An Active Vision System for Detecting, Fixating and Manipulating Objects in Real World

B. Rasolzadeh, M. Björkman, K. Huebner and D. Kragic

August 24, 2009

Abstract

The ability to autonomously acquire new knowledge through interaction with the environment is an important research topic in the field of robotics. The knowledge can be acquired only if suitable perception-action capabilities are present: a robotic system has to be able to detect, attend to and manipulate objects in its surrounding. In this paper, we present the results of our longterm work in the area of vision based sensing and control. The work on finding, attending, recognizing and manipulating objects in domestic environments is studied. We present a stereo based vision system framework where aspects of Top-down and Bottom-up attention as well as foveated attention are put into focus and demonstrate how the system can be utilized for robotic object grasping.

1 Introduction

Humans use visual feedback extensively to plan and execute actions. However, this is not a well-defined one way stream: how we plan and execute actions depends on what we already know about the environment we operate in (context), what we are about to do (task), and what we think our actions will result in (goal). A significant amount of human visual processing is not accessible to consciousness - we do not *experience* using optical flow to control our posture, (Sloman, 2001). By not completely understanding the complex nature of human visual system, what are the ways to model similar capabilities into robots?

Visual attention plays an important role when we interact with the environment, allowing us to deal with the complexity of everyday scenes. The requirements on artificial 'seeing' systems are highly dependent on the task and have historically been developed with this in mind. To deal with the complexity of the environment, prior task and context information have commonly been integrated with low level processing structures, the former being denoted as Top-down and latter Bottom-up principle.

In our research, tasks such as “Robot, bring me my cup” or “Robot, pick up this” are studied. Depending on the task or context information, different execution strategies task may be chosen. The first task is well defined in that manner that the robot already has the internal representation of the object - the *identity* of the object is known. For the second task, the spoken command is commonly followed by a pointing gesture - here, the robot does not know the *identity* of the object, but it can rather easy extract its *location*. A different set of underlying visual strategies are required for each of these scenarios being the most representative examples for robotic fetch-and-carry tasks. We have worked with different aspects of the above for the past several years, (Björkman and Eklundh, 2002; Björkman and Eklundh, 2006; Björkman and Kragic, 2004; Ekvall and Kragic, 2005; Ekvall *et al.*, 2007; Huebner *et al.*, 2008a; Kragic and Kyrki, 2006; Petersson *et al.*, 2002; Rasolzadeh *et al.*, 2006, 2007; Topp *et al.*, 2004). The work presented here continues our previous works, but the focus is put on the design and development of a vision system architecture for the purpose of solving more complex visual tasks. The overall goal of the system is to enable the understanding and modeling of space in various object manipulation scenarios. A schematic overview of the experimental platform is shown in Fig. 1. The different parts of this illustration are described in detail throughout this paper.

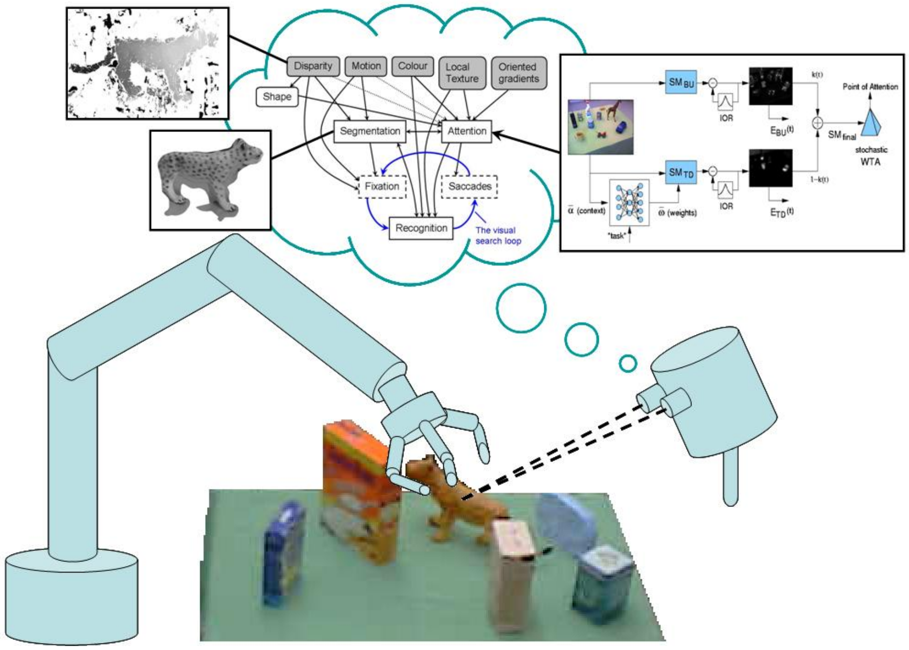


Figure 1: Illustration of the complete robotic platform that is the system described in this paper. See Fig. 2 and 4 for detailed illustrations, and Fig. 10 for an illustration of the actual setup.

The paper is organized as follows. In Section 2 system functionalities, considered tasks, system design and flow of information in the system are outlined.

This corresponds roughly to the diagram in the middle of Fig. 1. In Section 3, the details about the camera system and its calibration are given. Aspects of Bottom-up and Top-down attention are presented in Section 4 and foveated segmentation in Section 5. Section 6 describes how the visual system can be used to facilitate manipulation. Selected results of the experimental evaluation are presented in Section 7, where an evaluation of the attention-system and the recognition-system is first performed separately, followed by a find-and-remove-object task in Section 7.3. A discussion and summary finalizes the paper in Section 8.

2 Vision System Functionalities and Tasks

Similar to the human vision system, but unlike many systems from the computer vision community, robotic vision systems are embodied. Vision is used as a mean for the robot to interact with the world. The system perceives to act and acts to perceive. The vision system is not an isolated entity, but part of a more complex system. Thus, the system should be developed and evaluated as such. In fact, measuring the performance of system components in isolation can be misleading. The quality of a component depends on its ability to function within the rest of the system. Computational speed might sometimes be preferable to accuracy or vice-versa. As a designer, one has to take a step backwards and concentrate on the tasks the robotic system is supposed to perform and the world in which the system resides. What are the goals of the system? What can be expected from the world and what cannot be expected?

Examples of recent work taking the embodiment into account have been demonstrated in Björkman and Eklundh (2006); Ude *et al.* (2006). In these systems vision is embodied in a robotic system capable of visual search as well as simple object manipulation. The goal of the work presented here is to design a similar robotic system able to interact with the world through recognition and manipulation of objects. Objects can either be previously known or completely new to the system. Even if confusions do occur frequently, a human being is able to immediately divide the perceived world into different physical objects, seemingly without effort. The task is performed with such ease that the complexity of the operation is easily underestimated. For a robotic system to perform the same task, the visual percept has to be grouped into larger entities that have some properties in common, properties such as proximity and appearance. These perceptual entities might or might not correspond to unique physical objects in 3D space. It is not until the robot acts upon an entity, that the existence of a physical object can be verified. Without interaction the entity has no real meaning to the robot. We call these entities *things* to differentiate them from *objects* that are known to be physical objects, through interaction or other means. The idea of *things* and *objects* is the foundation of the project PACO-PLUS [1] and the recent work presented in Geib *et al.* (2006); Kraft *et al.* (2008); Wörgötter *et al.* (2009).

[1] www.paco-plus.org

For the visual system to be of use in robotic tasks, it needs the abilities to divide the world into *things*, represent these for later association and manipulation, and continuously update the representation as new data becomes available. A representation can either be short-lived and survive only during a short sequence of actions, or permanent, if interactions with the *thing* turn out to be meaningful. A meaningful action is an action that results in some change in the representation of the *thing*, such as a pushing action resulting in a change in position. From this stage on, the *thing* is considered an *object*.

The amount of perceptual data arriving through a visual system easily becomes overwhelming (Tsotsos, 1987). Since resources will always be limited in one way or the other, there is a need for a mechanism that highlights the most relevant information and suppresses stimuli that is of no use to the system. Instead of performing the same operations for all parts of the scene, resources should be spent where they are needed. Such a mechanism is called *visual attention.* Unfortunately, relevancy is not a static measure, but depends on the context, on the scene in which the robot acts and the tasks the robot is performing. Consequently, there is a need for the attentional system to adapt to context changes as further studied in Section 4. A static *thing* too large for the robot to manipulate might be irrelevant, while an independently moving *thing* of the same size can be relevant indeed, if it affects the robot in achieving its goals. Since sizes and distances are of such importance to a robotic system, a visual system should preferably consist of multiple cameras.

2.1 Flow of Visual Information

The visual system used in our study has a set of basic visual functionalities, the majority of which uses binocular cues, when such cues are available. The system is able to attend to and fixate on *things* in the scene. To facilitate object manipulation and provide an understanding of the world, there is support for figure-ground segmentation, recognition and pose estimation. All these processes work in parallel, but at different time frames, and share information through asynchronous connections. The flow of visual information through the system is summarized in Fig. 2. Information computed within the system is shown in rounded boxes. Squared boxes are visual processes that use this information. Grey boxes indicate information that is derived directly from incoming images. The camera control switches between two modes, fixation and saccades, as illustrated by the dashed boxes. The vision system generally works within the visual search loop that consists of a saccade to the current attentional foci, after which the system tries to fixate on that point, which in turn will yield more (3D) information for the recognition step. If the attended/fixated region is not the desired object we are searching for, the visual search loop continues.

The above-mentioned vision system has been implemented on the four-camera stereo head (Asfour *et al.*, 2006) shown in Fig. 10. The head consist of two foveal cameras for recognition and pose estimation, and two wide field cameras for attention. It has seven mechanical degrees of freedom; neck roll, pitch and yaw, head tilt and pan & tilt for each camera in relation to the neck. The

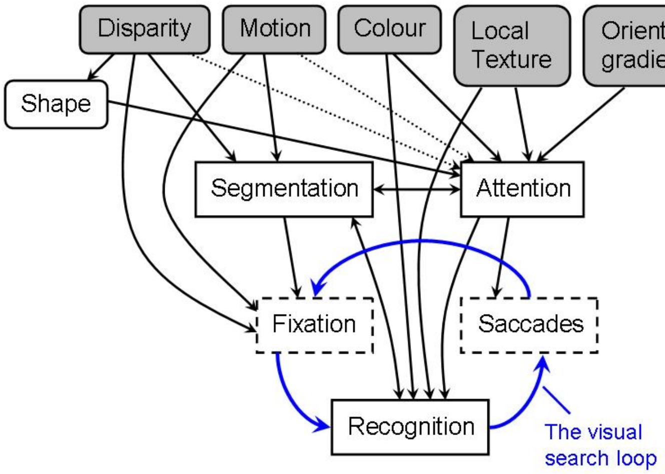


Figure 2: The flow of visual information.

attentional system keeps updating a list of scene regions that might be of interest to the rest of the system. The oculomotor system selects regions of interest from the list and directs the head so that a selected region can be fixated upon in the foveal views. Redirection is done through rapid gaze shifts (saccades). As a consequence, the camera system always strives towards fixating on some region in the scene. The fact that the system is always in fixation is exploited for continuous camera calibration and figure-ground segmentation.
Given the large focal lengths of the foveal cameras, the range of possible disparities can be very large, which complicates matching in stereo. With the left and right foveal cameras in fixation, we know that an object of interest can be found around zero disparity. This constrains the search range in disparity space, which simplifies stereo matching and segmentation.

2.2 Design Issues

We have chosen a design methodology that is biologically inspired, without the ambition to make our systems biologically plausible. Since computational and biological architectures are so fundamentally different, biologically plausibility comes at a cost. One critical difference is the relative costs of computation and communication of the estimated results. In biological systems, computations are done in neurons, with results communicated through thousands of synapses per neuron. This is much unlike computational systems in which the cost of communicating data, through read and write operations to memory, can be higher than that of computing the actual data. Unfortunately, computer vision tend to be particularly memory-heavy, especially operations that cover whole images. If one considers typical real-time computer vision systems, the cost of storage easily out-weight the cost of computation. Thus for a system to perform at real-time speed, biological plausibility easily becomes a hindrance. Even if biological systems inspire the design process, our primary interest is that of robotic manipulation in domestic environments.

