

# Biologically Motivated Visual Servoing and Grasping for Real World Tasks

D. Kragic and H. I. Christensen

Centre for Autonomous Systems, Royal Institute of Technology, Stockholm, Sweden  
{danik, hic}@nada.kth.se

**Abstract**—Hand-eye coordination involves four tasks: i) identification of the object to be manipulated, ii) ballistic arm motion to the vicinity of the object, iii) preshaping and alignment of the hand, and finally iv) manipulation or grasping of the object. Motivated by the operation of biological systems and utilizing some constraints for each of the above mentioned tasks, we are aiming at design of a robust, robotic hand-eye coordination system. Hand-eye coordination tasks we consider here are of the basic *fetch-and-carry* type useful for service robots operating in everyday environments. Objects to be manipulated are, for example, food items that are simple in shape (polyhedral, cylindrical) but with complex surface texture. To achieve the required robustness and flexibility, we integrate both geometric and appearance based information to solve the task at hand. We show how the research in human visuo-motor system can be facilitated to design a fully operational, visually guided object manipulation system.

## I. INTRODUCTION

Robotic visual servoing and manipulation has received significant attention during the past few years [1], [2]. Despite of this, most of the examples are still limited to final alignment where point-to-point visual servoing is employed. However, the alignment step is just one of the steps in an object manipulation sequence which should also include grasping or manipulation of objects. The integration of the building blocks required to perform the whole task has proven to be difficult and systems that can demonstrate it are very few (if any for general settings).

We believe that for service robot applications it is important to observe the complete task. Assuming basic *fetch-and-carry* tasks, there are varying demands for precision depending on the sub-task complexity. As proposed in [3], a key to solve hand-eye tasks efficiently and robustly is to identify how precise control is needed at a particular time during task execution. This should then be matched with appropriate sensory input as shown in Fig 1. This is also the main idea pursued in our work. Compared to the work presented in [3], where the main consideration was a control framework, we will concentrate more on the design of the vision and tactile based feedback systems for object manipulation and grasping.

In our previous work, we have studied mobile manipulation in real world scenarios such as a living room where the problem of sensor integration for mobile manipulation

was discussed [4]. Two major problems were considered: i) *robust* perception-action integration, and ii) *generic* models for systems integration. The problem of mobile manipulation was chosen as a test case, since it required addressing the following issues: a) navigation, b) object recognition, c) figure-ground segmentation, d) servoing to a position that allows grasping, e) grasp planning, f) control generation and g) integration of the above into a unified framework.

In this paper we extend the visual perception and grasping parts by improving the step d) servoing to a position from which a grasp can be executed and step e) grasp planning. Our goal here is to show how the human hand-eye coordination strategies can be used to design robotic systems with different levels of complexity. Consequently,

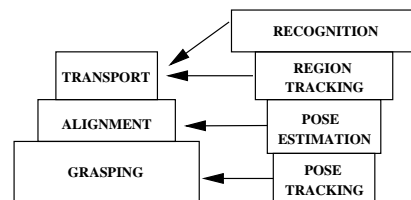


Fig. 1. Robot control versus vision hierarchy. The required complexity of visual feedback depends on the manipulation task.

we have studied reports of how the human visuo-motor system performs under similar conditions. The main motivation for our work was the research performed on human visuo-motor systems [7], [8], [23]. In [8] it was shown that kinematic parameters of the human grasp (transport velocity, path, grip aperture) are strongly correlated with the 3-dimensional geometric structure of the target object and not the 2-dimensional projected image of the object (which is commonly used in vision based control systems today). This research, however, concentrated mostly on the identification of human visuo-motor system performance whereas we are interested in the development of the real robotic system. In [23], it was demonstrated that visually controlled tasks in humans are to a large extent based on feed-forward control of the arm. Once a target has been identified/located, the control is predictive rather than based on strong feedback information. We will here

illustrate how these observations can be utilized in the design of a robotic system. The details about the control architecture can be found in [10].

The paper is organized as follows: in Section II we briefly present the system framework. In the remaining sections, the details about the estimation of visual feedback are given depending on the level of complexity of the current task. Section III presents a 2D tracking system and Section IV presents an appearance based approach used for initial pose estimation of the object. In Section V the design of the grasping system is presented. We conclude by a short discussion in Section VI. Since this paper covers a large amount of research, we cannot provide all the details about each of the topics. The emphasis of the paper is on the overall strategy and details can be found in the publications that are referenced at relevant places.

