Modules of a Vision–Based Robotic System: 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: An 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.

In this paper, two modules of such a system are described: A visual and (potentially) haptic attention mechanism, and an accumulation algorithm to extract stable object representations. 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 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. Note that our attention mechanism is planned not to be only vision–based. We are currently redeveloping a haptic sensor [17] which allows to explore an object haptically. Therefore, our attention mechanism potentially focuses visual *and* haptic attention to the new object. The 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 allow a self–calibration during the perception–action cycle [10, 18].

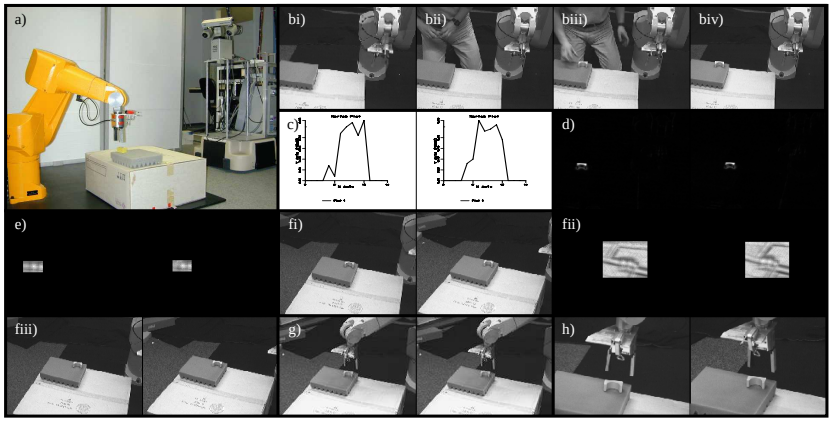

Fig. 1. a) Active binocular head with robot arm. bi-biv) Images of a person entering the scene, putting an object into the scene and leaving the scene. c) Graph indicating a dynamic period by the magnitude of differences between images. d-h) Stereo images: d) Difference image before and after the dynamic 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.

The second module (described 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 3) a more stable representation can be achieved by combining information from different images (see right hand side of figure 3). 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 perception–action–cycle.

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 [14]).

2 Attention mechanism based on visual–robotic Perception–Action Cycles

Our basic behavior aimin at a 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 shall "aim" 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 attention 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 depending on the actual system's goal. 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.
  In case that the person puts a new object into the scene, the object is detected by the difference in the images before and after the dynamic period (see

figure 1d). Since simple differences of grey–level images are unstable due to little movements of the camera or variation of illumination, we also compute the difference of the magnitude of Gabor wavelet responses. For efficient computation we make use of the separability of quaternionic Gabor wavelets [5].

- **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. 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 [13] 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). Object and grasper appear magnified and their relative distance can be used for grasper manipulation with high accuracy. It is expected that this relative distance can be extracted with higher accuracy than absolute distances from stereo images.

- **Tracking and self–calibration:** The system is equipped with a permanent grasper–tracking mechanism which is also based on the jet–representation in [13]. 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 [19]). Here again, the system's ability to measure the success of performing competences is of significant importance.

We would like to finish this section with the remark that, although in its current state the behavior pattern is to a huge degree predetermined, we do intend to achieve a more robust, more flexible and more–goal oriented behavior pattern in a complex system through learning. Self–calibration by grasper tracking already supports an even better estimate of internal and external parameters and therefore a more robust behavior. Furthermore, the module can measure its success of fixation (figure 1fi-iii) and is intended to detect the success of tactile contact and grasping. Therefore, this information can be used as feedback for a more global learning which may allow to achieve direct contact and successful grasping more frequently by optimizing free parameters of the system. Finally, after achieving robot control more complex object representations can be learned (see section 3) and the original reflex behavior can be transformed into a more goal–oriented behavior, e.g., an object is only grasped when it hasn't been learned so far.

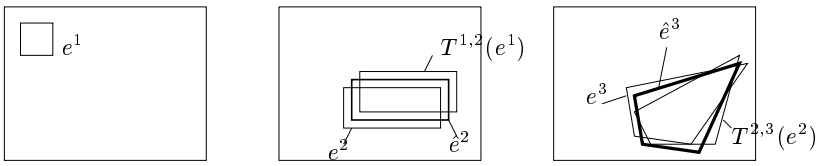

Fig. 2. The accumulation scheme. The entity e^1 (here represented as a square) is transformed to $T^{1,2}(e^1)$. Note that without this transformation it is nearly impossible to find a correspondence between the entities e^1 and e^2 because the entities show significant differences in appearance and position. Here a correspondence between $T^{1,2}(e^1)$ and e^2 is found because a similar square can be found close to $T^{1,2}(e^1)$ and both entities are merged to the entity $\hat{e}^2$. The confidence assigned to $\hat{e}^2$ is set to a higher value than the confidence assigned to e^1 indicated by the width of the lines of the square. The same procedure is then applied for the next frame for which again a correspondence has been found. By this scheme information can be accumulated to achieve robust representations.

