Biologically Motivated Hand-Eye Coordination for the Autonomous Grasping of Unknown Objects *

Alexa Hauck, Georg Passig, Johanna Rüttinger, Michael Sorg, Georg Färber

Institute for Real-Time Computer Systems
Technische Universität München
{hauck,passig,sorg}@rcs.ei.tum.de

Abstract. In the field of visually guided grasping, humans still outshine their robotic counterparts with respect to accuracy, speed, robustness, and flexibility. We therefore examined current neuroscientific models for the control of human reach-to-grasp movements and, based on one of them, developed a novel visual motion control strategy. This control strategy was integrated into a complete hand-eye system, including modules for the determination of suitable 3D grasping positions on unknown objects from the images of a stereo camera system. The modules were implemented and tested on the experimental hand-eye system MINERVA.

1 Introduction

The ability to grasp arbitrary objects will be an important component of future autonomous service robots. Humans acquire this ability very early; they achieve a performance that still outshines that of their robotic counterparts with respect to accuracy, speed, robustness, and flexibility. In the past, robotic approaches mostly used precisely calibrated hand-eye systems and known (CAD) models of the objects to grasp in an open-loop motion control ("look-then-move"), e.g. in [2]. Especially the first prerequisite, the exact calibration, is problematic: errors in the internal system model directly affect the endpoint accuracy of the grasping movement.

To overcome this problem, a new approach was proposed: so-called *visual servoing* systems (see [8] for a tutorial) use a continuous feedback of visual information about the position of the end-effector or about its distance to the target. Such systems are robust against errors in the internal models or do not even need a metric calibration (see e.g. [7, 11, 13] for examples of visual servoing systems that actually grasp objects). One of the main disadvantages of visual servoing approaches is, however, that visual information is needed during the whole movement and at a high rate.

* The work presented in this paper was supported by the *Deutsche Forschungsgemeinschaft* as part of the Special Research Program "Sensorimotor - Analysis of Biological Systems, Modeling, and Medical-Technical Application" (SFB 462).

In contrast, the results of neuroscience show that human reach-to-grasp movements cannot be explained by either of the two approaches described above; they rather seem to be the result of a combination of both. The main point is that robustness against model errors is reached *without* requiring continuous visual feedback. We therefore examined current neuroscientific models for the control of human reach-to-grasp movements with special regard to the visual control strategy used, and extended one in such a way that it fulfills the requirements of a robotic system. This novel control strategy combines the two robotic approaches described above in such a way that visual information about the position of object or gripper can be integrated asynchronously during the movement. Thus, errors in the internal models can be compensated without the need for continuous, high rate visual feedback. The resulting motion control module is described in Sec. 2.

The second limitation of most visual servoing systems actually setting out to grasp objects is that they need exact models of the objects (e.g. [13]) or make implicit assumptions about the shape of the objects (e.g. [7]). In [11], a method is proposed to find grasping points on the silhouette of unknown objects for the quasi-planar case, i.e. with the limitation that the objects are flat, lying and are viewed from above. In contrast, we have developed a method to find 3D grasping points with the help of a stereo camera system. This method will be described in Sec. 3.

Fig. 1 shows a block diagram of the complete sensorimotor system: information about the environment or the system itself is extracted using sensors (in our case two CCD cameras), further processed and interpreted. The resulting, more abstract information is used to plan motor actions; those actions are executed by translating the plans into commands for the actors. These commands affect the system and the environment; the changes can be observed again by sensors. We have implemented the necessary modules and integrated them on the robotic hand-eye system MinERVA; experimental results are given in Sec. 4.

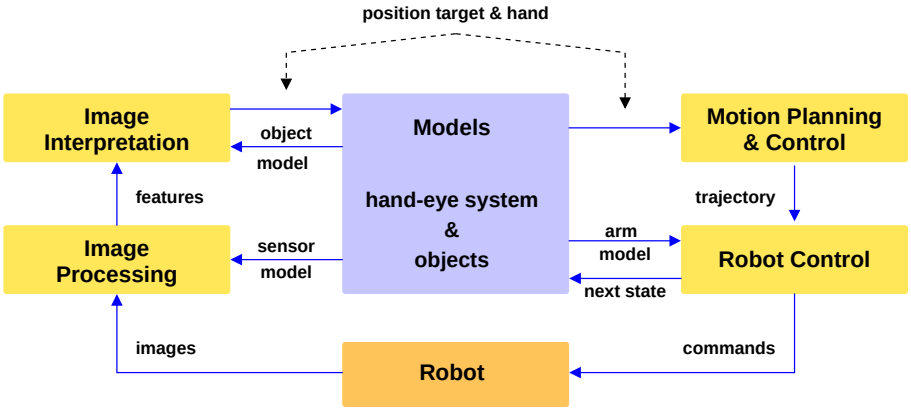


Fig. 1. Block diagram of the sensorimotor system.