## II. THE IMPLEMENTATION PLATFORM

To integrate basic building blocks, a Distributed Control Architecture (DCA) is used, [10]. The experimental platform is a Nomadic Technologies XR4000 equipped with a Puma 560 arm, see Fig. 2. The robot has two rings of sonars, a SICK laser scanner, a wrist mounted force/torque sensor (JR3), and a color CCD camera mounted on the gripper (Barrett hand). For the control of grasping, the Barrett hand has been equipped with two types of tactile sensors. The palm is equipped with a touch pad for detection of palm contacts. Each link of the three fingers has a pair of tactile sensors as well as each tip of a finger.

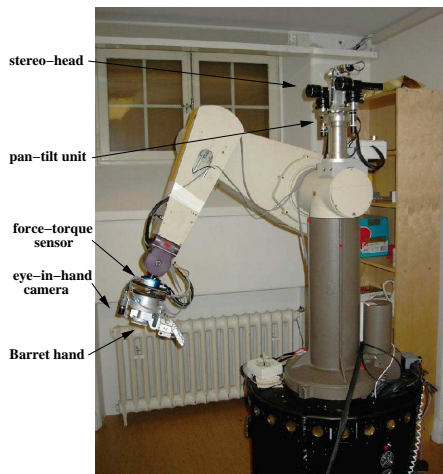


Fig. 2. The experimental platform.

### A. Visual servoing

The study on human reaching movement performed in [17], proposed that a reaching movement consists of two components: an *initial impulse propelling the hand*



Fig. 3. Some of the objects we want the robot to manipulate.

towards the target and a *current control to home in on the final position via successive approximations*.

Based on this, we consider different levels of performing an object manipulation task as presented in Fig 1. **Transport** or **pre-grasping** considers motion of the robot platform and/or the robot arm to the vicinity of the object. In our framework, the robot usually starts to search for the object at a well defined position, for example, on the dinner table. Given a task such as “pickup the cup from the dinner table”, the robot will navigate to the vicinity of the table. A visual search/object recognition is then performed. After the object has been located, the robot moves towards it keeping it in the field of view. For this purpose, an image based tracking (2D) approach suffices. This is explained in more detail in Section III. To prepare for grasping, the robot has to **align** the hand with the object or reach a pose from which the **grasping** can easily be performed. For this purpose, we use both geometric and appearance based models to estimate pose of the object as explained in Section IV and Section V.

## III. TRANSPORT/PRE-GRASPING

To transport the arm to the vicinity of the object, the object is first located. After that, an image based tracking system is used to keep the object in the field of view while approaching it. There are two alternatives to this approach. The first is to estimate the 3D pose of the object and perform a ballistic movement to a position close to the object. The second is to use a stereo based recognition to provide a rough estimate of the object’s pose. Both of these approaches have been demonstrated in [6].

### A. Recognition

Recent research on human vision has clearly shown that object recognition can be efficiently modeled as a view based process [24], [25], which motivates our use of an Support Vector Machines (SVM) based method for recognition, [20]. The recognition step delivers the image position and approximate size of the image region occupied by the object.

## B. 2D integration for tracking

For decades, neuroscientists have tried to answer the question of how the brain integrates information from different cues into coherent percepts [9]. Lots of perceptual experiments support the idea that when it comes to aspects of visual scenes, the most frequently used cues are color, form and motion. There is a belief that information about form, color, motion and depth is processed separately in the visual system. However, it has also been shown that the segregation is not complete and there is a cross-talk among different cues [12].

The integration of information from multiple cues has extensively been studied in computer vision [13]. There are three basic principles for integration: i) one source *dominates* all other sources, ii) there is a *compromise* or *weak fusion* between sources or iii) there is some level of *interaction* between cues to arrive to an optimal solution.

We use a voting principle [5] where responses from four different visual cues (motion, color, correlation and texture) are fused using weighted super-position:

$$\delta(a) = \sum_{i=1}^n w_i O_{c_i}(a) \quad (1)$$

where  $n$  is the number of cues,  $O_{c_i}$  is the output of a cue  $i$  and  $w$  its reliability. The most appropriate action is selected according to a winner-take-all strategy:

$$a' = \operatorname{argmax}\{\delta(a) | a \in \mathbf{A}\} \quad (2)$$

The advantage of the voting approach for integration is that the response from different cues can be easily combined without the need for explicit models as it is for example the case in Bayesian approaches. The voting based fusion is at the same time general enough to allow implementation of dominating, compromising and/or weak fusion. What we have found important and therefore investigated in more detail were: i) the choice of voting space, and ii) the estimation of cues' reliability.