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 (see figures 2 and 3). Feature extraction faces the problem that semantic information extracted by artificial systems from a single image or stereo images even under optimal conditions is necessarily imperfect. For instance, although there exist a large amount of edge detectors none of them is comparable to human performance. Moreover, we see it as an important problem to extract object representations in *real* situations and not in artificially adapted conditions (such as homogeneous background, controlled pose etc.), i.e., we intend to fullfill the requirements *situatedness* formulated by Brooks [3]. One

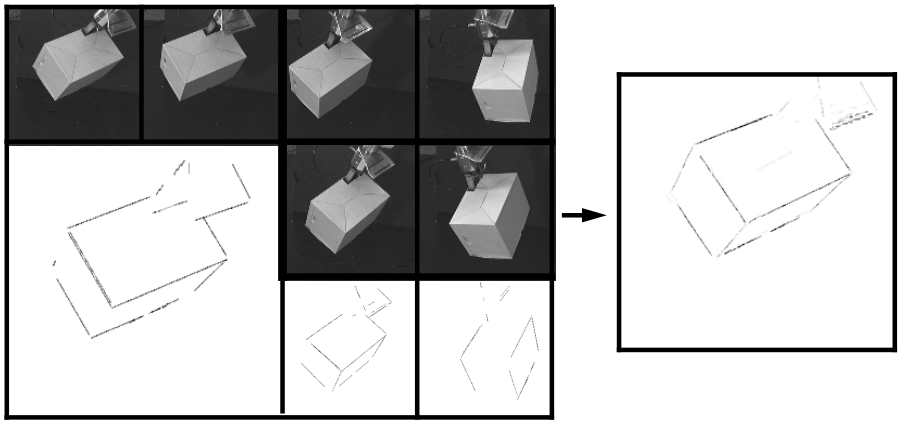

Fig. 3. left) top: left and right image of an object. bottom: the projected 3D representation extracted from the stereo images. **middle)** Two pairs of stereo images (top: left camera image, middle: right camera image) and the the projected 3D representation (bottom). **right)** Projected 3D Representation 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).

important reason for the extremely good performance of humans on these tasks in even very difficult situations is that the human visual system applies *constraints* to interpret a certain scene or situation [7, 11]. An important constraint is the utilization of the coherence of objects during a rigid body motion which allows to accumulate information over time. Furthermore, in an active vision–based robot system we are able, instead of only passively perceiving a certain situation , to support learning by our own actions. This corresponds to *embodyment* as another requirement formulated by Brooks [3].

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 (such as rigid body motion) can be applied. This accumulation algorithm is an extension of an algorithm introduced in [12, 15] which has only dealt with 2D representation and translational motion.

Let $e \in E$ be an entity used to describe objects (for instance a 2D–line segment, a structure tensor [9] extracted from an image, 3D–line segments extracted from a stereo image pair or any other kind of object descriptor) and $d(e, e')$ be a distance measure on the space of entities E. Furthermore, let T be a transformation or equivalence relation, for instance a rigid body motion or the projective map corresponding to a rigid body motion. If e^i is an entity extracted from frame i of a sequence of events then $T^{i,i+1}(e^i)$ is the transformation $T^{i,i+1}$ from the i–th to the $(i+1)$–th frame applied to e^i.

Let e^{i+1} be an entity extracted from the $(i+1)$–th frame of the sequence. We say that e^i and e^{i+1} are likely to correspond to each other if $d(T(e^i), e^{i+1})$

is small. Often it might not be possible to find an exact correspondence with $d(T(e^i), e^{i+1}) = 0$. For example, if we want to compare local image patches in two images knowing the exact projective transformation corresponding to the rigid body motion of an object from the first to the second frame, the corresponding image patches can not be expected to be exactly equal because of factors such as noise during the image acquisition, changing illumination, non–Lambertian surfaces or discretization errors, i.e., the features are quasi–invariant. The problem may even become more severe when we extract more complex entities such as 3D or 2D line segments or 3D–surface patches. Therefore it is advantageous to formalize a confidence of correspondence by using a metric.