2 Motion Planning and Control

Motion planning and control is an extensive field of research in robotics; the same applies to neuroscientific research on human motion control. In this section, we will limit ourselves to describing the biological background briefly in Sec. 2.1. Then, our new control strategy is presented (Sec. 2.2) and validated in simulation (Sec. 2.3). More information can be found in [6, 4].

2.1 Biological Background

It has been consistently reported that the invariant features of human multi-joint arm movements are (1) that the path of the hand is a roughly straight line in Cartesian coordinates and (2) that the profile of tangential (Cartesian) hand velocity is bell-shaped. Thus, one has to conclude that human reaching movements are planned in spatial coordinates[1], not in joint space. Another interesting phenomenon is that humans correct their movements very smoothly in the case that the target position changes during the movement ("double-step target") or if visual information is distorted, e.g. by prism glasses.

Concerning the control structure, a number of different models have been developed. As the problem of trajectory planning is heavily underconstrained, a popular approach steming from optimal control theory is to look for an objective function that human motion control might be optimizing. However, none of these models allows for the integration of information about the hand position, i.e. (visual) feedback[2].

Following a different line of reasoning, Goodman et al. [3] proposed a control strategy that, similarily to visual servo control, computes the current velocity from the remaining distance to the target $\mathbf{x}_{T1}$, as described by the differential equation

$$\dot{\mathbf{x}}(t) = \frac{1}{\tau_1} \cdot \dot{g}(t) \cdot (\mathbf{x}_{T1} - \mathbf{x}(t)), \qquad g(t) = t^3 \tag{1}$$

which can be solved to yield the path of the hand over time, resulting in a straight line in space. In the case of changes of target position at times t_i, corrective movements based on a similar differential equation are superimposed, resulting in the following path function:

$$\mathbf{x}(t \geq t_n) = \mathbf{x}_{Tn} - \sum_{i=1}^{n} \mathbf{D}_i \cdot e^{-\frac{1}{\tau_i} g(t-t_i)}, \qquad \mathbf{D}_i = \mathbf{x}_{Ti} - \mathbf{x}_{Ti-1} \tag{2}$$

In its original version, this model did not address the case of visual feedback, though. Because of its similarity to visual servoing structures, we set out to extend it suitably.

[1] The variable measured is the position of the hand; orientation is thought to be controlled by a process running in parallel.

[2] One of them does, but is difficult to realize on a robotic system.

2.2 New Control Strategy

The model of Goodman et al. has two drawbacks: First, superposition in the case of double-step targets was limited to the level of position control (Eq. 2), and secondly, the question of visual feedback was not addressed. In a first step, we therefore generalized the model in order to allow the control of velocity similarly to a visual servoing control law even in the case of double-step targets. After rearranging Eq. 2 and substituting part of it into its derived version (details in [4]), one arrives at the following velocity control law

$$\dot{\mathbf{x}}(t \geq t_n) = \frac{\dot{g}(t-t_n)}{\tau_n} \cdot (\mathbf{x}_{Tn} - \mathbf{x}(t)) + \sum_{i=1}^{n-1} \mathbf{D}_i \cdot e^{-\frac{1}{\tau_i} g(t-t_i)} \cdot \left(\frac{\dot{g}(t-t_i)}{\tau_i} - \frac{\dot{g}(t-t_n)}{\tau_n} \right) \tag{3}$$

which consists of a "feedback" term containing $\mathbf{x}(t)$, and a sum of corrective "feedforward" terms that decline exponentially with time.