## C. Voting Space

The basic idea for this investigation comes from two selection processes: attention (covert selection) and eye movements (overt selection) [12]. The experiments on the human visual system indicate that attention precedes eye-movements as an integral step in preparing to move the eyes. In our experiments, we have tried to mimic these two systems by choosing two different voting approaches: i) *response fusion*, and ii) *action fusion*.

The first approach makes the use of “raw” responses from the individual cues. This can be seen as the *spatial attention mechanism* where each of the cues votes for a particular region in the image. The “winner” region is then the image position for “where-to-look-next”. The second approach uses a different action space represented

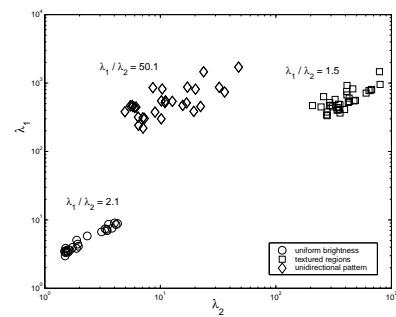


Fig. 4. Eigenvalues of the matrix in (Eq. 3) and their ratios for uniform brightness, textured and regions with unidirectional pattern.  $\lambda_1/\lambda_2$  represent mean value of ratios for each cluster.

by *direction* and *speed*. This can be seen as an *eye-movement mechanism* where the compromised result is then “where-to-move-the-eyes” or gaze selection.

## D. Reliability and Adaptability

The reliability of each cue which is used as a weight in Eq. 1 can be preset or selected adaptively depending on scene contents. In our previous work, [5] we have investigated four different weighting strategies. For the first two, the weights are preset and kept constant during tracking:

**Uniform weighting** - all cues have equal weights.

**Texture based weighting** - weights depend on textural properties of the object. If the object is uniform in color, the color cue is weighted more heavily. A simple method to determine textural properties is to estimate the image gradient in a local region[11]:

$$\mathbf{D} = \begin{bmatrix} \sum I_x^2 & \sum I_x I_y \\ \sum I_x I_y & \sum I_y^2 \end{bmatrix} \quad \text{with} \quad (\lambda_1, \lambda_2) = \operatorname{eig}(\mathbf{D}) \quad (3)$$

where  $I_x$  and  $I_y$  are spatial image intensity gradients. This matrix should be above noise-level and well-conditioned. If the eigenvalues are small, there is a roughly constant intensity profile within a window. A large and a small eigenvalue represent a uni-directional pattern while large eigenvalues represent corners, salt-and-pepper textures and are assumed to represent a region that can be reliably tracked. Fig. 4 shows a few examples of eigenvalues and their ratio for uniform brightness, textured and regions with unidirectional pattern. The values are obtained for window sizes between 15 and 45 pixels and normalized by window size.

It has been shown that observers' cue-integration strategies are adaptable but still little is known about the adaptation process itself. In [9], it was experimentally shown that reweighting in humans takes place in about 1s. Therefore, we have designed the tracking system so to adapt the weights automatically based on the performance of each cue. This has also provided mechanisms for suppressing

cues when they are considered to be unreliable, which is determined by the level of agreement with the general consensus. The two approaches investigated were:

**One-step distance weighting** - We have investigated whether a very fast reweighting of cues, as for example demonstrated in case of human tracking [9], can be beneficial to emphasize the information provided by the consensus between the cues. This “short-term-memory” approach estimates the absolute distance difference between the state estimated by each of the cues and the state estimated after the integration between two frames. Then, the reliability of a cue is inverse proportional to this difference.

**History based weighting** - It has been demonstrated that visual cue-integration strategies are adaptable in an experience-dependent manner where the adaptation happens on a relatively long time scale (hours or days). Consequently, we allow the system to adapt each cue depending on the cues’ performance during the entire tracking task. The difference measure is similar to the previous approach, but instead of using two consecutive frames, we compare the distance for all the frames up to the current one.

Results reported in [5], where we have experimentally tested these approaches, showed that texture based weighting resulted in lowest tracking error (measured by the absolute distance in pixels from the ground truth value). This is in accordance to the results reported in [19] where a comparison between weighting approaches showed that assigning equal weights to all cues performed best. As argued, this is due to the many sudden changes usually occurring in real-world sequences. This is also one of the strong arguments for considering an integration approach for tracking in real-world settings since non of the cues will be reliable throughout the entire tracking sequence.