3 Camera System

For a robot to successfully react to sudden changes in the environment the attentional system ought to cover a significant part of the visual field. Recognition and vision-based navigation, however, place another constraint on the visual system, i.e. a high resolution. Unfortunately, these two requirements are hard to satisfy for a system based on a single stereo pair. A biological solution, exemplified by the human vision system, is a resolution that varies across the visual field, with the highest resolution near the optical centers. There are some biologically-inspired artificial systems (Kuniyoshi *et al.*, 1995; Sandini and Tagliasco, 1980) that use similar approaches. However, non-uniform image sampling leads to problems that make these systems less practical. Binocular cues can be beneficial for a large number of visual operations, from attention to manipulation, and with non-uniform sampling stereo matching becomes hard indeed. Furthermore, the reliance on specialized hardware makes them more expensive and less likely to be successfully reproduced. Another possible solution is the use of zoom lenses (Paulus *et al.*, 1999; Ye and Tsotsos, 1999). While the robot is exploring the environment the lenses are zoomed out in order to cover as much as possible of the scene. Once an object of interest has been located, the system zooms in onto the object to identify and possibly manipulate it. However, while the system is zoomed in it will be practically blind to whatever occurs around it.

Other than the obvious reasons of cost and weight, there is nothing conceptually preventing us from using more than just two cameras, which is the case in solutions based on either zoom-lenses or non-uniform sampling. Instead, one could use two sets of stereo pairs (Scassellati, 1998), a wide-field set for attention and a foveal one for recognition and manipulation. The most important disadvantage is that the calibration process becomes more complex. In order to relate visual contents from one camera to the next, the relative placement and orientation of cameras have to be known.

Sets of four cameras can be calibrated using the quadrifocal tensor (Hartley and Zisserman, 2000), or the trifocal tensor if sets of three are considered at a time. However, the use of these tensors assumes that image features can be successfully extracted and matched between all images considered. Depending on the camera configuration and observed scene, it may not at all be the case. For example, due to occlusions the visual fields of the two foveal images might not overlap. Furthermore, since the visual fields of the foveal cameras are so much narrower than those of the wide-field ones, only large scale foveal features can be matched to the wide-field views. The largest number of matchable features is found if only two images are considered at a time and the corresponding focal lengths are similar in scale. Thus for the system presented in this paper, we use

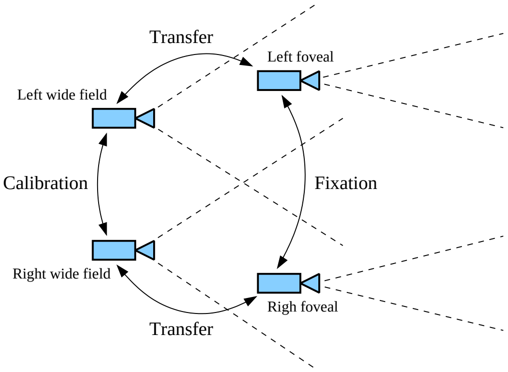


Figure 3: Two sets of cameras, a wide-field camera set for attention and a foveal one for recognition and manipulation, with external calibration performed between pairs.

pair-wise camera calibration as illustrated by the arrows in Fig. 3.

3.1 Wide-Field Calibration

Since external calibration is inherently more stable if visual fields are wide, we use the wide-field cameras as references for the foveal ones. This calibration is an on-going process, where previous estimates are exploited for feature matching in the subsequent frames, assuming a limited change in relative camera orientation from one frame to the next. The purpose of the calibration is two-fold. First, given a known baseline (the distance between the cameras) it provides a metric scale to objects observed in the scene. Second, calibration provides parameters necessary for rectification of stereo images, so that dense disparity maps can be computed, as it will be shown in Section 4.4.

In our study we related the wide-field cameras to each other using an iterative approach based on the optical flow model (Longuet-Higgins, 1980):

$$\begin{pmatrix} dx \\ dy \end{pmatrix} = \frac{1}{Z} \begin{pmatrix} 1 & x \\ 0 & y \end{pmatrix} \begin{pmatrix} t_x \\ t_z \end{pmatrix} + \begin{pmatrix} 0 & 1+x^2 & -y \\ 1 & xy & x \end{pmatrix} \begin{pmatrix} r_x \\ r_y \\ r_z \end{pmatrix} \tag{1}$$

In an earlier study (Björkman and Eklundh, 2002), we have shown this model to more gracefully degrade in cases of image noise and poor feature distributions, than the more popular essential matrix model (Longuet-Higgins, 1981). The rotational (r_x, r_y, r_z) and translational (t_x, t_z) parameters are determined iteratively from matches of corner features. We use Harris' corner features (Harris and Stephens, 1988) for the purpose and apply random sampling (Fischler and Bolles, 1981) to reduce the influence from outliers in the dataset. Matching is done using normalized cross-correlation of 8×8 pixel image patches. As quality measure we use a left-to-right and right-to-left matching consistency check, rather than thresholding on matching scores. Once the above parameters are known, the metric distance to the current fixation point is computed from the baseline and the vergence angle r_y. This distance is later used as a reference for distance and scale measurements, as well as for camera control.

3.2 Wide-field to Foveal Transfer

Once an object of interest has been found through the attentional process (explained in Section 4), the cameras are directed so that the object is placed in fixation in the foveal views. This is done using affine transfer (Fairley *et al.*, 1998), which is based on the fact that if the relations between three different views are known, the position of a point given in two views can determined in the third. In our case a new fixation point is found in the wide-field views and the problem is to transfer the same point to each foveal view. To relate a foveal view position $\mathbf{x_f}$ to the corresponding wide-field position $\mathbf{x_w}$, we use the affine epipolar constraint $\mathbf{x_w^\top F_A x_f} = 0$ and the affine essential matrix

$$\mathbf{F_A} = \begin{pmatrix} 0 & 0 & a_3 \\ 0 & 0 & a_4 \\ a_1 & a_2 & a_5 \end{pmatrix} \qquad (2)$$

Here a_1, a_2, a_3 and a_4 encode the relative orientation and scale between the wide-field and foveal views, while a_5 is the difference in y-wise position between the optical centers. Similarly to wide-field calibration, the parameters are determined using feature matching of Harris' corner features (Harris and Stephens, 1988) and random sampling (Fischler and Bolles, 1981). With wide-field views related to each other using Equation (1) and foveal views to their wide-field counterparts using Equation (2), a new fixation point can be transferred from the wide-field to the foveal views. The cameras can then be moved using rapid gaze shifts, so that the new point is placed in the center of the foveal images.

3.3 Fixation

Once a transfer has been completed and a saccade (rapid gaze shift) executed towards the new attention point, the system tries to fixate onto the new region in the center of the foveal views. This fixation is kept until another region of interest has been found through the attentional system. Thus the camera control can be in either of two modes, saccade or fixation. However, since a saccade occurs in a fraction of a second, the cameras are almost always in fixation. This is beneficial to more high-level processes. With regions of interest in fixation, binocular information can be extracted, information that can be useful for segmentation, object recognition and manipulation. We will see examples of this in later sections.

The relative orientations of the left and right foveal views are constantly kept up-to-date, much like the wide-field external calibration in Section 3.1. Harris' corner features (Harris and Stephens, 1988) are extracted from both views and features are matched using random sampling (Fischler and Bolles, 1981). The cameras are then related by an affine essential matrix $\mathbf{F_A}$, similar to the one used for wide-field to foveal transfer in Equation (2). Even if $\mathbf{F_A}$ is just an approximation of a general essential matrix, it is applicable to our case, since focal lengths are large and views narrow. Binocular disparities are measured along the epipolar lines and the vergence angle of the stereo head is controlled such that the highest density of points are placed at zero disparity. For temporal filtering we use Kalman filters, but ignore frames for which not enough matches are found.

4 Bottom-Up and Top-Down Attention

The best way of viewing attention in the context of a robotic system is as a selection mechanisms serving the higher level tasks such as object recognition and manipulation. Biological systems may provide a good basis for solving some of the modeling issues. However, due to computational issues mentioned earlier, these studies serve as a mere inspirational source and should not be restricting the computational implementation. We know that humans tend to perform a subconscious ranking of the "interestingness" of the different components of a visual scene. This ranking depends on the observers goals as well as the components of the scene; how the components in the scene relate to their surroundings (Bottom-up) and to our objectives (Top-down) (Itti, 2000; Li, 2002). In humans, the attended region is then selected through dynamic modifications of cortical connectivity or through the establishment of specific temporal patterns of activity, under both Top-down (task dependent) and Bottom-up (scene dependent) control (Olshausen *et al.*, 1993). In this work we will define the Top-down information as consisting of two components: 1) task dependent information which is usually volitional, and 2) contextual scene dependent information.

We propose a simple, yet effective, Artificial Neural Network approach that learns the optimal bias of the Top-down saliency map (Koch and Ullman, 1985). The most novel part of the approach is a dynamic combination of the Bottom-up and Top-down saliency maps. Here an information measure (based on entropy measures) indicates the importance of each map and thus how the linear combination should be altered over time. The combination will vary over time and will be governed by a differential equation. Together with a mechanism for Inhibition-of-Return, this dynamic system manages to adjust itself to a balanced behavior, where neither Top-down nor Bottom-up information is ever neglected.

4.1 Biased Saliency for Visual Search Tasks

Current models of how the attentional mechanism is incorporated in the human visual system generally assume a Bottom-up, fast and primitive mechanism that biases the observer toward selecting stimuli based on their saliency (most likely encoded in terms of center-surround mechanisms) and a second slower, Top-down mechanism with variable selection criteria, which directs the 'spotlight of attention' under cognitive, volitional control (Treisman and Gelade, 1980). In computer vision, attentive processing for scene analysis initially largely dealt with salience based models, following (Treisman and Gelade, 1980) and the influential model of Koch and Ullman (1985). However, several computational approaches to selective attentive processing that combine Top-down and Bottom-up influences have been presented in recent years.

Koike and Saiki (2002) propose that a stochastic Winner-Take-All (WTA) enables the saliency based search model to cause the variation of the relative saliency to change search efficiency, due to stochastic shifts of attention. Ramström and Christensen (2004) calculate feature and background statistics to be used in a game theoretic WTA framework for detection of objects. Choi *et al.* (2004) suggest learning the desired modulations of the saliency map (based on the model by Itti *et al.* (1998)) for Top-down tuning of attention. Navalpakkam and Itti (2003) enhance the Bottom-up salience model to yield a simple, yet powerful architecture to learn target objects from training images containing targets in diverse, complex backgrounds. Earlier versions of their model did not learn object hierarchies and could not generalize, although the current model could do that by combining object classes into a more general super class.

Lee *et al.* (2003) showed that an Interactive Spiking Neural Network can be used to bias the Bottom-up processing towards a task (in their case in face detection). However, their model was limited to the influence of user provided Top-down cues and could not learn the influence of context. In Frintrop's VOCUS-model (Frintrop, 2006) there are two versions of the saliency map; a Top-down map and a Bottom-up one. The Bottom-up map is similar to that of Itti and Koch's, while the Top-down map is a tuned version of the Bottom-up one. The total saliency map is a linear combination of the two maps using a fixed user provided weight. This makes the combination rigid and non flexible, which may result in loss of important Bottom-up information. Oliva *et al.* (2003) show that Top-down information from visual context can modulate the saliency of image regions during the task of object detection. Their model learns the relationship between context features and the location of the target during past experience in order to select interesting regions of the image. Many of the computational models study the attention mechanism itself but there have also been approaches that demonstrate robotic applications.

In the work of Breazeal and Scassellati (1999) a computational implementation of the visual search model described by Wolfe (1994), is created. They use this system on a robotic platform where they integrate perception with inhibition of return and other internal effects. The result is a context-dependent attention map they use to determine the gaze direction of the robot. On their humanoid platform, Vijayakumar *et al.* (2001) explored the integration of oculomotor control (biologically inspired) with visual attention. Their attention module consisted of a neural network implementing a WTA-network (Tsotsos *et al.*, 1995), in which different visual stimuli could compete for shift of gaze in their direction. Nickerson *et al.* (1998) created within the framework of the ARK project, mobile robots that used attention-based space segmentation for navigating within industrial environments without the need for artificial landmarks. They too used the WTA-network of (Tsotsos *et al.*, 1995). Clark and Ferrier (1992) suggested early on an robotic implementation of a visual control system based on models of the human oculomotor control. They did that by attentive control through specifications of gains in parallel feedback loops. Their modal control was through saliency map calculated as weighted combination of several feature maps. In a recent work, Siagian and Itti (2007) use the attention system of Itti *et al.* (1998) in combination with a "gist" model of the scene, to direct an outdoor robot toward the most likely candidate locations in the image, thus making the time-consuming process of landmark identification more efficient. In their human-machine interaction system, Heidemann *et al.* (2004) recognize hand gestures in parallel with computing context-free attention maps for the robot. Allowing an interaction between the human and the robot where, according to the authors, a balanced integration of bottom-up generated feature maps and top-down recognition is made. One of the few recent works that incorporates a computational mechanism for attention into a humanoid platform is the work of Moren *et al.* (2008). A method called Feature Gating is used to achieve Top-down modulation of saliencies.