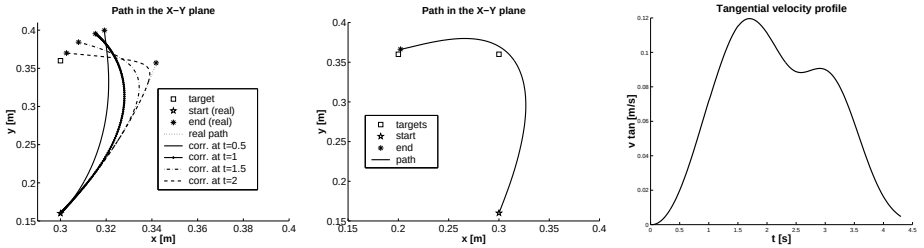
As as next step, we wanted to generalize the superposition scheme in order to include the case of sparse, *asynchronous* visual feedback. This means that visual information is not used continuously to compute $\mathbf{x}(t)$, but only integrated from time to time, while $\mathbf{x}(t)$ is computed from the generated commands, i.e. in a feedforward fashion. Due to the inevitable errors in the internal models, a visually measured hand position, $\mathbf{x}_v(t)$ will differ from the estimated one, $\mathbf{x}_e(t)$, resulting in an apparent "jump" of hand position. The main idea of our approach now is to treat this jump analogously to a target jump, by superimposing a corrective movement with the amplitude $\mathbf{D} = \mathbf{x}_e(t) - \mathbf{x}_v(t)$. Thus, a smooth adaptation of the motion to the new information is assured.

2.3 Simulation Results

We first tested the control strategy by simulating movements of a 2 d.o.f. robot, and on a simulation model of our hand-eye system MinERVA (see Sec. 4.1)[3]. Fig. 2 shows how errors in the kinematic model of the two-link robot can be compensated: Without correction, an error of 20% in the first link length, i.e. $L_1^* = 0.8L_1$, results in a endpoint error of about $4cm$ for a movement amplitude of $20cm$; a single corrective movement based on visual feedback can reduce this error significantly (Fig. 2a). Note, that visual feedback seems to get "more effective" when integrated near the end of the movement. Generally, the effectiveness depends on how similar the error measured at the current hand position is to the error measured at the target position (details in [4]).

Fig. 2b,c depict the path and the tangential velocity profile of a movement with two feedback corrections and a target jump, again with $L_1^* = 0.8L_1$. The movement remains smooth, and the error is almost completely compensated.

[3] Simulations were realized using *MATLAB* and *Simulink*.



a) b) c)

Fig. 2. Effect of corrective motion for a disturbed internal model ($L_1^* = 0.8L_1$): (a) Dependency on time of measurement. (b) Path and (c) tangential velocity profile with two measurements at $t = 1s$ and $t = 2s$ and a target jump at $t = 1.5s$.

To evaluate the performance of the new control strategy in more detail, we simulated the reaction to errors in all parameters of the geometric model of our hand-eye system MinERVA (transformations between arm, head, and cameras, intrinsic camera parameters, arm model). In the case of most parameters, errors can be corrected quite well with only one corrective movement[4]. As a sort of worst case, the intrinsic camera parameters were disturbed with a variance of 2.5%, all other parameters with a variance of 5%, resulting in a mean terminal error of $8.85cm$, with a maximum of $20cm$. Integrating visual feedback at a rate of only $1Hz$ reduces this error already to a mean of $4mm$ (maximum: $1cm$). For more details see [4].

3 Image Processing and Interpretation

The vision modules are to provide the motion control module with information about the position of target and hand. We therefore developed prototypical modules for determining the position of the gripper (Sec. 3.3) and of the target. The latter task was addressed in more detail, as the goal was to enable the robot to determine suitable grasping positions on unknown objects.

Without knowledge about the object to grasp and with only one view of it, there is only one place to look for grasping points: its silhouette, or *apparent contour*[5]. There already exist methods to determine grasping positions on the silhouette, heuristic (e.g. [11]) and analytical ones (e.g. [12]), but they all operate on images from a single camera, and therefore need additional context knowledge to be applicable to 3D tasks.

[4] Exceptions are e.g. focal length and the horizontal pixel size, as they affect the 3D reconstruction algorithm in a non-linear way.

[5] Another advantage of the silhouette is that it is a global feature useful for object recognition as well. In fact, parts of the algorithms presented in the following were originally developed for the task of object recognition [1].

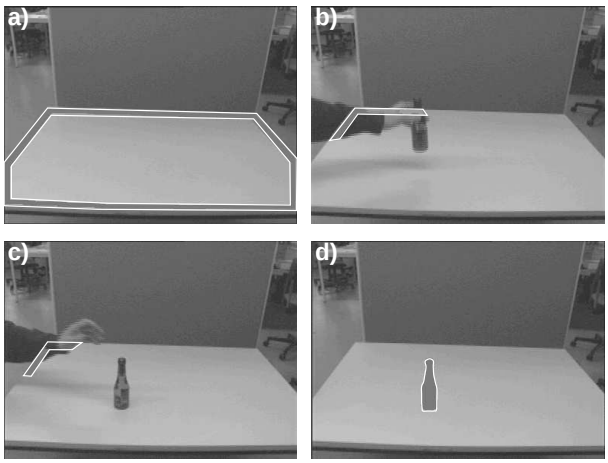


Fig. 3. Scene in front of the robot: (a) empty table with region of attention, (b) before placing the object, (c) after placing the object, (d) segmented object.