#### E. Eye/Camera Control

To position an object on the fovea and keeping it there, humans use eye, neck and body movements. Although humans have a rich variety of eye movements [26], for our application we consider two types: fast saccades and smooth pursuit. Due to the high bandwidth of the arm (500Hz) and low bandwidth of the base (10Hz), it was natural to use the two last joints of the arm for mimicking the saccadic motion to center the object on the image plane. The smooth pursuit is implemented as an integral part of the visual servoing loop. The base motion is used to approach the object since the initial distance from the object is often more than 1.5m which is more than the arm can be stretched. A control approach for this type of servoing was presented in our previous work, [21].

### IV. ALIGNMENT

To align the robot hand with the object, prior knowledge about objects geometrical properties can be used. This

way, the approximate position and orientation of the object relative to the camera or some other coordinate system can be estimated. It is obvious that for manipulation of everyday objects the geometrical representation will not suffice due to the complex textural properties. We deal with brightness images that are functions of both shape and intrinsic scene properties (reflectance, illumination). So, rather than relying only on shape properties, we integrate them with an appearance based method.

#### A. Subspace Methods

The question of how the human visual system exploits the statistical structure of “natural images” has been studied extensively. The term “natural images” refers to retinal images that arise from natural environments under natural viewing conditions [12]. A number of experiments have shown that certain cells of human brain are optimized for extracting information according to particular computational principles where two of them are: **compact coding** and **sparse distributed coding**. Compact coding assumes that the image should be represented by a minimum number of units. A typical example of this coding type is principle component analysis (PCA) that identifies a reduced set of basis functions that capture the maximum variance of the subsets of points corresponding to a set of images. Sparse distributed coding, on the other hand, represents data with the minimum number of *active* units where redundancies in images are usually removed by, for example, unsupervised learning algorithms [12].

Our approach is mainly motivated by ideas proposed in [18] where PCA was exploited to estimate three pose parameters used to move a robotic arm to a predefined pose with respect to the object. Compared to our approach, where the pose is expressed relative to the camera coordinate system, they express the pose relative to the current arm configuration, making the approach unsuitable for robots with different number of degrees of freedom. In our system, during the learning stage each image is projected as a point to the eigen-space and the corresponding pose of the object is stored with each point. For each object, we have used 96 training images (8 rotations for each angle on 4 different depths). One of the reasons for choosing this low number of training images is the workspace of the PUMA560 robot which is fairly limited and for our applications this discretization was adequate. To enhance the robustness with respect to variations in intensity, all images are normalized. At this stage, the size of the training samples is  $100 \times 100$  pixels. The training procedure takes about 3 min on a Pentium III 550MHz.

This pose estimation procedure provides only an approximation to the real pose as shown in Fig.5. To estimate the true pose of the object, the pose initialization step is followed by a local fitting method using a geometric model of the object. Section V-A provides details about this step.

A number of researchers have reported that the invariant features of human multi-joint arm movements are that i) the path of the hand is roughly straight line in Cartesian coordinates, and that ii) the profile of tangential (Cartesian) hand velocity is bell-shaped [16]. In addition, it has been shown in [7] that visuo-motor actions such as grasping use the actual size of the object and that the position and orientation are computed in egocentric frames of reference. Thus, human reaching movements are planned in spatial coordinates, not in joint space.

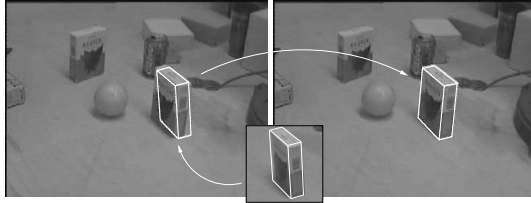


Fig. 5. Small image is the training image used to estimate the closest pose given the current image. Left) the initial pose overlaid, and right) the final pose obtained after local refinement.

Since an accurate pose of the target is available together with a good arm model (which is true in our case), we use the ideas proposed in [22] to generate human-like arm trajectories. Here, a feedback scheme was proposed that, similarly to visual servo control, estimates the current velocity from the remaining distance to the target.