Our framework is based on the notion of saliency maps (SMs), (Koch and Ullman, 1985). To define a Top-down SM, $SM_{TD}(t)$, t denoting time, we need a preferably simple search system based on a learner that is trained to find objects of interest in cluttered scenes. In parallel, we apply an unbiased version of the same system to provide a Bottom-up SM, $SM_{BU}(t)$. In the following we will develop a way of computing these two types of maps and show that it is possible to define a dynamic active combination where neither one always wins, i.e. the system never reaches a static equilibrium, although it sometimes reaches dynamic one. The model (illustrated in Fig. 4) contains a standard Saliency Map (SM_{BU}) and a Saliency Map biased with weights (SM_{TD}). The Top-down bias is achieved by weight association (our Neural Network). An Inhibition-of-Return mechanism and stochastic WTA-network gives the system its dynamic behavior described in Section 4.3. Finally the system combines $SM_{BU}\,(t)$ and $SM_{TD}\,(t)$ with a linear combination that evolves over time t. Our model applies to visual search and recognition in general, and to cases in which new visual information is acquired in particular.

Several computational models of visual attention have been described in the literature. One of the best known systems is the *Neuromorphic Vision Toolkit* (NVT), a derivative of the model by Koch and Ullman (1985) that was (and is) developed by the group around Itti et al. (Itti and Koch, 2001; Itti *et al.*, 1998; Navalpakkam and Itti, 2003). We will use a slightly modified version of this system for our computations of saliency maps. Some limitations of the NVT have been demonstrated, such as the non robustness under translations, rotations and reflections, shown by Draper and Lionelle (2003). However, our ultimate aim is to develop a system running on a real time active vision system and we therefore seek to achieve a fast computational model, trading off time against precision. NVT is suitable in that respect. Similarly to Itti's original model, we use color, orientation and intensity features, with the modification that we have complemented these with a texture cue that reacts to the underlying texture of regions, not to outer contours. The details of how this texture cue, based on the eigen-values of small patches in the image, are calculated can be found in

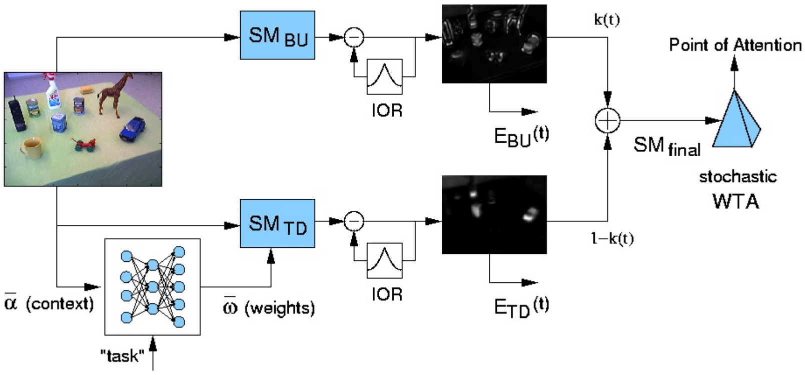


Figure 4: An attentional model that combines Bottom-up and Top-down saliency, with Inhibition-of-Return and a stochastic Winner-Take-All mechanism, with context and task dependent Top-down weights.

(Rasolzadeh *et al.*, 2007).

4.2 Weight Optimization and Contextual Learning

As mentioned above we base both Top-down and Bottom-up salience on the same type of map. However, to obtain the Top-down version we bias this conspicuity map. In our approach, which otherwise largely follows Frintrop (2006), the weighting is done differently. This has important consequences, as it will be shown later. The four broadly tuned color channels R, G, B and Y, all calculated according to the NVT-model, are weighted with the individual weights (ω_R, ω_G, ω_B, ω_Y). The orientation maps (O_{0°, O_{45°, O_{90°, O_{135°) are computed by Gabor filters and weighted with similar weights (ω_{0°, ω_{45°, ω_{90°, ω_{135°) in our model. Following the original version, we then create scale pyramids for all 9 maps (including the intensity map I) and form conventional center-surround-differences by across-scale-subtraction and apply Itti's normalization operator. [2] This leads to the final conspicuity maps for intensity $(\bar{I})$, color $(\bar{C})$, orientation $(\bar{O})$ and texture $(\bar{T})$. As a final set of weight parameters we introduce one weight for each of these maps, (ω_I, ω_C, ω_O, ω_T). To summarize the calculations:

$$RG(c,s) = |(\omega_R \cdot R(c) - \omega_G \cdot G(c)) \ominus (\omega_R \cdot R(s) - \omega_G \cdot G(s))|$$
$$BY(c,s) = |(\omega_B \cdot B(c) - \omega_Y \cdot Y(c)) \ominus (\omega_B \cdot B(s) - \omega_Y \cdot Y(s))|$$
$$O_\theta(c,s) = \omega_\theta \cdot |O_\theta(c) \ominus O_\theta(s)|$$
$$\bar{C} = \bigoplus_c \bigoplus_s N(RG(c,s)) - N(BY(c,s))$$
$$\bar{O} = \sum_\theta N(\bigoplus_c \bigoplus_s N(O_\theta(c,s)))$$

[2] The center-surround-differences are a computational model of the center-surround receptive fields composed by ganglion cells in the retina. For details on the across-scale subtraction we refer to Itti's original work.

$$\bar{I} = \bigoplus_c \bigoplus_s N(|I(c) \ominus I(s)|)$$
$$\bar{T} = \bigoplus_c \bigoplus_s N(|T(c) \ominus T(s)|)$$
$$SM_{TD} = \omega_I \bar{I} + \omega_C \bar{C} + \omega_O \bar{O} + \omega_T \bar{T}$$

Here $\ominus$ denotes the across-scale-subtraction, $\bigoplus$ the across-scale-summation. The center scales are $c \in \{2, 3, 4\}$ and the surround scales $s = c + \delta$, where $\delta \in \{3, 4\}$ as proposed by Itti and Koch. We call the final modulated saliency map the Top-down map, SM_{TD}. The Bottom-up map, SM_{BU} can be regarded as the same map with all weights being 1.

As pointed out by Frintrop, the number of introduced weights in some sense represents the degrees of freedom when choosing the "task" or the object/region to train on. A relevant question to pose is: how much "control" do we have over the Top-down map by changing the weights? As previously stated, we divide Top-down information in two categories; i) task and ii) context information. To tune and optimize the weight parameters of the SM for a certain task, we also have to examine what kind of context information is important. For instance, the optimal weight parameters for the same task typically differ from one context to the other. These two issues will be considered in the remaining part of the section.

4.2.1 Optimizing for the ROI

First we need to formalize the optimization problem. For a given Region Of Interest (ROI) characteristic for a particular object, we define a measure of how the Top-down map differs from the optimum as:

$$e_{ROI}(\bar{\omega}) = \frac{\max SM_{TD} - \max\left(SM_{TD}|_{ROI}\right)}{\max SM_{TD}}$$

where $\bar{\omega} = (\omega_I, \omega_O, \omega_C, \omega_T, \omega_R, \omega_G, \omega_B, \omega_Y, \omega_{0^\circ}, \omega_{45^\circ}, \omega_{90^\circ}, \omega_{135^\circ})$ is the weight vector. The optimization problem will then be given by $\bar{\omega}_{opt} = \arg\min e_{ROI}(\bar{\omega})$. $\bar{\omega}_{opt}$ maximizes peaks within the ROI and minimizes peaks outside ROI. This optimization problem can be solved with the Levenberg-Marquardt algorithm. The optimal set of weights (the optimal weight-vector) are thus obtained. With this set of weights, we significantly increase the probability of the winning point being within a desired region. To summarize; given the task to find a certain (type) of ROI we are able to find a good set of hypotheses by calculating the Top-down map $SM_{TD}(\bar{\omega}_{opt})$. The method used to do this optimization for a given ROI, is described in (Rasolzadeh *et al.*, 2006).

4.2.2 Learning Context with a Neural Network

The weight optimization above is in principle independent of context. In order to include the correlation between the optimal weights and the context (environmental Top-down information), we have to know both types of Top-down information (context and task) in order to derive the set of optimal weights as a function of context and task.

There are a large number of different definitions of context in the computer vision literature (Rabinovich *et al.*, 2007; Strat and Fischler, 1989, 1995). In our model we will keep the definition simple enough to serve our purpose of visual search. A simple example is that a large weight on the red color channel would be favorable when searching for a red ball on a green lawn. However, the same weighting would not be appropriate when searching for the same ball in a red room! We therefore represent context by the total energy of each feature map, in our case a 11-dimensional contextual vector, here denoted as $\bar{\alpha}$. This will give us a notion of "how much" of a certain feature we have in the environment, and thus how discriminative that feature will be for a visual search task.

Obviously we cannot perform this non-convex (time-consuming) optimization every time we need to find the optimal weights (minimizing $e_{ROI}(\bar{\omega})$), in order to find a ROI, i.e. have maximal saliency within the ROI. Instead, we collect the optimized weights and the corresponding contextual vectors for a large set of examples. Given that data set, we train an artificial neural network (Haykin, 1994) to associate between the two: i.e. given a contextual vector, what will the optimal set of weights be like. This is performed for each type of ROI, thus there will be one trained neural network (NN) for each object. Each of these NNs can automatically correlate the context information with the choice of optimal weight parameters *without* any optimization. Fore more details on how this training is done we refer to our previous works (Rasolzadeh *et al.*, 2006).

4.3 Top-Down / Bottom-Up Integration

So far we have defined a Bottom-up map $SM_{BU}(t)$ representing the unexpected feature based information flow and a Top-down map $SM_{TD}(t)$ representing the task dependent contextual information. We obtain a mechanism for visual attention by combining these into a single saliency map that helps us to determine where to "look" next.

In order to do this, we rank the "importance" of saliency maps, using a measure that indicates how much value there is in attending that single map at any particular moment. We use an energy measure (E-measure) similar to that of Hu et al, who introduced the *Composite Saliency Indicator* (CSI) for similar purposes (Hu *et al.*, 2004). In their case, however, they applied the measure on each individual feature map. We use the same measure, yet we use it on the summed up saliency maps. The Top-down and Bottom-up energies E_{TD} and E_{BU} are defined as the density of saliency points divided by the convex hull of all points.

Accordingly, if a particular map has many salient points located in a small area, that map might have a higher E-value than one with even more salient points, yet spread over a larger area. This measure favors saliency maps that contain a small number of very salient regions.

4.3.1 Combining SM_{BU} and SM_{TD}

We now have all the components needed to combine the two saliency maps. We may use a regulator analogy to explain how. Assume that the attentional system contains several (parallel) processes and that a constant amount of processing power has to be distributed among these. In our case this means that we want to divide the attentional power between $SM_{BU}(t)$ and $SM_{TD}(t)$. Thus the final saliency map will be a linear combination

$$SM_{final} = k \cdot SM_{BU} + (1-k) \cdot SM_{TD}.$$

Here the k-value varies between 0 and 1, depending on the relative importance of the Top-down and Bottom-up maps, according to the tempo-differential equation

$$\frac{dk}{dt} = -c \cdot k(t) + a \cdot \frac{E_{BU}(t)}{E_{TD}(t)} \quad , \quad 0 \leq k \leq 1$$

The two parameters c and a, both greater than 0, can be viewed as the amount of *concentration* (devotion to search task) and the *alertness* (susceptibility for Bottom-up info) of the system. The above equation is numerically solved between each attentional shift.

The first term represents an integration of the second one. This means that a saliency peak needs to be active for a sufficient number of updates to be selected, making the system less sensitive to spurious peaks. If the two energy measures are constant, k will finally reaches an equilibrium at aE_{BU}/cE_{TD}. In the end, SM_{BU} and SM_{TD} will be weighted by aE_{BU} and $\max(cE_{TD} - aE_{BU}, 0)$ respectively. Thus the Top-down saliency map will come into play, as long as E_{TD} is sufficiently larger than E_{BU}. Since E_{TD} is larger than E_{BU} in almost all situations when the object of interest is visible in the scene, simply weighting SM_{TD} by E_{TD} leads to a system dominated by the Top-down map.