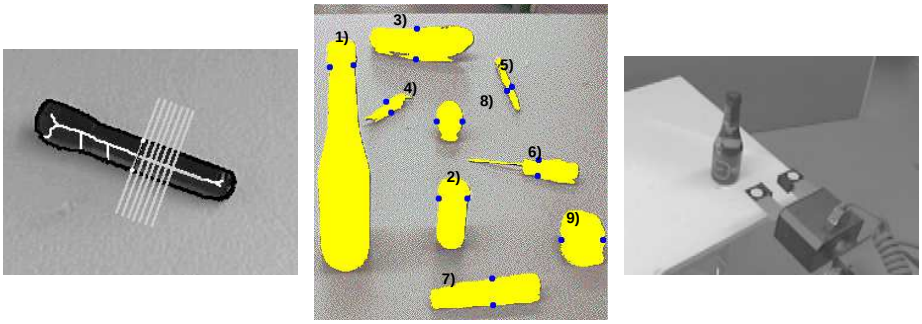
The task of determining a grasp position can be divided into the following sub-tasks: First, the object to grasp has to be detected and segmented in the image(s) (Sec. 3.1). To facilitate 3D reconstruction, the segmented image region has to be represented in a suitable way for determining point-to-point correspondences (details can be found in [1, 4]). With the help of a heuristic algorithm, potential grasping points for a two-finger gripper are determined on the 2D silhouette (Sec. 3.2). Using the corresponding grasping points in the two images, the 3D grasp position can be computed via triangulation (details in [5, 4]).

3.1 Object Detection and Segmentation

The main problem when working on the apparent contour of an object is that a very precise *segmentation* is required. Robust segmentation is a problem in itself, therefore researchers often resort to putting dark objects on white tables. We are no real exception to this rule. However, as one of our scenarios sees the robot in front of a table on which the objects to grasp are placed, we developed a method specialized on detecting any change in the scene and thereby segmenting the object to grasp. The principal idea is to acquire new images continuously and to "wait" for a change. To reduce run-time computation, a *region of attention* is defined (Fig. 3a). When placing an object into the scene, the hand will first enter the region of attention; only after it has left the region again, the inner region is checked for changes.

3.2 Determination of Grasping Points

After being detected in the stereo image pair, objects are first classified as *lying* or *standing* by estimating the rough position of special points on the boundary via



a) b) c)

Fig. 4. (a) Determining 2D grasping points on a flashlight. (b) Resulting grasping points for: (1) bottle, (2) film box, (3) zucchini, (4) pepper, (5) pen, (6) screwdriver, (7) white board marker, (8) onion, (9) walnut. (c) Gripper with artificial markers.

triangulation. Quasi-spherical objects form a class of their own. The algorithm for the determination of 2D grasping points is based on the *symmetry* of the object silhouette, which is evaluated using the *skeleton* (see e.g. [9]). Long, straight parts of the skeleton indicate a potential grasping area. Starting at the longest straight part of the skeleton, the contour is iteratively intersected with a line perpendicular to the skeleton segment (see Fig. 4a) until a computed stability measure[6] meets a given threshold.

After determining the corresponding points in the second image, the 3D grasping position is then computed via triangulation. Fig. 4b shows the extracted grasping points for a set of different objects.

3.3 Hand Position

The goal of integrating visual feedback is to compensate errors in the chain of transformations from visual input to motor output. Ideally, the algorithm providing the visual feedback, i.e. the module estimating the position of the gripper, should be affected by exactly the same model errors in exactly the same way as the algorithm estimating the position of the target. Then, the errors would cancel out completely.

For this reason, the choice of methods for estimating the hand position was limited. As a prototypical solution, we put markers on the gripper fingers that can be easily extracted from the images (see Fig. 4b). The centroids of the circular marks form the input of the triangulation algorithm.

[6] This stability measure is computed using the following criteria in addition to symmetry: (1) the distance between the two points, (2) the angle between the line connecting the two points and the horizontal plane, and (3) the distance of this line to the area centroid. The importance of the criteria depends on the object class. For more details see [5].