## V. GRASPING

After objects pose is available, grasping is performed. It has been shown in [8] that the 3D geometric properties of the target affect the kinematic parameters of human grasping movements. In addition, the maximum grip aperture is correlated with the actual size of the target rather than the size of its projected retinal image as demonstrated in [14]. This implies that the use of geometric models is more typical to human grasping than use of, for example, bounding contours and similar approaches.

### A. Control of corrective movements

The final positioning for grasping in case of humans is achieved through a number of successive approximation movements, [17]. When a perfect kinematic model of the arm/object is not available or if the object is not static, there is a need for visual feedback. We have therefore implemented a model based tracking system for pose estimation, see Fig.8. Lie group and Lie algebra formalism are used as the basis for representing the motion of a rigid body and pose estimation, as proposed in [15]. The same principle is used for the local fitting step.

Using the available pose estimate and tactile feedback, our grasping system is able to compensate for minor errors in the pose estimate. The grasping strategy is formulated using finite state machines, [27]. Using the general idea proposed in [28], the basic states mimic the human grasping procedure as shown in Fig. 6. The details about the grasp state minimization and implementation are provided in [29].

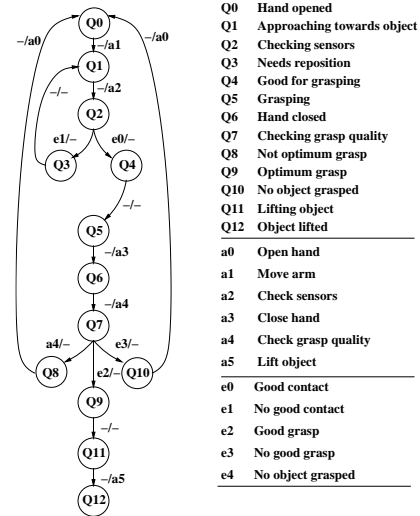


Fig. 6. Abstract representation of grasping.

## VI. SUMMARY AND CONCLUSIONS

Using human visuo-motor system principles for designing robotic systems have been demonstrated previously in [16]. It was argued that sparse, asynchronous visual feedback will suffice for most of the object manipulation tasks. This research was mainly concentrated on development of control models to generate human-like grasping trajectories and use of both feedforward and feedback strategies. Our current research has demonstrated that through careful consideration of the evidence available in the physiological and cognitive literature it is possible to design efficient strategies for control of the full hand-eye coordination process from recognition to actual grasping.

Considering manipulation of objects, we have designed a system that provides visual feedback at different levels of complexity - from a simple image (2D) position estimation to complete 3D pose estimate of the object. Using only an eye-in-hand camera configuration and its relatively small field of view ( $40^\circ$ ), it was not possible to continuously estimate the objects' pose when the hand was less than 20cm from the object. Our current work therefore includes also a stereo head which allows for observing both the hand and the object during the alignment task. We are currently investigating the idea of a hybrid control



Fig. 7. Our approach: after the object is recognized, 2D tracking is used to approach the object. After that, the appearance based approach followed by a local fitting stage is used to estimate the current pose of the object. After that, the grasping can be performed.

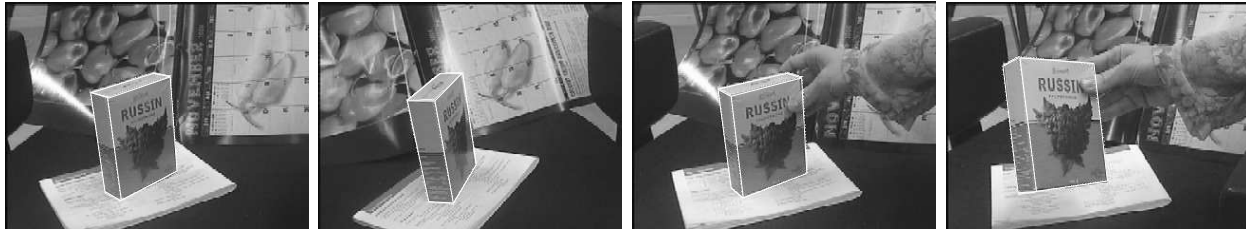


Fig. 8. A tracking example: a fairly textured object against a textured background. The estimated pose of the object is overlaid in white. During this experiment a 6mm lens was used and the object was at the distance of approximately 50cm from the camera.

structure in order to combine the advantages of “look-then-move” and “look-and-move” control approaches.