The dynamics of the system comes as a result of integrating the combination of saliencies with Inhibition-of-Return (IOR). The kind of IOR we talk about here is in a covert mode, where the eyes or the head are not moving at all (overt shifts) and there is essentially only a ranking of the various saliency peaks within the same view. Of course, if the desired object is not found within the current set of salient points, the system will continue to an overt shift where the head and the eyes focus on a different point in space. If a single salient Top-down peak is attended to (covertly), saliencies in the corresponding region will be suppressed, resulting in a lowered E_{TD} value and less emphasis on the Top-down flow, making Bottom-up information more likely to come into play. However, the same energy measure will hardly be affected if there are many salient Top-down peaks of similar strength. Thus the system tends to visit each Top-down candidate before attending to purely Bottom-up ones. This, however, depends on the strength of each individual peak. Depending on *alertness*, strong enough Bottom-up peaks could just as well be visited first. The motivation for a balanced combination of the two saliency-maps based on two coefficients

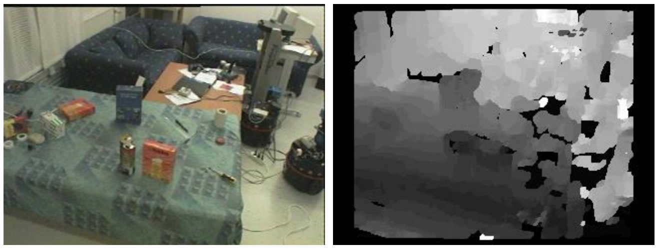


Figure 5: Disparity map (right) of a typical indoor scene (left).

(*alertness* and *concentration*) comes from neuroscientific studies on the dorsal and ventral pathways. In a larger system, this attentive mechanism we propose here can thus easily be integrated with the rest of the reasoning system, with the dorsal stream activity indicating a pragmatic mode, whereas the the ventral stream activity indicates a semantic mode.

4.4 Binocular Cues for Attention

Since the attentional system described above is generic with respect to the visual task, it may just as well deliver regions of interest corresponding to things that are either too large or too far away to be manipulated. It is clear that in our scenario, size and distance needs to be taken into account for successful interaction with the environment. Now, even if the projective size of a region can be measured, its real-world size is unknown, since the projective size depends on the distance from the camera set. One of the benefits of a binocular system, such as the one described in Section 3, is that sizes and distances can be made explicit. Therefore, we complement the attentional system with binocular information in order to make the system more likely to pop-out regions of interest suitable for manipulation.

With wide-field cameras calibrated as described in Section 3.1 disparity maps, such as the one to the right in Fig. 5, are computed. Disparity maps encode distances to 3D points in the scene. A point distance is given by $Z = bf/d$, where b is the baseline (the distance between the cameras), f is the focal length and d the respective binocular disparity. Before a peak is selected from the saliency map, the saliency map is sliced up in depth into a set of overlapping layers, using the disparity map. Each layer corresponds to saliencies within a particular interval in depth. A difference of Gaussian (DoG) filter is then run on each layer. The sizes of these filters are set to that of the expected projected sizes of manipulable things. Thus for saliency layers at the distance the DoGs are smaller than for layers closer to the camera head. As a result you will get saliency peaks similar to those in Fig. 6, with crosses indicating the expected size of things in the scene.

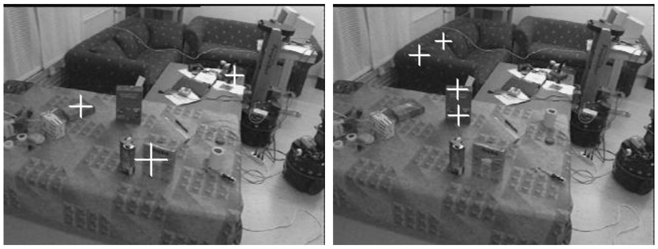


Figure 6: Saliency peaks with saliency maps computed using top-down tuning for the orange package (left) and the blue box (right). The crosses reflect the sizes derived from the attentional process.

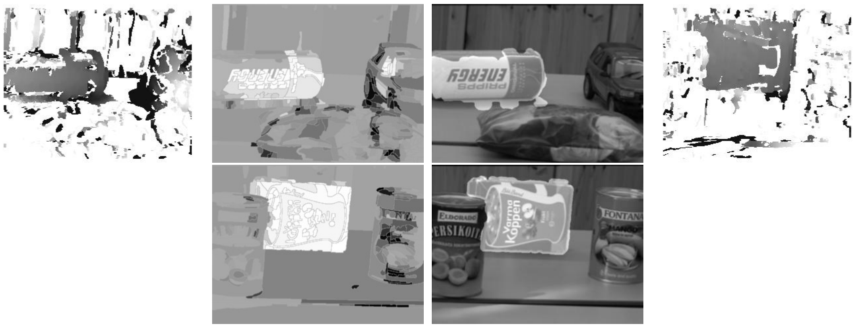


Figure 7: Disparity maps (right), prior foreground probabilities (middle) and posteriori figure-ground segmentation (left).

5 Foveated Segmentation

After a region of interest has been selected by the attentional system, the camera system is controlled such that the region is placed at zero disparity in the center of the foveal views. It is now ready to be identified and possibly manipulated. Before this is done, it would be beneficial if it could also be segmented from other nearby objects in the scene. Both recognition and pose estimation are simplified if the object is properly segmented. In our system we do this with the help of binocular disparities extracted from the foveal views.

In the foveated segmentation step, the foreground probability of each pixel is computed in a probabilistic manner. From area based correlation we estimate a measurement for each pixel, that are then used to estimate the prior probability of a pixel belonging to the foreground. Examples of foreground priors can be seen in Fig. 7 (middle).

By modeling the interaction between neighboring patches and computing the posteriori foreground probabilities using graph-cuts, pixels are finally labeled as being part of either the *foreground* or *background*. Since there are only two possible labels, the exact posteriori solution is given in a single graph-cut operation

(Grieg *et al.*, 1989). The resulting segmentation may look like the two images in Fig. 7 (right). These segmentations are then passed to either recognition or pose estimation. For more information on the precise modeling and motivations see (Björkman and Eklundh, 2006).

5.1 From 3D Segments to Shape Attributes

In order to have segmentation that is appropriate for manipulation image data needs to be grouped into regions corresponding to possible objects in the 3D scene. Disparities can be considered as measurements in 3D space, clustered around points of likely objects. These clusters are found by applying a kernel-based density maximization method, known as Mean Shift (Comaniciu and Meer, 2002). Clustering is done in image and disparity space, using a 3D Gaussian kernel with a size corresponding to the typical 3D size of objects that can be manipulated. The maximization scheme is iterative and relies on initial center point estimates. As such estimates we use the hypotheses from the attentional system. Examples of segmentation results using this approach can be seen in the second row of Fig. 9.

One major drawback of the mean shift algorithm is the fact that an object can not be reliably segregated from the surface it is placed on, if there is no evidence supporting such a segregation. Without any additional assumption on surface shape or appearance there is no way of telling the surface from the object. However, in many practical scenarios (including ours) it might be known to the robotic system that objects of interest can in fact be expected to be located on flat surfaces, such as table tops.

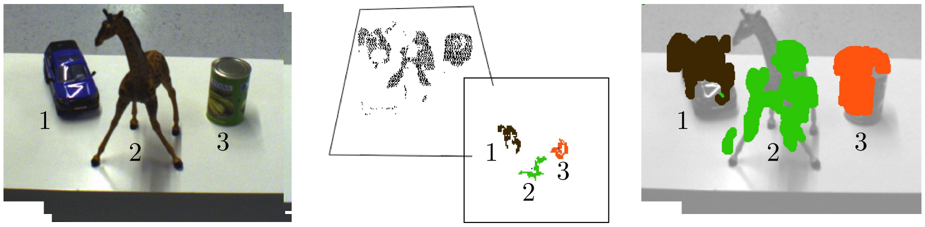
We therefore complement our approach with a table plane assumption. Using a textured surface, it is possible to find the main plane and cut it with a 3D version of the Hough transform as in (Huebner *et al.*, 2008a). Following the table assumption the 3D points are mapped onto a 2D grid to find segments and basic shape attributes.

The result of transformation and clipping on the scene given in Fig. 8(a) can be seen in Fig. 8(b). The segmentation of objects is computed on the 2D grid (Fig. 8(c)) with a simple region growing algorithm grouping pixels into larger regions by expanding them bottom up. Since the grid is thereby segmented, simple shape-based attributes of each segment can be determined and the segments reprojected to 3D points or to the image plane (illustrated in Fig. 8(d)) [3].

5.2 Associated Attributes

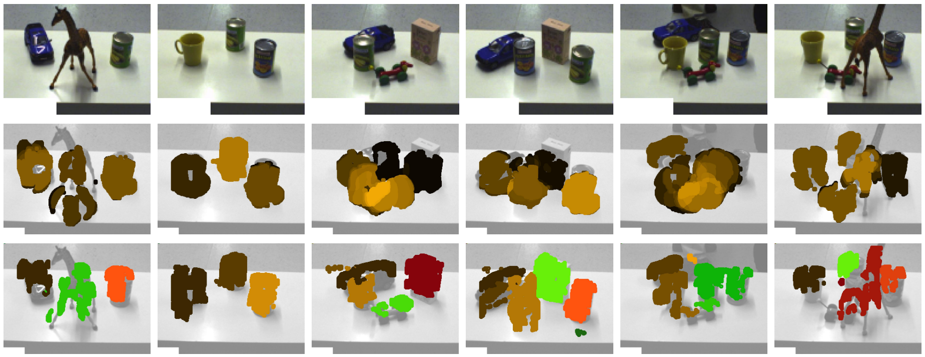
The generated segments are just *things*, as the step to an *object* longs for semantics. One way to identify the semantics of a thing in order to derive an object

[3] Note that dilation has been applied for the reprojected segments for the later application of point-based object hypotheses verification. The dilation, the grid approach, as also noisy and incomplete data from stereo causing that reprojections are often little larger or not completely covering the bodies.



(a) Stereo images (b) 3D points (c) 2D segmentation (d) Reprojection

Figure 8: Segmentation using the table plane assumption. Disparity information from the stereo images (a) produces 3D points (b). Having defined the dominant plane, the points can be projected onto this plane, where distinctive segments are computed (c) and reprojected to the image (d).



(a) (b) (c) (d) (e) (f)

Figure 9: Sample scenario segmentation (best viewed in color). Original images are shown in the first row. The second row shows results using the Mean Shift segmentation, the bottom row those using the table plane assumption (mentioned in Section 5.1). In the latter, (a) and (b) seem well segmented and in (c) there is just some noise at the table edge. Problems arise for (d)-(f): (d) two segments for the car, (e) one segment for two cans, and (f) the dog underneath the giraffe is not detected.

is to associate attributes to it. The attributes can be of two kinds, *intrinsic* and *extrinsic*. Intrinsic attributes are object-centered and thereby theoretically viewpoint-independent (e.g. shape, color, mass). Extrinsic attributes describe the viewpoint-dependent state of an object (e.g. position, orientation). In our system, the basic intrinsic attributes of covered area, length (along the dominant axis), width (perpendicular to the dominant axis) and height can be qualitatively determined for each segment. The discretization, i.e. if an object is *small* or *large* in size, is adapted to our table-top manipulation scenario at hand. Additionally, the centroid position of a segment is calculated. Its 3D point cloud is kept available for the subsequent operations, e.g. pose estimation (as we will show later in Section 6.2) or shape approximation and grasping, as we proposed in (Huebner *et al.*, 2008b).

6 Object Manipulation

To achieve real cognitive capabilities, robotic systems have to be able to interact with the surrounding. Grasping and manipulation of objects is one of the basic building blocks of such a system. Compared to humans or primates, the ability of today's robotic grippers and hands is surprisingly limited and their dexterity cannot be compared to human hand capabilities. Contemporary robotic hands can grasp only a few objects in constricted poses with limited grasping postures and positions.

Grasping, as a core cognitive capability, has also been posed as one of the key factors of the evolution of the human brain. This is founded in convergent findings of brain researchers. For example, 85% of axons in visual cortex do not come from the retina, but other brain areas including what is thought to be higher brain regions (Sigala and Logothetis, 2002). Lately, anatomical and physiological investigations in non human primates, together with brain imaging studies in humans, have identified important cortical pathways involved in controlling visually guided prehension movements. In addition, experimental investigations of prehension movements have begun to identify the sensorimotor transformations and representations that underlie goal directed action. It has been shown that attentional selection of the action related aspects of the sensory information is of considerable importance for action control, (Castiello, 2005; Riddoch *et al.*, 2001). When a grasp is being prepared, the visual system provides information about the egocentric location of the object, its orientation, form, size, and the relevant environment. Attention is particularly important for creating a dynamic representation of peripersonal space relevant for the specific tasks.