4 Experimental Validation

As part of our project on "Human and Robotic Hand-Eye Coordination" (TP $\mathbf{C_1}$, SFB 462), an experimental robotic platform was sought after that resembled its human counterpart regarding geometry and kinematics, in order to facilitate the transfer of knowledge. We therefore designed and integrated the hand-eye system MinERVA in an anthropomorphic fashion (Sec. 4.1). On this platform, we implemented and tested the modules described in the previous sections.

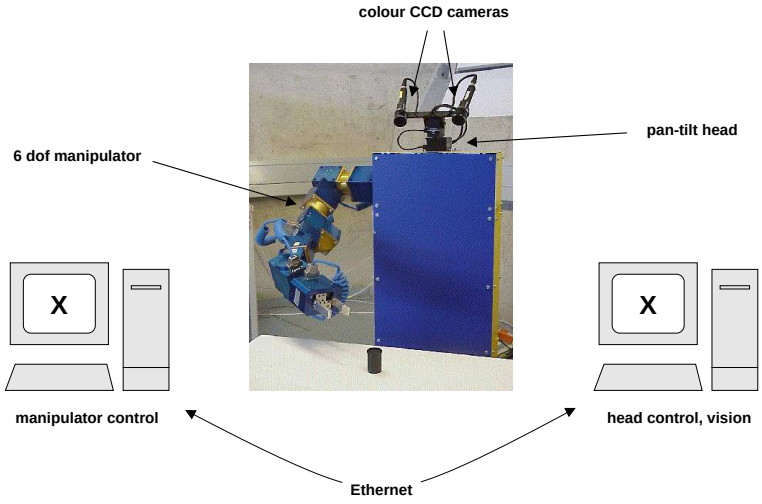


Fig. 5. The hand-eye system MinERVA

4.1 The Robotic Hand-Eye System MinERVA

MinERVA[7] consists of a 6 d.o.f. manipulator[8] and a pan-tilt head on which two color CCD cameras are mounted with a small vergence angle (Fig. 5). The system is controlled with the help of two PCs (Pentium 133 and Pentium 166) running *Linux*; processes communicate via Ethernet. Image processing including the interaction with the framegrabbers (one Matrox Meteor for each camera) was realized using the image analysis system *HALCON*, an extensive domain-independent software library providing low-level and medium-level image processing operators [9].

The camera parameters and the camera-head relation were determined using *HALCON*'s calibration method; the other parameters of the system model were extracted from the manufacturer's specifications or measured by hand.

[7] Manipulating Experimental Robot with Visually guided Actions

[8] The manipulator (*amtec*, Germany) consists of separate modules, which enables a later reconfiguration or extension. For more details on the real-time trajectory control see [10].

4.2 Results

Before the hand-eye coordination experiments, we verified that the robot actually executes the trajectories as commanded. Fig. 6a shows a setup at the *Klinikum Grohadern*: On the left side of the image one can see the receiver part of a motion registration system that tracks a small ultrasonic sender mounted on MINERVA's hand. Fig. 6b shows the measured velocity profile for a movement $20cm$ to the front. Another setup [10] consisted of a sort of writing tray on which MINERVA "wrote down" her movements (Fig. 6c).

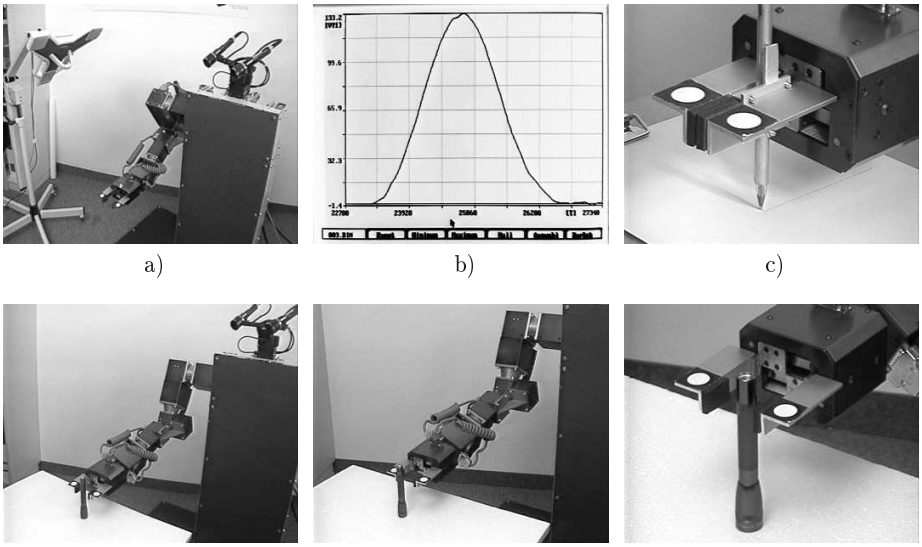


Fig. 6. (a) Setup for motion registration. (b) Measured velocity profile. (c) Writing tray. End of: (d) undisturbed, (e) disturbed, and (f) corrected movement.