The accumulation of information can now simply be achieved by the following update rule: If there exists an entity e^{i+1} in the $(i+1)$–th frame for which $d(T(e^i), e^{i+1})$ is small (i.e., a correspondence is likely), then merge $T(e^i)$ and e^{i+1} by some kind of average operator, $\hat{e}^{i+1} = \text{merge}(T(e^i), e^{i+1})$, and set the confidence for $\hat{e}^{i+1}$ to a higher value than the confidence assigned to e^i. If there exists no entity e^{i+1} in the $(i+1)$–th frame for which $d(T(e^i), e^{i+1})$ is small, the confidence for entity e^i to be part of the object is decreased. In Figure 2 a schematic representation of the algorithm is shown for two iterations.

The accumulation scheme could also be interpreted as an iterative clustering scheme with an in build equivalence relation to compensate the motion of the object. It is also related to, so called 'dynamic neural nets' [6, 4], in which cells appear or vanish according to some kind of confidence measure.

Figure 3 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 (i.e., $T^{i,i+1}(e)$) and a metric can be computed explicitly (for details see [1]). Up to now, only one aspect of an object can be accumulated because correspondences are needed which are not granted when occlusion does occur. That means, that when the robot rotates the object by a larger degree, it is likely that new edges occur in the stereo images and other edges disappear. In the current state we ensure that the same aspect is presented to the system by only allowing movements within a small subspace of the space of rigid body motions. To define such a subspace of possible rigid–body motions we make explicitly use of the metric defined on the space of unit–quaternions corresponding to rotations in Euclidean space [2].

4 Outlook

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 [10, 18].

Important components of such a system are still missing, such as performing grasping of the object after the 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. A further important problem is the application of our extracted representations to recognition and

grasping tasks. In [16] we could successfully apply one of our accumulated representations to the tracking problem.

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. W. Blaschke. *Kinematik und Quaternionen.* VEB Deutscher Verlag der Wissenschaften, 1960.
3. R.A. Brooks. Intelligence without reason. *International Joint Conference on Artificial Intelligence*, pages 569–595, 1991.
4. J. Bruske and G. Sommer. Dynamic cell structure learns perfectly topology preserving map. *Neural Computation*, 7(4):845–865, 1995.
5. T. Bülow and G. Sommer. Quaternionic gabor filters for local structure classification. In A.K. Jain, S. Venkatesh, and B.C. Lovell, editors, *14th International Conference on Pattern Recognition, ICPR'98*, volume 1, pages 808–810. Brisbane, Australia, August 16-20, IEEE Computer Society, 1998.
6. B. Fritzke. Growing cell structures – a self-organizing network for unsupervised and supervised learning. *Neural Networks*, 7(9):1441–1460, 1994.
7. S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58, 1995.
8. M. Hansen. *Stereosehen - ein verhaltensbasierter Ansatz.* PhD thesis, Inst. f. Inf. u. Prakt. Math. der Christian-Albrechts-Universität Kiel, 1998.
9. B. Jähne, editor. *Digitale Bildverarbeitung.* Springer, 1997.
10. J.J Koenderink. Wechsler's vision: An essay review of computational vision by Harry Wechsler. *Ecological Psychology*, 4:121—128, 1992.
11. N. Krüger. *Visual Learning with a priori Constraints.* (PhD Thesis) Shaker Verlag, Germany, 1998.
12. 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.
13. 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.
14. KiViGraP (Homepage of the Kieler Vision and Grasping Project). http://www.ks.informatik.uni-kiel.de/~kivi/kivi.html.
15. M. Pötzsch, N. Krüger, and C. von der Malsburg. A procedure for automatic analysis of images and image sequences based on two–dimensional shape primitives. *U.S. Patent Application*, 1999.
16. B. Rosenhahn, N. Krüger, T. Rabsch, and G. Sommer. Automatic tracking with a novel pose estimation algorithm. *accepted for Robot Vision 2001*, 2001.
17. Peer Schmidt. Entwicklung und Aufbau von taktiler Sensorik für eine Roboterhand. *Institut für Neuroinformatik Bochum (Internal Report)*, 2000.
18. 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.
19. D. Wendorff. ¨ *Diplomarbeit (in progress)*, in progress.