Regarding implementation in robots, grasp modeling and planning is difficult due to the large search space resulting from all possible hand configurations, grasp types, and object properties that occur in regular environments. The dominant approach to this problem has been the model based paradigm, in which all the components of the problem (object, surfaces, contacts, forces) are modeled according to physical laws. The research is then focused on grasp analysis, the study of the physical properties of a given grasp; and grasp synthesis, the computation of grasps that meet certain desirable properties, (Bicchi and Kumar, 2000; Coelho Jr. *et al.*, 1998; Namiki *et al.*, 2003; Platt Jr. *et al.*, 2002; Shimoga, 1996). More recently, it has been proposed to use vision as a solution to obtain the lacking information about object shapes or to use contact information to explore the object (Kragic *et al.*, 2005; Morales *et al.*, 2001; Platt Jr. *et al.*, 2002). Another trend has focused on the use of machine learning approaches to determine the relevant features that indicate a successful grasp (Coelho *et al.*, 2001; Kamon *et al.*, 1998; Morales *et al.*, 2004). Finally, there have been efforts to use human demonstrations for learning grasp tasks (Ekvall and Kragic, 2004).

One of the unsolved problems in robot grasping is grasping of unknown objects in unstructured scenarios. For general settings, manipulation of unknown objects has almost not been pursued in the literature and it is commonly as-

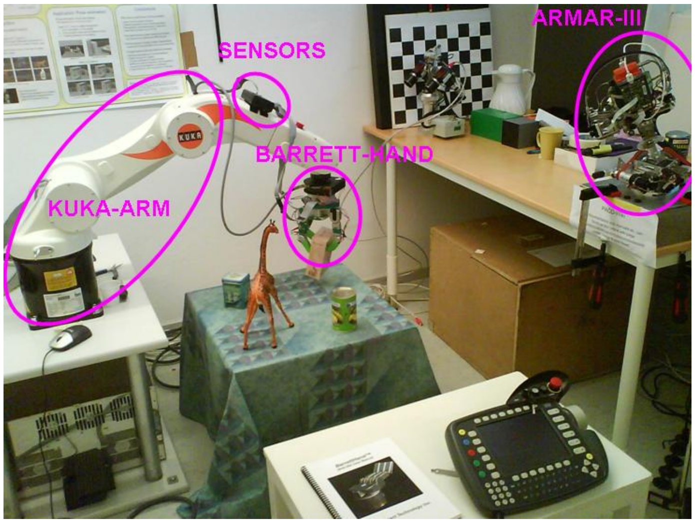


Figure 10: Our robotic setup.

sumed that objects are easy to segment from the background. In the reminder of this section, we concentrate on an example of how the presented visual system can be used to provide grasping hypotheses of objects for which the identity/geometry is not known in advance. We acknowledge that this approach is not valid in all situations, but it is one of the possible directions to pursue.

6.1 Experimental Platform

Our robotic setup consist of the Armar-III humanoid head described in Section 2.1, a BH8-series BarrettHand mounted on a KUKA KR5 R850 6-DOF robot, Fig. 10. The hand-yye calibration is performed using the classical approaches of (Shiu and Ahmad, 1989; Tsai and Lenz, 1988). The integration of the different parts of this robotic platform is achieved using a modularized software system; containing interacting modules for frame grabbing, camera calibration, visual front end modules, head control, arm control, hand control and sensory reading. Modules are implemented as CORBA processes that run concurrently and generally on different machines.

6.2 Model-Free Manipulation

In general, we will not have a precise geometrical model for all objects the robot is supposed manipulate. One a new object hypothesis is made based on the visual modules described so far, different attributes may be attached to it. These attributes are intrinsic (size, shape) and extrinsic (pose) and are stored as a part of object representation for later indexing. We refer to (Huebner *et al.*, 2008a) for more details.

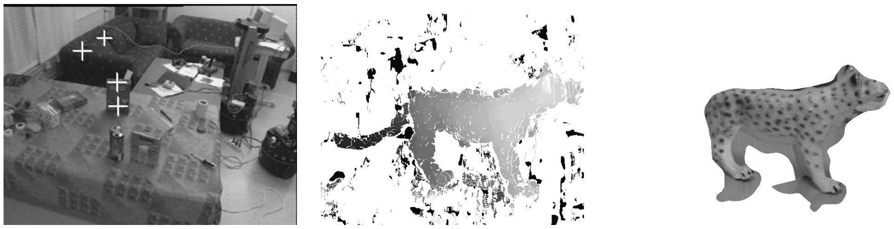


Figure 11: Left) A left manipulation camera image, Middle) The corresponding disparity map, Right) Segmentation from mean shift in 3D space.

Before the 3D position of an object, as well as its orientation can be determined, it has to be segmented from its surrounding, which in our system is done using a dense disparity map as explained in Section 5, and exemplified by the images in Fig. 11. In the current system, we thus use the generated object hypotheses in combination with the orientation estimation described below, to apply top-grasps on the objects. Given the segmentation (with table-plane assumption), and 3D coordinates, a plane is mapped to the 3D coordinates of all points within the segmented object. Since only points oriented toward the cameras are seen, the calculated orientation tends to be somewhat biased toward fronto-parallel solutions. However, the BarrettHand is able to tolerate some deviations from a perfectly estimated orientation. With the 3D points denoted by $\mathbf{X}_i = (X_i, Y_i, Z_i)^\top$, we iteratively determine the orientation of a dominating plane using a robust M-estimator. The normal of the plane at iteration k is given by the least eigenvector $\mathbf{c}_k$ of

$$\mathbf{C}_k = \sum_i \omega_{i,k}(\mathbf{X}_i - \bar{\mathbf{X}}_\mathbf{k})(\mathbf{X}_i - \bar{\mathbf{X}}_k)^\top, \tag{3}$$

where the weighted mean position is $\bar{\mathbf{X}}_\mathbf{k}$.

Points away from the surface are suppressed through the weights

$$\omega_{i,k} = t^2/(t^2 + \delta_{i,k}^2), \tag{4}$$

where $\delta_{i,k} = \mathbf{c}_{k-1}^\top(\mathbf{X}_i - \bar{\mathbf{X}})$ is the distance from the point $\mathbf{X}_i$ to the plane of the previous iteration. Here t is a constant reflecting the acceptable variation in flatness of the surface and is set to about a centimeter. More details on the implementation can be found in (Kragic *et al.*, 2005).

7 Experimental Results

The following sections present several experiments related to the different aspects of the vision system. Section 7.1 presents qualitative and quantitative experiments of the attention system, such as the weight optimization process and the neural network learning. Section 7.2 presents results on the object recognition module and Section 7.3 gives an example of the integrated modules solving an object detection and manipulation task.

7.1 Top-Down and Bottom-Up Attention

As described in Section 4, our attentional model consists of three main modules:

- The optimization of Top-down weights (offline);
- The Neural Network which associates context and weight (online); and
- The dynamical combination of SM_{BU} and SM_{TD}.

The experiments presented below are designed to show how these different modules affect the performance of the model. We present results from the experiments on the contextual learning, since it is the most crucial part for our visual search tasks. In Fig. 12 (top), the ten objects used in the experiments are shown. These are all associated with a set of intrinsic attributes that consist of 3D size, appearance, and feasible grasps. To represent the appearance, we use SIFT descriptors (Lowe, 1999) and color histograms (Gevers and Smeulders, 1999). Detection experiments using these can be found in Section 7.2. The graph in Fig. 12 shows the Top-down (TD) weights deduced for the four cues from one particular image. The cues with high weights for most materials are color and texture. We can see that some cues are almost completely suppressed for several objects. The resulting set of triplets $\{ROI, \bar{\omega}_{opt}, \bar{\alpha}\}$ were used for training the neural networks.

7.1.1 Weight Optimization

The non-convex optimization, solved with the Levenberg-Marquardt method, maximizes the saliency value SM_{TD} within the RIO. RIO represents the desired target object and the process of optimization is based on manipulating the weight-vector $\bar{\omega}$. However, it is important to note that, even if one may reach a global minimum in the weight optimization, it does not necessarily mean that our Top-down map is "perfect", as in Fig. 13. In other words, the Top-down map may not rank the sought ROI the highest, in spite of $e_{ROI}(\bar{\omega})$ being at its global minimum for that specific image and object. What this implies is that for some objects $\min[e_{ROI}(\bar{\omega}_{opt})] \neq 0$, or simply that our optimization method failed to find a proper set of weights for the Top-down map at the desired location as, for example, in Fig. 14.

Another observation worth mentioning is the fact that there may be several global optima in weight space each resulting in different Top-down maps. For example, even if there exists many linear independent weight vectors $\bar{\omega}_i$ for which $e_{ROI}(\bar{\omega}_i) = 0$, the Top-down maps $SM_{TD}(\bar{\omega}_i)$ will in general be different from one another (with different E_{CSI}-measure).

7.1.2 Artificial Neural Network (ANN) Training

When performing the pattern association on the neural network that is equivalent with context learning, it is important that the training data is "pure". This means that only training data that gives the best desired result should be

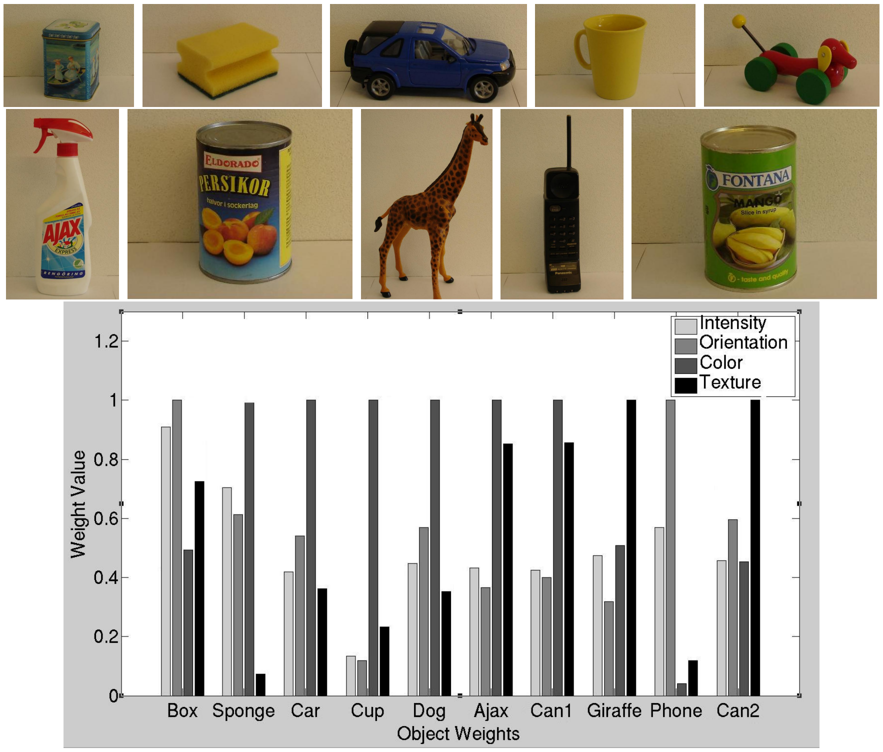


Figure 12: A set of objects used for experiments (left) and the four TD-weights $\bar{\omega}_I, \bar{\omega}_O, \bar{\omega}_C, \bar{\omega}_T$ for each object in one particular image (right).

included. Thus only examples $\{ROI, \bar{\omega}_{opt}, \bar{\alpha}\}$ where $e_{ROI}(\bar{\omega}_{opt}) = 0$ were used. To examine the importance of our context information we created another set of NNs trained without any input, i.e. simple pattern learning. For the NN calculations this simply leads to an averaging network over the training set $\{ROI, \bar{\omega}_{opt}\}$. Quantitative results of these experiments are shown in Fig. 15. There were from a training set of 96 images taken of 10 different objects on 4 different backgrounds (table-cloths) in two different illumination situations. In each of the 96 images, the location of each of the 10 objects is annotated, thus yielding 960 annotated locations (ROIs). See database online (Rasolzadeh, 2006). Results using optimized weights (last row) in some sense represent the best performance possible, whereas searches using only the Bottom-up map perform the worst. One can also observe the effect of averaging (learning weights without context) over a large set; you risk to *always* perform poor, whereas if the set is smaller you may at least manage to perform well on the test samples that resemble some few training samples. Each NN had the same structure, based on 13 hidden neurons, and was trained using the same number of iterations. Since all weights (11) can be affected by all context components (9) and

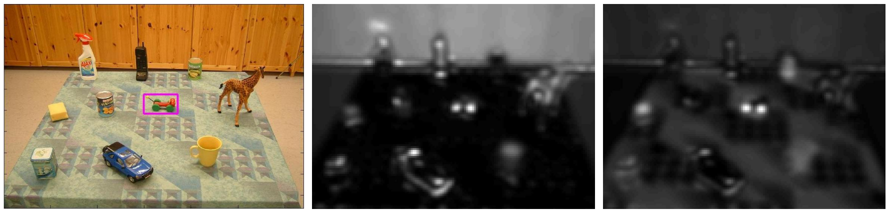


Figure 13: An example of successful optimization; the ROI is marked in the left image. Without optimization (unitary weights) the saliency map is purely Bottom-up (middle). However, an optimization that minimizes $e_{ROI}(\bar{\omega})$ (in this case to 0) the optimal weight vector $\bar{\omega}_{opt}$ clearly ranks the ROI as the best hypothesis of the Top-down map (right).