## VII. REFERENCES

- [1] S.Hutchinson,G.D.Hager,P.Corke ‘A tutorial on visual servo control’, *IEEE TRA* 12(5), 651–670, ’96.
- [2] D. Kragic. “Visual Servoing for Manipulation: Robustness and Integration Issues”, *PhD thesis*, Royal Institute of Technology, Stockholm, Sweden, 2001.
- [3] Z. Dodds and M. Jägersand and G. Hager and K. Toyama, “A hierarchical vision architecture for robotic manipulation tasks”, *ICVS99*, 312-331
- [4] L. Petersson, P. Jensfelt, D. Tell, M. Strandberg, D. Kragic and H. Christensen, “Systems Integration for Real-World Manipulation Tasks”, *ICRA02*,3:2500-2505
- [5] D. Kragic and H. Christensen, “Weak models and cue integration for real-time tracking”, *ICRA02*,3:3044-3049
- [6] D. Kragic and H. Christensen, “A framework for visual servoing”, *ICVS03*, Graz, Austria
- [7] Y. Hu and M. Goodale, “Constraints in Human visuomotor systems”, *IROS00*,2:1633-1638
- [8] Y. Hu, R. Eagleson and M. Goodale, “Human visual servoing for reaching and grasping: The role of 3D geometric features”, *ICRA99*,3:3209-3216
- [9] J.Triesch, D.H.Ballard and R.A.Jacobs, “Fast temporal dynamics of visual cue integration”, *Perception*02,3:421-434
- [10] L.Petersson, D.Austin, and H.I.Christensen. “DCA: A Distributed Control Architecture for Robotics”, *IROS02*,3:2361-2368
- [11] J.Shi, C.Tomasi, “Good features to track”, *CVPR94*, 593-600
- [12] S.E. Palmer, “Vision Science: Photons to Phenomenology”, *MIT Press*, Cambridge, MA, 1999
- [13] Clark and Yuille, “Data fusion for sensory information processing systems”, *Kluwer Academic Publisher*, 1990
- [14] P. Mamassian, “Prehension of objects oriented in 3D space”, *Experimental Brain Research*, 114:235-245, ’97
- [15] T.Drummond and R.Cipolla. Real-time tracking of multiple articulated structures in multiple views. *ECCV00*, 2:20-36
- [16] A.Hauck, M.Sorg, T.Schenk and G.Färber, “What can be Learned from Human Reach-To-Grasp Movements for the Design of Robotic HandEye Systems?”, *ICRA99*, 2521-2526
- [17] R. Woodworth, “The accuracy of voluntary movement”, *Psychological Review*,3:1-114, 1899
- [18] S.K. Nayar, S.A. Nene, and H. Murase. “Subspace methods for robot vision”, *IEEE TRA*, 12(5):750–758, 1996.
- [19] J. Triesch and C. von der Malsburg, “Self-Organized Integration of Adaptive Visual Cues for Face Tracking”, *Int. Conf. on Automatic Face and Gesture Recognition00*
- [20] D.Roobaert, “Pedagogical Support Vector Learning: A Pure Learning Approach to Object Recognition”, PhD thesis, CVAP, KTH, Stockholm, Sweden, 2001
- [21] H.Sidenbladh, D.Kragic and H.Christensen, “A person following behavior for a mobile robot” *ICRA99*, 1:670-675
- [22] S.R.Goodman and G.G.Gottlieb, “Analysis of kinematic invariances of multijoint reaching movements”, *Biological Cybernetics*, 73:311-322, ’95
- [23] R.Johansson,G.Westling,A.Bäckström,J.Flanagan, “Eye-hand coordination in object manipulation.” *Jour. Neurosci.* 21:6917.-6932, 2001
- [24] M.Tarr, H.Bulthoff, “Object recognition on man, monkey and machine”, *Intl. Jour of Cognitive Science*, 69:1-2,1998.
- [25] S.Edelman, “Representation and Recognition in Vision”, *MIT Press*, Cambridge, MA. 1999.
- [26] R.H.S.Carpenter, “Movement of the Eyes”, *Pion Ltd*, 1998.
- [27] R.H.Katz, “Contemporary logic design”, *Benjamin Cummings/Addison Wesley Publishing Company*, 1993
- [28] I.Horswill, “Behaviour-Based Robotics, Behaviour Design”, *Tech. report CS395*, Northwestern University, 2000
- [29] D.Kragic, S.Crinier, D.Brunn and H.I.Christensen, “Vision and Tactile Sensing for Real World Tasks”, *ICRA03*.