The rest of Fig. 6 shows snapshots from real hand-eye experiments. In these experiments, a flashlight was placed on the table in front of MINERVA and automatically detected by the vision module. Then, grasping points were determined and sent to the motion control module. Fig. 6d shows the end of a reaching movement towards the flashlight with a normally calibrated model. In the movement of Fig. 6e, the internal model of the head was disturbed by increasing the right vergence angle by 5%, resulting in an endpoint error of about $3cm$ in each direction. Then, visual feedback is allowed once, at the beginning of the movement. The close up in Fig. 6f shows that the error can be compensated quite well.

5 Conclusion

To summarize, we have developed two components important for future autonomous service robots: First, a strategy that allows the flexible and economic use of visual information for the control of reach-to-grasp movements, and secondly, a method to determine 3D grasping points on unknown objects. We integrated and validated the components on our hand-eye system MinERVA.

Current and future work is directed at three extensions: First, the motion control part is to be extended by a module controlling hand orientation. In this context, we will extend MinERVA's arm to 7 d.o.f. Secondly, visually detected model errors should also be used to adapt the internal models. And thirdly, the system is to be extended to allow the grasping of moving objects. Together, these components will lead the way towards autonomous robotic hand-eye coordination.

References

1. T. Bandlow, A. Hauck, T. Einsele, and G. Färber. Recognising Objects by their Silhouette. In *IMACS Conf. on Comp. Eng. in Systems Appl. (CESA'98)*, pages 744–749, Apr. 1998.
2. S. Blessing, S. Lanser, and C. Zierl. Vision-based Handling with a Mobile Robot. In M. Jamshidi, F. Pin, and P. Dauchez, editors, *International Symposium on Robotics and Manufacturing (ISRAM)*, volume 6, pages 49–59. ASME Press, 1996.
3. S. R. Goodman and G. G. Gottlieb. Analysis of kinematic invariances of multijoint reaching movement. *Biological Cybernetics*, 73:311–322, 1995.
4. A. Hauck. *Vision-Based Reach-To-Grasp Movements: From the Human Example to an Autonomous Robotic System.* PhD thesis, TU München. submitted.
5. A. Hauck, J. Rttinger, M. Sorg, and G. Frber. Visual Determination of 3D Grasping Points on Unknown Objects with a Binocular Camera System. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS'99)*, Oct. 1999.
6. A. Hauck, M. Sorg, T. Schenk, and G. Frber. What can be Learned from Human Reach-To-Grasp Movements for the Design of Robotic Hand-Eye Systems? In *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA'99)*, pages 2521–2526, May 1999.
7. N. Hollinghurst and R. Cipolla. Uncalibrated Stereo Hand-Eye Coordination. *Image and Vision Computing*, 12(3):187–192, 1994.
8. S. Hutchinson, G. D. Hager, and P. I. Corke. A Tutorial on Visual Servo Control. *IEEE Trans. on Robotics and Automation*, 12(5):651–670, Oct. 1996.
9. MVTec Software GmbH. *HALCON – The Software Solution for Machine Vision Applications.* http://www.mvtec.com/halcon/.
10. G. Passig. Optimierung der Manipulatoransteuerung des Hand-Auge-Systems MinERVA. Master's thesis, TU München, Apr. 1999.
11. P. Sanz, A. del Pobil, J. Inesta, and G. Recatalá. Vision-Guided Grasping of Unknown Objects for Service Robots. In *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA'98)*, pages 3018–3025, May 1998.
12. M. Taylor, A. Blake, and A. Cox. Visually guided grasping in 3d. In *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA'94)*, pages 761–766, 1994.
13. M. Tonko, J. Schurmann, K. Schäfer, and H.-H. Nagel. Visually Servoed Gripping of a Used Car Battery. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS'97)*, pages 49–54, Sept. 1997.