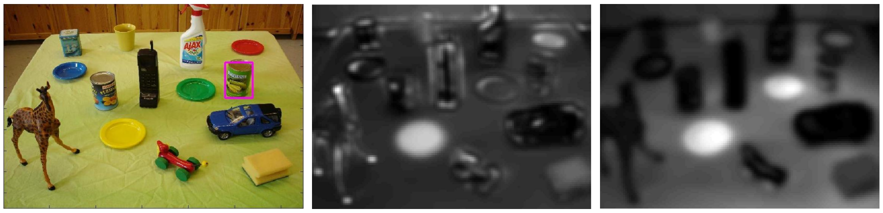


Figure 14: An example of poor optimization; although the optimization may reach a global minimum for $e_{ROI}(\bar{\omega})$ (in this case >0) the optimal weight vector $\bar{\omega}_{opt}$ *doesn't* rank the ROI as the best hypothesis of the Top-down map (right).

since each weight can be increased, decreased or neither, a minimum number of 12 hidden units is necessary for good learning.

7.2 Multi-Cue Object Detection and Hypotheses Validation

Relying on a single object detection method is difficult due to a large variety of objects a robot is expected to manipulate is difficult. Using a combinations of methods seems therefore as a suitable strategy. Without providing an extensive study of all possible methods and combinations, we give an example that shows the benefit of foveated segmentation and multiple cues object recognition. For this purpose, we have selected two methods that show different strengths and weaknesses. The first method is based on color histograms (Gevers and Smeulders, 1999) and the other on scale and rotation invariant SIFT features (Lowe, 1999). Histogram based methods are suited for both uniformly colored and textured objects, but tend to be problematic when objects are not easy to distinguish from the background. Feature based method, work well in cluttered environments, but break down when too few features are extracted due to limited texture.

We selected a set of 24 objects, similar to those in Fig. 12. We performed

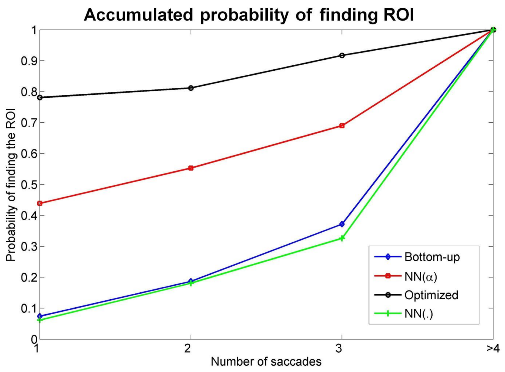


Figure 15: The estimated accumulated probability of finding the ROI. The results were averaged over the entire test set of objects(ROI:s). BU is purely Bottom-up search, $NN_i(\bar{\alpha})$ is Top-down search guided by a Neural Network (trained on i% of the training data available) choosing context dependent weights, and $NN_i(.)$ is the same without any context information.

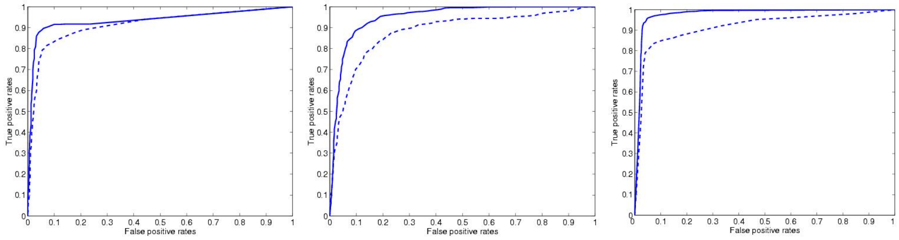


Figure 16: ROC curves for SIFT based (left), color histogram based (middle) and combined (right) object detection, with (solid) and without (dashed) foveated segmentation.

886 object recognition tasks using images provided in real-time using the binocular attention system described in earlier sections. The ROC curves in Fig. 16 illustrate the recognition performance with and without segmentation for both methods individually, as well as for a combination. The combination is done using a binary operator that is learned using a support vector machine (SVM) approach, (Björkman and Eklundh, 2006).

Since we are also interested in object manipulation, we combine the results of appearance and shape recognition where the shape here is represented by the width, breadth and height of the object. Thus, we bind the object identity to its known intrinsic attributes. This binding serves two purposes: i) it boosts the recognition rate by disregarding more false positives, ii) it allows for substitution of objects with other "visually similar" objects. This opens up for broader Object-Action-Complex (OAC) categorization of objects and is discussed further in (Huebner *et al.*, 2008a) as in more detail in (Geib *et al.*, 2006; Kraft *et al.*, 2008; Wörgötter *et al.*, 2009). Since "action" here implies possible (stable) grasps, this binding of identity with intrinsic attributes leads to a scenario where objects that resemble each other (in appearance and shape) may be grasped similarly.

7.3 Object Grasping: an Example

Several object grasping trials were performed and the overall performance of the system was qualitatively evaluated in a tabletop scenario. The goal for the robot was to find a desired object or object type and move it to a predefined location. The first task is to find the object of interest. Here the attention system was tuned by our NN, that selected appropriate weights for the SM_{TD} based on task (i.e. object) and context (scene). That gave us hypotheses of where the object of interest might be. Fig. 17 shows two such examples of SM_{TD} when searching for the 'UncleBen' object and the 'yellowCup' object, respectively. Given any of these hypotheses of location, a saccade was performed to redirect the robot's focus to that particular point in the environment. Consequently the binocular system tried to fixate on that point by the fixation mechanisms described earlier.

Next, a segmentation based on disparities, using the table-plane assumption mentioned in Section 5.1, was made on the "thing" of interest. Segmentation results can be viewed as the enclosed regions in the foveal views of the four examples in Fig. 18. One consequence of real world conditions such as noise, varying illumination etc., is that the segmentation are far from perfect. However, following the OAC-concept mentioned earlier, it is not our goal to gather information about the state of the object solely through vision. Instead we want to complement this sensory information through interactions with the object. Therefore, this imperfection is of minor importance, if the grasping yields a successful result.

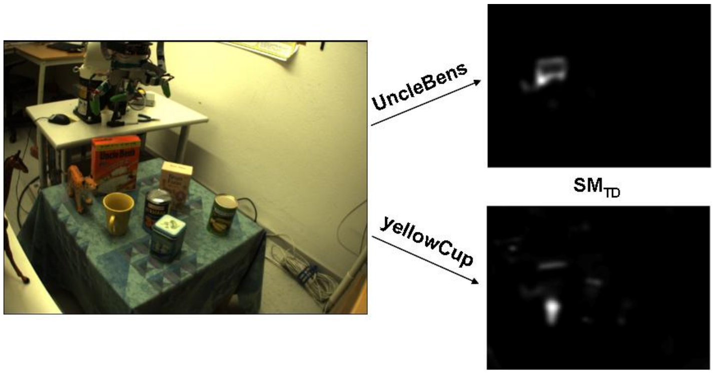
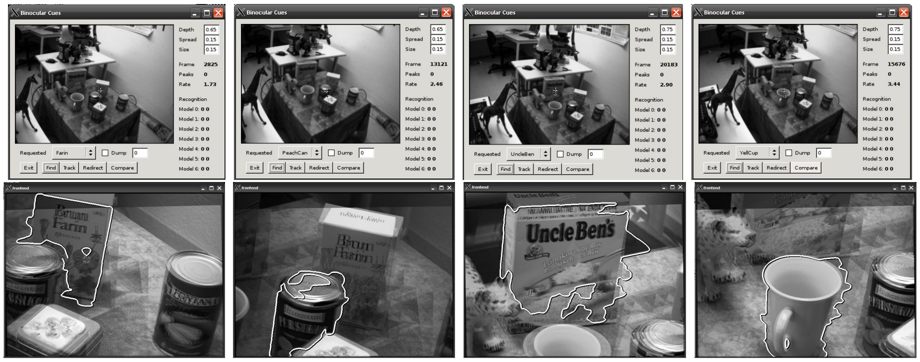


Figure 17: Example with Top-down tuned saliency maps (UncleBens & yellowCup)



Depth 0.65
Spread 0.15
Size 0.15
Frame 2825
Peaks 0
Rate 1.73
Recognition
Model 0: 0 0
Model 1: 0 0
Model 2: 0 0
Model 3: 0 0
Model 4: 0 0
Model 5: 0 0
Model 6: 0 0
Requested Farin
□ Dump 0

Depth 0.65
Spread 0.15
Size 0.15
Frame 13121
Peaks 0
Rate 2.46
Recognition
Model 0: 0 0
Model 1: 0 0
Model 2: 0 0
Model 3: 0 0
Model 4: 0 0
Model 5: 0 0
Model 6: 0 0
Requested PeachCan
□ Dump 0

Depth 0.75
Spread 0.15
Size 0.15
Frame 20183
Peaks 0
Rate 2.90
Recognition
Model 0: 0 0
Model 1: 0 0
Model 2: 0 0
Model 3: 0 0
Model 4: 0 0
Model 5: 0 0
Model 6: 0 0
Requested UncleBen
□ Dump 0

Depth 0.75
Spread 0.15
Size 0.15
Frame 15676
Peaks 0
Rate 3.44
Recognition
Model 0: 0 0
Model 1: 0 0
Model 2: 0 0
Model 3: 0 0
Model 4: 0 0
Model 5: 0 0
Model 6: 0 0
Requested YellCup
□ Dump 0

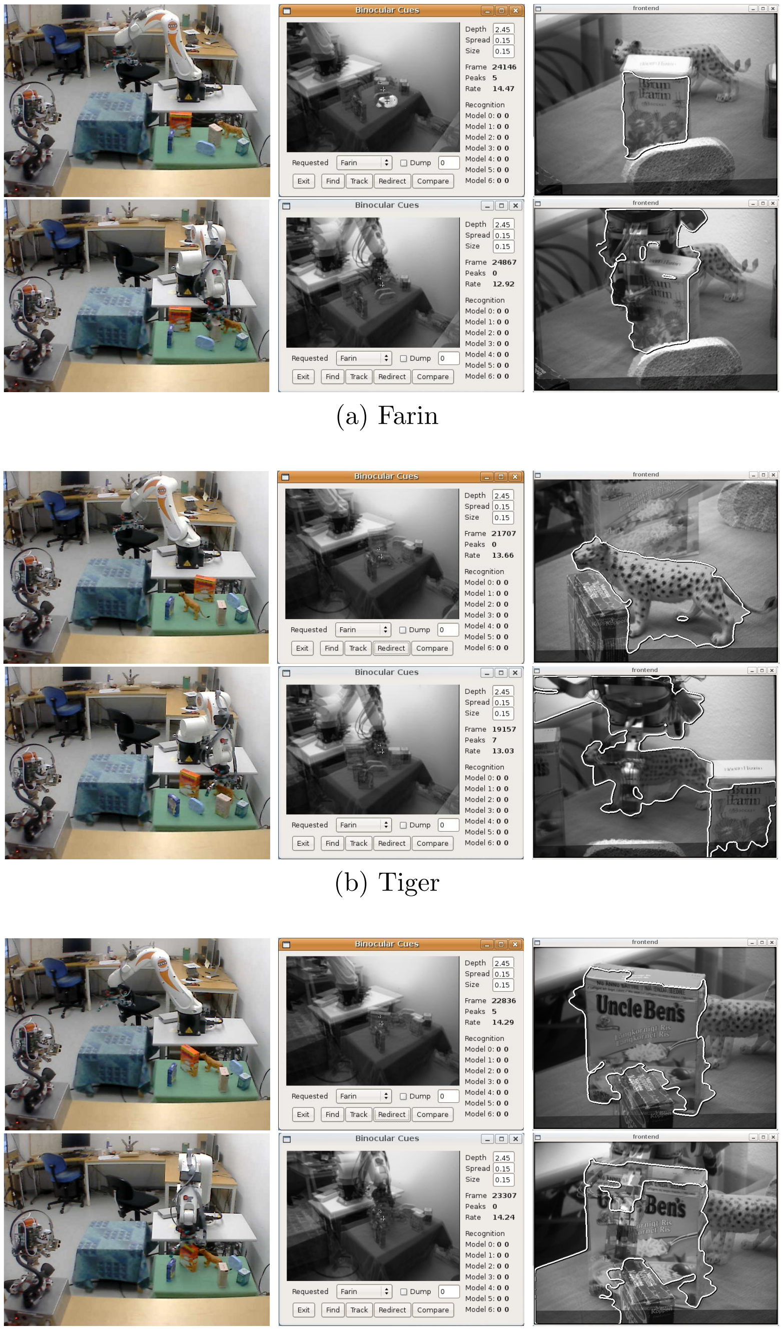
Figure 18: The visual front-end. The top row shows the wide-field view where the visual search selection is made. The bottom row shows the foveal view in which the binocular segmentation and recognition as well as validation is done.

If the segmented region contains the object sought for based on the appearance and intrinsic attributes, the estimated position and orientation is sent to the manipulator. The system then chooses an appropriate grasp based on the intrinsic and extrinsic attributes of the object.

A couple of examples are shown in Fig. 19. The images show the scene before (top rows) and during grasping (bottom rows). One interesting detail seen in these images, is that when the gripper enters the foveal view the fixation-loop adapts to its presence and tries to re-fixate on the point in the center of the image, now being closer to the eyes.

One important detail about this particular implementation is that we have here not included the Bottom-up cues (SM_{BU}) nor the temporal linear combination of the two saliency maps. The reason for this is simply that we were only interested in the Top-down performance of the system. The more dynamic combination of the two saliency maps will be further examined in our future work, where a more "natural" environment with clutter and distractors that might be of importance, will be used.

Imperative in the context is that this is just one example to expose the qualitative properties of the system. A potential quantitative and objective evaluation can be difficult to perform for the complex real-world applications that we are facing. Thus other than a separated part-wise evaluation of the different components of this assembled system, we will not here present any quantitative performance results. However, we do intend to create such evaluation processes in the future to more exactly measure the performance of the system as a whole.



Depth 2.45
Spread 0.15
Size 0.15
Frame 24146
Peaks 5
Rate 14.47
Recognition
Model 0: 0 0
Model 1: 0 0
Model 2: 0 0
Model 3: 0 0
Model 4: 0 0
Model 5: 0 0
Model 6: 0 0
Requested Farin
Dump 0

Depth 2.45
Spread 0.15
Size 0.15
Frame 24867
Peaks 0
Rate 12.92
Recognition
Model 0: 0 0
Model 1: 0 0
Model 2: 0 0
Model 3: 0 0
Model 4: 0 0
Model 5: 0 0
Model 6: 0 0
Requested Farin
Dump 0

(a) Farin

Depth 2.45
Spread 0.15
Size 0.15
Frame 21707
Peaks 0
Rate 13.66
Recognition
Model 0: 0 0
Model 1: 0 0
Model 2: 0 0
Model 3: 0 0
Model 4: 0 0
Model 5: 0 0
Model 6: 0 0
Requested Farin
Dump 0

Depth 2.45
Spread 0.15
Size 0.15
Frame 19157
Peaks 7
Rate 13.03
Recognition
Model 0: 0 0
Model 1: 0 0
Model 2: 0 0
Model 3: 0 0
Model 4: 0 0
Model 5: 0 0
Model 6: 0 0
Requested Farin
Dump 0

(b) Tiger

Depth 2.45
Spread 0.15
Size 0.15
Frame 22836
Peaks 5
Rate 14.29
Recognition
Model 0: 0 0
Model 1: 0 0
Model 2: 0 0
Model 3: 0 0
Model 4: 0 0
Model 5: 0 0
Model 6: 0 0
Requested Farin
Dump 0

Depth 2.45
Spread 0.15
Size 0.15
Frame 23307
Peaks 0
Rate 14.24
Recognition
Model 0: 0 0
Model 1: 0 0
Model 2: 0 0
Model 3: 0 0
Model 4: 0 0
Model 5: 0 0
Model 6: 0 0
Requested Farin
Dump 0

(c) UncleBen

Figure 19: Finding and manipulating three different objects. In each of the three examples, the top row shows the state of the system before grasping and the bottom row shows the attempted grasp. Best viewed in color.

8 Conclusions

The goal for the future development of intelligent, autonomous systems is to equip them with the ability to achieve cognitive proficiency by acquiring new knowledge through interaction with the environment and other agents, both human and artificial. The base for acquiring new knowledge is the existence of a strong perception-action components where flexible and robust sensing plays a major role. Visual sensing has during the past few years proved to be the sensory modality that offers the richest information about the environment. Despite this it has typically been used for well defined, specific tasks for the purpose of coping with the complexity and noise effects.

For the past several years, our work has concentrated on the development of general systems and their applications in navigation and object manipulation applications. The work presented here is in line with the development of such a system, except that we have kept our attention on the design and development of a vision system architecture that allows for more general solutions in service robot settings.

Our system uses binocular cues extracted from a system that is based on two sets of cameras: a wide field for attention and a foveal one for recognition and manipulation. The calibration of the system is performed online and facilitates the information transfer between the two sets of cameras. The importance and role of Bottom-up and Top-down attention is also discussed and shown how biased saliency for visual search tasks can be defined. Here, intensity, color, orientation and texture cues facilitate the context learning problem. The attentional system is then complemented with binocular information to make the system more likely to pop out regions of interest suitable for manipulation. We have also discussed how the attentional system can adapt to context changes.

In relation to manipulation, we show and discuss how the system can be used for manipulation of objects for which geometrical model is not known in advance. Here, the primary interest is to pick up an object and retrieve more information about it by obtaining several new views. Finally, we present experimental results of each, and give an example of how the system has been used in one of the object pick-up scenarios. As mentioned, this was just one example to expose the qualitative properties of the system. In real-world applications it is in general difficult to perform extensive experiments thus evaluation different modules in a number of benchmarking tasks may be one of the solutions. Current directions in the area of robotics and different competitions of mobile and manipulation settings are pointing in this direction. However, there are still very few systems that use active vision.

Regarding the limitations of the presented systems we first need to touch upon the issue of using four-camera setup. As we discussed, the ability to use wide-field and narrow-field cameras is good but it is not necessary in all applications. The area of robot navigation and localization, which is currently going more into direction of using visual sensing, may not necessary require such a setup. In addition, one may argue that if a camera system is placed on a mobile base, the robot can move toward the object for achieving a better view of the object. Our opinion is still that changing between the cameras may be faster and alleviates the need for obstacle avoidance and path planning that moving a platform commonly requires.

Another aspect is the comparison with a zooming camera system. Our opinion here is that changing zoom and fixating on a target results in loosing the wide-field coverage that may be necessary when re-detection of objects for tracking is attempted.

An interesting research issue is related to further learning of object attributes and affordances. The affordances need to be meaningful and related to tasks a robot is expected to execute. In addition, using a interactive setup where a robot can grasp objects, offers more freedom in terms of what attributes can be verified, e.g. hollowness, or extracted, e.g. weight.

The current system does not perform any long-term scene representation, i.e. there is no real memory in the system apart from storing the individual object's attributes. One aspect of future research is therefore to investigate large-scale spatial/temporal representations of the environment. Some aspects of our previous work in the area of Simultaneous Localization and Mapping as well as semantic reasoning will be exploited here.

In the current work, we study the issue of calibration between the head cameras but the hand-eye calibration is not really tackled. The system is complex so that there are also neck motions that could be taken into account for online learning of the hand-head-eye calibration. This process is also related to the issue of smooth pursuit once moving objects are considered. An application may be just the classical object tracking or observation of human activities. An active vision system allows for fixation on parts of human that are important for the task at hand: fixating on mouth when a human is peaking or fixating on the hands when a human is manipulating objects. In this case, the interplay between the saccades and the neck motions is an interesting problem and the solution can be biologically motivated: fast movement of eyes followed by a slower movement of the neck and the mutual compensation.

One of the aspects not studied is vision based closed-loop control of arm motions: visual servoing. It would be interesting to explore, similarly to the partitioned control approaches based on the integration of image based and position based control, to what extent the change between using one of the four cameras at the time can cope with the problems of singularities and loss of features that are inherit to the image based and position based visual servoing approaches.

The most immediate extension of the system is the integration of the object attributes that are extracted based purely on visual input and the ones that are further extracted once the object has been picked up by the robot arm. These include both more detailed visual representation such as several views of the unknown object and the attributes that are extracted by other sensors on the robot: force-torque sensor in the wrist and haptic sensors on the fingers of the robot hand.

Appendix A: Index to Multimedia Extensions

The multimedia extensions to this article can be found online by following the hyperlinks from www.ijrr.org.

Table 1: Index to Multimedia Extensions

Extension	Media Type	Description
1	Video	The grasping experiment described in Section 7.3.
2	Images	Large size multipage(6)-Tiff of Fig. 19(a).
3	Images	Large size multipage(6)-Tiff of Fig. 19(b).
4	Images	Large size multipage(6)-Tiff of Fig. 19(c).

Acknowledgments

This work has been supported by EU through the project PACO-PLUS, FP6-2004-IST-4-27657 and GRASP, IST-FP7-IP-215821.

References

Asfour, T., Regenstein, K., Azad, P., Schroder, J., Bierbaum, A., Vahrenkamp, N., and Dillmann, R. (2006). ARMAR-III: An Integrated Humanoid Platform for Sensory-Motor Control. In *6th IEEE-RAS International Conference on Humanoid Robots*, pages 169–175.

Bicchi, A. and Kumar, V. (2000). Robotic Grasping and Contact: A Review. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 348–353.

Björkman, M. and Eklundh, J.-O. (2002). Real-Time Epipolar Geometry Estimation of Binocular Stereo Heads. *IEEE Trans. Pattern Analysis and Machine Intelligence*, **24**(3), 425–432.

Björkman, M. and Eklundh, J.-O. (2006). Vision in the real world: Finding, attending and recognizing objects. *International Journal of Imaging Systems and Technology*, **16**(5), 189–209.

Björkman, M. and Kragic, D. (2004). Combination of Foveal and Peripheral Vision for Object Recognition and Pose Estimation. In *IEEE Int. Conf. on Robotics and Automation, ICRA'04*, volume 5, pages 5135–5140.

Breazeal, C. and Scassellati, B. (1999). A Context-Dependent Attention System for a Social Robot. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 1146–1153. Morgan Kaufmann Publishers Inc.

Castiello, U. (2005). The Neuroscience of Grasping. *Nature Neuroscience*, **6**, 726–736.

Choi, S., Ban, S., and Lee, M. (2004). Biologically Motivated Visual Attention System using Bottom-Up Saliency Map and Top-Down Inhibition. *Neural Information Processing - Letters and Review*, **2**, 19–25.

Clark, J. J. and Ferrier, N. J. (1992). Attentive Visual Servoing. In *Active Vision*, pages 137–154. MIT Press.

Coelho, J., Piater, J., and Grupen, R. (2001). Developing Haptic and Visual Perceptual Categories for Reaching and Grasping with a Humanoid Robot. *Robotics and Autonomus Systems*, **37**, 195–218.

Coelho Jr., J., Souccar, K., and Grupen, R. (1998). A Control Basis for Haptically-Guided Grasping and Manipulation. Technical Report CMP-SCI Technical Report 98-46, Dept. Computer Science, University of Massachusetts.

Comaniciu, D. and Meer, P. (2002). Mean Shift: A Robust Approach toward Feature Space Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(5), 603–619.

Draper, B. and Lionelle, A. (2003). Evaluation of Selective Attention under Similarity Transforms. In *Proc. International Workshop on Attention and Performance in Computer Vision*, pages 31–38.

Ekvall, S. and Kragic, D. (2004). Interactive Grasp Learning Based on Human Demonstration. In *IEEE/RSJ International Conference on Robotics and Automation*, pages 3519–3524.

Ekvall, S. and Kragic, D. (2005). Receptive Field Cooccurrence Histograms for Object Detection. In *Proc. IEEE/RSJ International Conference Intelligent Robots and Systems, IROS'05*, pages 84–89.

Ekvall, S., Kragic, D., and Jensfelt, P. (2007). Object Detection and Mapping for Service Robot Tasks. *Robotica*, **25**, 175–187.

Fairley, S., Reid, I., and Murray, D. (1998). Transfer of Fixation Using Affine Structure: Extending the Analysis to Stereo. *International Journal of Computer Vision*, **29**(1), 47–58.

Fischler, M. and Bolles, R. (1981). Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, **24**(6), 381–395.

Frintrop, S. (2006). VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search. *Lecture Notes in Computer Science*, **3899**.

Geib, C., Mourao, K., Petrick, R., Pugeault, N., Steedman, M., Krüger, N., and Wörgötter, F. (2006). Object Action Complexes as an Interface for Planning and Robot Control. In *Workshop: Towards Cognitive Humanoid Robots at IEEE RAS Int Conf. Humanoid Robots*.

Gevers, T. and Smeulders, A. (1999). Color Based Object Recognition. *Pattern Recognition*, **32**(3), 453–464.

Grieg, D., Porteous, B., and Seheult, A. (1989). Exact Maximum A Posteriori Estimation for Binary Images. *Journal of Royal Statistical Society - B*, **51**(2), 271–279.

Harris, C. and Stephens, M. (1988). A Combined Corner and Edge Detector. In *Proceedings of the of the 4th Alvey Vision Conference*.

Hartley, R. and Zisserman, A., editors (2000). *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY.

Haykin, S. (1994). *Neural Networks: A Comprehensive Foundation*. Prentice Hall, Upper Saddle River, NJ.

Heidemann, G., Rae, R., Bekel, H., Bax, I., and Ritter, H. (2004). Integrating Context-Free and Context-Dependent Attentional Mechanisms for Gestural Object Reference. *Machine Vision and Applications*, **16**(1), 64–73.

Hu, Y., Xie, X., Ma, W.-Y., Chia, L.-T., and Rajan, D. (2004). Salient Region Detection using Weighted Feature Maps based on the Human Visual Attention Model. In *IEEE Pacific-Rim Conference on Multimedia*, pages 993–1000.

Huebner, K., Björkman, M., Rasolzadeh, B., Schmidt, M., and Kragic, D. (2008a). Integration of Visual and Shape Attributes for Object Action Complexes. In *6th International Conference on Computer Vision Systems*, volume 5008 of *Lecture Notes in Artificial Intelligence*, pages 13–22. Springer-Verlag.

Huebner, K., Ruthotto, S., and Kragic, D. (2008b). Minimum Volume Bounding Box Decomposition for Shape Approximation in Robot Grasping. In *IEEE International Conference on Robotics and Automation*, pages 1628–1633.

Itti, L. (2000). *Models of Bottom-Up and Top-Down Visual Attention*. Ph.D. thesis, California Institute of Technology, Pasadena, CA, USA.

Itti, L. and Koch, C. (2001). Computational Modeling of Visual Attention. *Nature Reviews Neuroscience*, **2**(3), 194–203.

Itti, L., Koch, C., and Niebur, E. (1998). A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**(11), 1254–1259.

Kamon, I., Flash, T., and Edelman, S. (1998). Learning Visually Guided Grasping: A Test Case in Sensorimotor Learning. *IEEE Transactions on Systems, Man and Cybernetics*, **28**(3), 266–276.

Koch, C. and Ullman, S. (1985). Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry. *Human Neurobiology*, **4**(4), 219–227.

Koike, T. and Saiki, J. (2002). Stochastic Guided Search Model for Search Asymmetries in Visual Search Tasks. In *2nd International Workshop on Biologically Motivated Computer Vision*, volume 2525 of *Lecture Notes in Computer Science*, pages 408–417.

Kraft, D., Baseski, E., Popovic, M., Krüger, N., Pugeault, N., Kragic, D., Kalkan, S., and Wörgötter, F. (2008). Birth of the Object: Detection of Objectness and Extraction of Object Shape through Object Action Complexes. *International Journal of Humanoid Robotics*, **5**, 247–265.

Kragic, D. and Kyrki, V. (2006). Initialization and System Modeling in 3-D Pose Tracking. In *In IEEE International Conference on Pattern Recognition 2006*, pages 643–646.

Kragic, D., Björkman, M., Christensen, H., and Eklundh, J.-O. (2005). Vision for Robotic Object Manipulation in Domestic Settings. *Robotics and Autonomous Systems*, **52**(1), 85–100.

Kuniyoshi, Y., Kita, N., Sugimoto, K., Nakamura, S., and Suehiro, T. (1995). A Foveated Wide Angle Lens for Active Vision. In *International Conference on Robotics and Automation (ICRA95)*, pages 2982–2988.

Lee, K., Buxton, H., and Feng, J. (2003). Selective Attention for Cueguided Search using a Spiking Neural Network. In *International Workshop on Attention and Performance in Computer Vision*, pages 55–62.

Li, Z. (2002). A Saliency Map in Primary Visual Cortex. *Trends in Cognitive Sciences*, **6**(1), 9–16.

Longuet-Higgins, H. (1980). The Interpretation of a Moving Retinal Image. In *Philosophical Trans. Royal Society of London, B-208*, pages 385–397.

Longuet-Higgins, H. (1981). A Computer Algorithm For Reconstructing a Scene From Two Projections. *Nature*, **293**, 133–135.

Lowe, D. (1999). Object Recognition From Local Scale-Invariant Features. In *IEEE International Conference on Computer Vision*, pages 1150–1157.

Morales, A., Recatalá, G., Sanz, P., and del Pobil, A. (2001). Heuristic Vision-Based Computation of Planar Antipodal Grasps on Unknown Objects. In *IEEE International Conference on Robotics and Automation*, pages 583–588.

Morales, A., Chinellato, E., Fagg, A., and del Pobil, A. (2004). Using Experience for Assessing Grasp Reliability. *International Journal of Humanoid Robotics*, **1**(4), 671–691.

Moren, J., Ude, A., Koene, A., and Cheng, G. (2008). Biologically-Based Top-Down Attention Modulation for Humanoid Interactions. *International Journal of Humanoid Robotics*, **5**(1), 3–24.

Namiki, A., Imai, Y., Ishikawa, M., and Kaneko, M. (2003). Development of a High-speed Multifingered Hand System and Its Application to Catching. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2666–2671.

Navalpakkam, V. and Itti, L. (2003). Sharing Resources: Buy Attention, Get Recognition. In *International Workshop Attention and Performance in Computer Vision*.

Nickerson, S. B., Jasiobedzki, P., Wilkes, D., Jenkin, M., Milios, E., Tsotsos, J., Jepson, A., and Bains, O. N. (1998). The ARK Project: Autonomous Mobile Robots for Known Industrial Environments. *Robotics and Autonomous Systems*, **25**(1–2), 83–104.

Oliva, A., Torralba, A., Castelhano, M., and Henderson, J. (2003). Top-Down Control of Visual Attention in Object Detection. In *International Conference on Image Processing*, pages 253–256.

Olshausen, B., Anderson, C., and van Essen, D. (1993). A Neurobiological Model of Visual Attention and Invariant Pattern Recognition based on Dynamic Routing of Information. *Journal of Neuroscience*, **13**(11), 4700–4719.

Paulus, D., Ahrlichs, U., Heigl, B., Denzler, J., Hornegger, J., Zobel, M., and Niemann, H. (1999). Active Knowledge-Based Scene Analysis. *Proceedings of the First International Conference on Computer Vision Systems*, **1542**, 180–199.

Petersson, L., Jensfelt, P., Tell, D., Strandberg, M., Kragic, D., and Christensen, H. I. (2002). Systems Integration for Real-World Manipulation Tasks. In *IEEE International Conference on Robotics and Automation, ICRA'02*, volume 3, pages 2500–2505.

Platt Jr., R., Fagg, A., and Gruppen, R. (2002). Nullspace Composition of Control Laws for Grasping. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1717–1723, Lausanne, Switzerland.

Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., and Belongie, S. (2007). Objects in Context. In *International Conference on Computer Vision*, pages 1–8.

Ramström, O. and Christensen, H. (2004). Object Detection using Background Context. In *Proc. International Conference of Pattern Recognition*, pages 45–48.

Rasolzadeh, B. (2006). KTH Attention-Table Dataset. In *http://www.e.kth.se/ babak2/database.htm*.

Rasolzadeh, B., Björkman, M., and Eklundh, J. (2006). An Attentional System Combining Top-Down and Bottom-Up Influences. In *Proc. International Cognitive Vision Workshop (ICVW06).*

Rasolzadeh, B., Targhi, A. T., and Eklundh, J.-O. (2007). An Attentional System Combining Top-Down and Bottom-Up Influences. In *WAPCV*, pages 123–140.

Riddoch, M., Humphreys, G., Edwards, S., Baker, T., and Wilson, K. (2001). Seeing the Action: Neuriopsychological Evidence for Action-Based Effects on Object Selection. In *Nature Neuroscience, 4*, pages 84–88.

Sandini, G. and Tagliasco, V. (1980). An Anthropomorphic Retina-like Structure for Scene Analysis. *Computer Graphics and Image Processing*, **14**(3), 365–372.

Scassellati, B. (1998). A Binocular, Foveated, Active Vision System. Technical report, MIT AI Memo 1628.

Shimoga, K. (1996). Robot Grasp Synthesis: A Survey. *International Journal of Robotics Research*, **3**(15), 230–266.

Shiu, Y. and Ahmad, S. (1989). Calibration of Wrist-Mounted Robotic Sensors by Solving Homogeneous Transform Equations of the Form Ax = Xb. *IEEE Transactions on Robotics & Automation*, **5**(1), 16–29.

Siagian, C. and Itti, L. (2007). Biologically-Inspired Robotics Vision Monte-Carlo Localization in the Outdoor Environment. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1723–1730.

Sigala, N. and Logothetis, N. (2002). Visual Categorization Shapes Feature Selectivity in the Primate Temporal Cortex. *Nature*, **415**, 318–320.

Sloman, A. (2001). Evolvable Biologically Plausible Visual Architectures. In *British Machine Vision Conference, BMVC'01*, pages 313–322.

Strat, T. and Fischler, M. (1989). Context-Based Vision: Recognition of Natural Scenes. In *23rd Asilomar Conference on Signals, Systems & Computers*, pages 532–536.

Strat, T. and Fischler, M. (1995). The Use of Context in Vision. In *IEEE Workshop on Context-Based Vision*.

Topp, E. A., Kragic, D., Jensfelt, P., and Christensen, H. I. (2004). An Interactive Interface for Service Robots. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'04)*, pages 3469–3475, New Orleans.

Treisman, A. and Gelade, G. (1980). A Feature Integration Theory of Attention. *Cognitive Psychology*, **12**, 97–136.

Tsai, R. and Lenz, R. (1988). Real Time Versatile Robotics Hand/Eye Calibration Using 3D Machine Vision. In *IEEE International Conference on Robotics and Automation*, pages 554–561.

Tsotsos, J. (1987). Analyzing Vision at the Complexity Level: Constraints on an Architecture, An Explanation for Visual Search Performance, and Computational Justification for Attentive Processes. Technical report.

Tsotsos, J. K., Culhane, S. M., Winky, W. Y. K., Lai, Y., Davis, N., and Nuflo, F. (1995). Modeling Visual Attention via Selective Tuning. *Artificial Intelligence*, **78**(1-2), 507–545.

Ude, A., Gaskett, C., and Cheng, G. (2006). Foveated Vision Systems with Two Cameras per Eye. In *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, pages 3457–3462.

Vijayakumar, S., Conradt, J., Shibata, T., and Schaal, S. (2001). Overt Visual Attention for a Humanoid Robot. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2332–2337.

Wörgötter, F., Agostini, A., Krüger, N., Shylo, N., and Porr, B. (2009). Cognitive Agents – a Procedural Perspective Relying on the Predictability of Object-Action-Complexes. *Robotics and Autonomous Systems*, **57**(4), 420–432.

Ye, Y. and Tsotsos, J. (1999). Sensor Planning in 3D Object Search. *Computer Vision and Image Understanding*, **73**(2), 145–168.

List of Figures

List of Footnotes

[1] The center-surround-differences are a computational model of the center-surround receptive fields composed by ganglion cells in the retina. For details on the across-scale subtraction we refer to Itti's original work.

[2] Note that dilation has been applied for the reprojected segments for the later application of point-based object hypotheses verification. The dilation, the grid approach, as also noisy and incomplete data from stereo cause that reprojections are often little larger or not completely covering the bodies.

[3] The notion of shape is here simplified into 3D size, meaning the approximate width, breadth and height of the object as listed in the intrinsic attribute list.